AI-GENERATED IMAGE DETECTION

By Emereole Chinduji

THURSDAY, 09 MAY 2024

Unit Leader: Hari Pandey

Table of Contents

Abstract

This paper addresses the growing challenge of distinguishing real images from artificially generated ones using artificial intelligence (AI). With the increasing production of AI-generated content in various domains, questions about their authenticity and trustworthiness have become crucial. The main objective addressed in this study is to investigate if whether an effective AI model can be developed to accurately detect AI generated images.

The hypothesis suggests that deep learning techniques will outperform traditional machine learning techniques in identifying AI-generated images based on distinct visual features. The aim is to investigate the importance of developing such a model for ensuring digital content authenticity and integrity.

This project findings reveal that the proposed AI approach achieves high accuracy in detecting AI generated images and surpasses machine learning approaches. This research has significant implications, it sets the stage for future development of digital content verification systems and ethical concerns related to AI-generated media.

Keywords: Artificial Intelligence, Image Detection, Deep Learning, Digital Content Authenticity.

Introduction

In today's digital age, the proliferation of deepfakes and manipulated images poses a significant challenge for individuals and organizations seeking to verify the authenticity of digital content. This issue become increasingly relevant as AI-generated media gains popularity and sophistication, making it harder to distinguish between real and fake images (Uzun 2023).

In this context, the research focuses on developing an effective AI approach to address this real-world problem in the domain of image detection. By achieving high accuracy in detecting AI-generated images, the findings could lead to future advancements in digital content verification systems and raise important ethical considerations regarding AI-generated media.

The first section of this paper will address the significance of the real-world problem and provide a description of how AI will address and solve it. Next, the paper will define the project objectives, then describe the adopted AI approaches and their implementation details. The paper's conclusion will be achieved by reviewing the project and analysing its future research.

Real World Problem

With the proliferation of fake visual content on social media and news platforms it has become increasingly easy for individuals or groups with malicious intentions to generate fake images. The manipulation of image content can result in the spread of misinformation, deception of the public, defamation of individuals or organizations, and even incite violence (David Rozado 2023).

Moreover, AI generated images raises concerns about intellectual property rights in creative industries. This poses a significant threat to creators who rely on the sale or licensing of their original content for income. For instance, an artist could have their work copied by an AI model, which then generates multiple variations of it, leading to widespread distribution and potential loss of revenue (Chesterman 2024).

The process of distinguishing between a real or fake image can take an important amount of time if done through the analyse of human eyes. Ineffective results are commonly observed, particularly as AI algorithms improve their ability to generate convincing images (Yang 2023).

To solve this problem AI must be utilised to develop an effective system capable of distinguishing between real and artificial images, thereby addressing this pressing issue in our increasingly digitized world.

Project Objectives

The project aims to predict if an image is AI-generated or is Real.

To solve the problem with the use of AI the objectives are:

1. **Literature review**: Investigating previews research in the field of image classification using AI image classifier.
2. **Finding a suitable dataset**: The dataset must consist of two sets of images, real images, and ai-generated images.
3. **Data cleaning & EDA (Data exploration)**
4. **Image preprocessing and preparation**
5. **Modelling & Training**: This stage will require selecting different types of algorithms and to train it with the pre-processed/prepared data.
6. **Compare Performance & Selecting the best model**.
7. **Investigate potential issues in the model**: The selected model might experience issues such as underfitting or overfitting during training. This project will investigate potential challenges related to these issues and explore methods for addressing them.
8. **Evaluate model on validation set**.

Adopted AI Approaches

Image classification belongs to a sub-field of AI named computer vision, and the purpose of it is to classify images based on their content (Li and Guo 2013). There are many factors in an image that can affects the process of image classification. For example, image orientation, image scale variation, image deformation, image brightness, or image visibility, present a challenge in the field of computer vision (Hasabo 2020). A literature review is necessary to explore the AI approaches for image classification and get a comprehensive analysis of existing knowledge on image classification (Lu and Weng 2007).

I. Supervised Learning Approach

Traditional machine learning algorithms have generally been used for image classification using supervised learning and have played a significant role in this research area (Wang et al. 2021). The following algorithms are example which are mainly used in image classification:

- **K-Nearest Neighbours (KNN)**: In image classification, KNN, classifies an image by finding the k nearest image that shares the same properties of the nearby image (Huang et al. 2019).
- **Support Vector Machines (SVM):** SVM operates by finding the optimal hyperplane in high-dimensional space that separates data images from different classes with the largest margin, ensuring good generalization performance (Mammone et al. 2009).

Deep learning algorithms have recently demonstrated outstanding results in image classification, previous research have even shown that deep learning models like CNNs have achieved state-of-the-art performance in various image classification tasks (Alom et al. 2019).

- **Convolutional Neural Network**: A Convolutional Neural Network (CNN) is a type of deep learning model designed for processing grid-like data, such as images. It uses a series of filters or "convolutions" to extract features from the input data by convolving the filter over the image with some padding and pooling layers down sampling the feature maps (Gupta et al. 2022).

II. Unsupervised Learning Approach

Unsupervised learning approaches are also used in image classification, usually when the images are not labelled. Unsupervised learning techniques have also displayed good results in image classification and even outperformed supervised techniques in some rare cases (Schmarje et al. 2021).

- **Deep Belief Networks (DBN)**: DBN is a deep learning algorithm that can learn a hierarchical representation of image features from unlabelled image data (Wang et al. 2021).

Hypothesis

The belief is that deep learning techniques will be superior to traditional machine learning techniques in classifying images.

Introduction to the dataset

CIFAKE is a large dataset which contains 120,000 images in total at the time of writing this report. The dataset is composed of two classes which are REAL and FAKE, with 60,000 real and 60,000 ai-generated images. REAL images were collected from CIFAR-10 dataset and the synthetic images were generated using Stable Diffusion version 1.4 using the CIFAR-10 images as input (Bird and Lotfi n.d.).

The purpose behind the creation of this dataset was due to the increasing concerns about the authenticity of digital content and its potential misuse. CIFAKE dataset aims to address these issues by providing this dataset (Bird and Lotfi n.d.).

Due to computational limitations of processing a large number of images, only 30% of the dataset images was used in this project, which represents 36297 images in total, with 17934 ai generated images and 17834 real images.

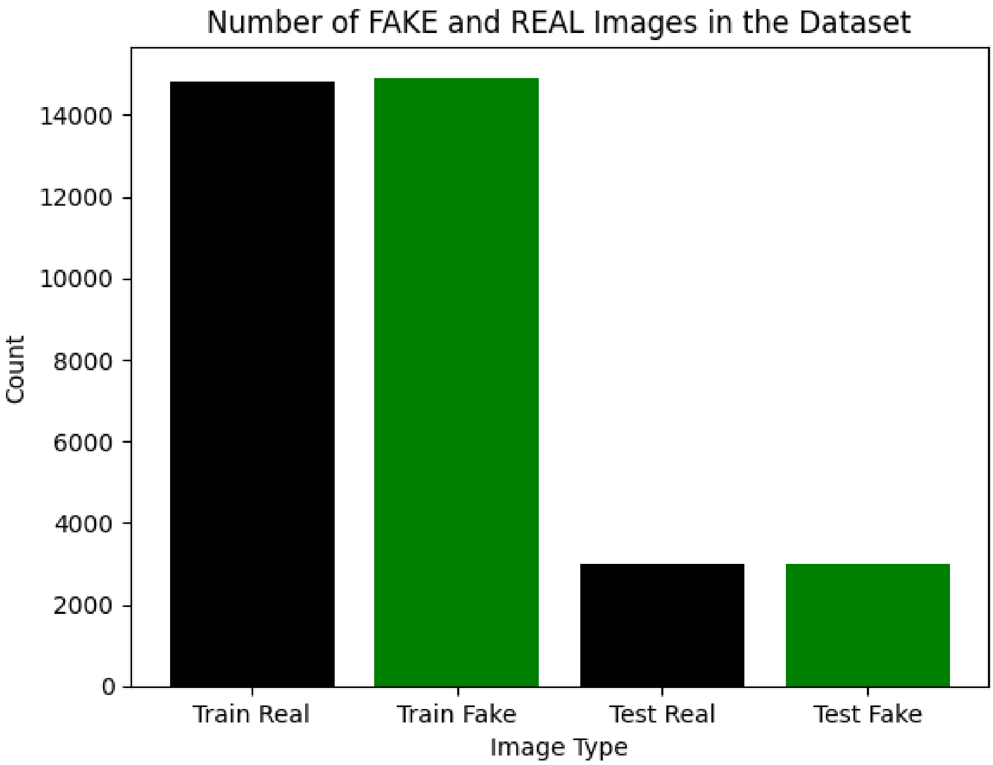
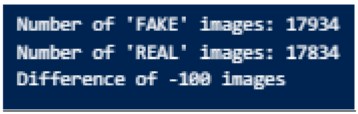


Figure: Number of images inside the dataset



```
Number of 'FAKE' images: 17934
Number of 'REAL' images: 17834
Difference of -100 images
```

Figure: Difference Fake and Real image number inside the dataset

AI approach Implemented.

Only one AI approach was implemented in this project.

I. Supervised learning approach

a. Support Vector Machines

In this project, the LinearSVC algorithm was used. LinearSVC is an implementation of SVM, and it was specifically used in this project because it solves optimisation problem by using a method called Lagrangian. This makes it more efficient and suitable for handling large datasets compared to traditional SVMs (Rosipal et al. 2003).

b. Stochastic Gradient Boosting

The main advantage of using SGB is that it is flexible in handling various types of data. However, it is computationally expensive due to its need to use multiple model evaluation during training (Ye et al. 2009).

c. Convolutional Neural Network

CNNs are particularly effective at recognizing patterns in image data due to their hierarchical representation learning, which allows to learn low-level features like edges and high-level features like objects or shapes. However, it requires significant computational resources, and the training phase is usually longer compared to traditional machine learning algorithms (Gupta et al. 2022).

II. Chosen AI algorithm.

After comparing the performances of the AI models, Convolutional Neural Networks (CNN) was selected. The CNN model demonstrated a more accurate and consistent result compared to the other algorithms. It shows that CNNs are highly effective at learning hierarchical feature representations directly from raw image data without requiring explicit preprocessing or engineering of features.

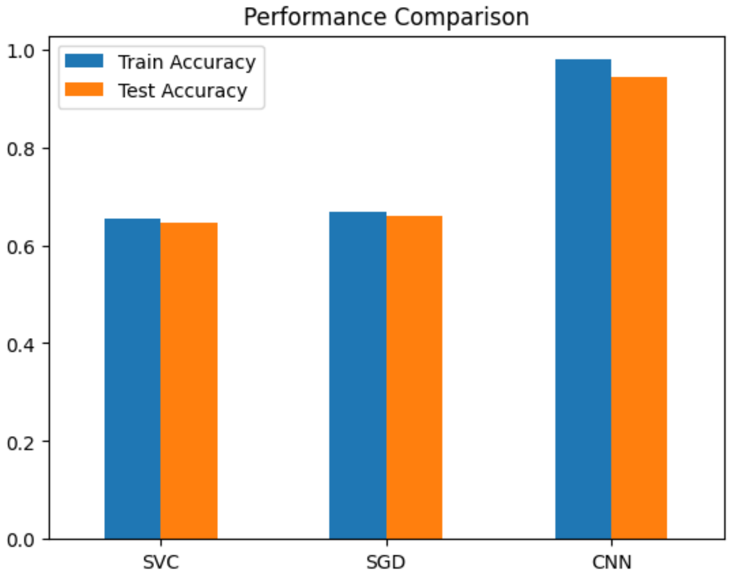


Figure: Models performances comparison

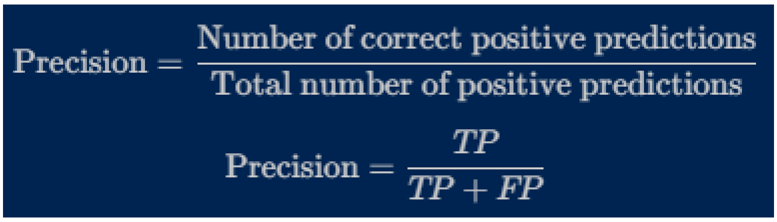
Evaluation, Results and Discussion

1. Evaluation

Multiple metrics were used to evaluate the model. The metrics used are precision score, Recall, Accuracy, and confusion matrix. These metrics were calculated using a sample of the test dataset which contains 32 images (pre-processed images).

I. Precision metric

Precision is a measurement metric that calculates the proportion of all positive results returned by the machine learning model (Davis and Goadrich 2006).

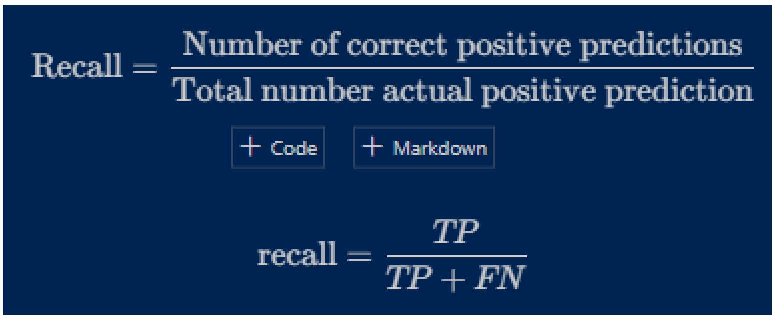


$$\text{Precision} = \frac{\text{Number of correct positive predictions}}{\text{Total number of positive predictions}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

II. Recall metric.

Recall is a measurement metric that measures the proportion of actual positives results identified by the model (Davis and Goadrich 2006).

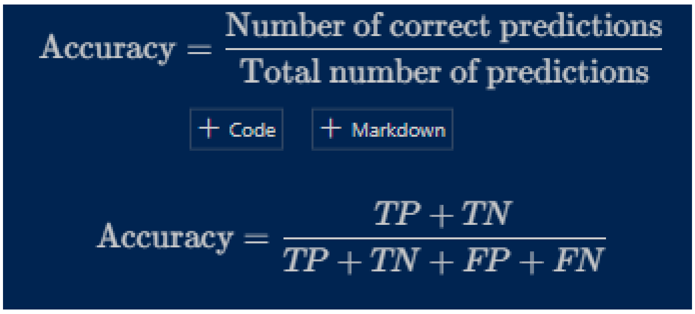


$$\text{Recall} = \frac{\text{Number of correct positive predictions}}{\text{Total number actual positive prediction}}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

III. Accuracy metric

Accuracy is a measurement metric that measures the proportion of correct predictions out of all the instances inside the dataset (Davis and Goadrich 2006).

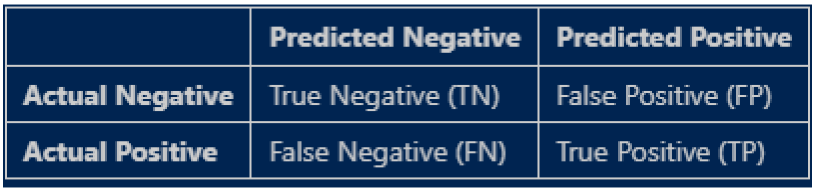


$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

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

IV. Confusion matrix.

A confusion matrix is a table layout that allows for the visualization of an algorithm's performance. The matrix shows a row for each instance in a predicted value, and the column shows the actual value, or vice versa (Javaheri et al. 2013).

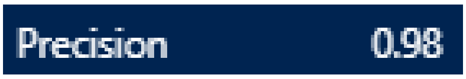


	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

2. Results

I. Precision result.

The model achieved a precision result of:

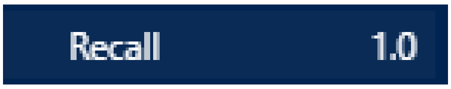


Precision 0.98

The precision achieved a score of 98%. This indicates that out of every 32 instances of image labelled as "positive" or "REAL", 31 were correctly identified as such.

II. Recall result.

The model achieved a recall result of:

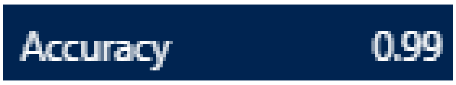


Recall 1.0

The recall value achieved a score of 100%, which means that the model has identified all the positive (REAL) instances of image.

III. Accuracy result

The model achieved an accuracy result of:

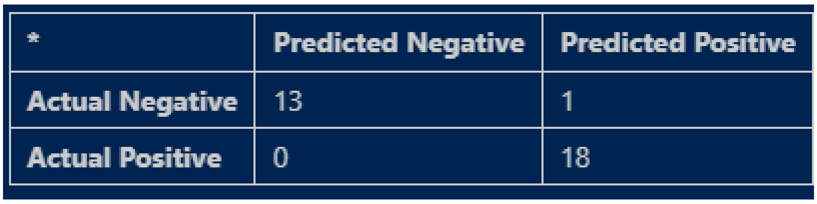


Accuracy 0.99

The accuracy achieved a score of 99%. It indicates that the overall performance of the model predicts REAL and FAKE images with a 99% of accuracy.

IV. Confusion matrix result.

The model confusion matrix result is (see picture 1 in appendix):



*	Predicted Negative	Predicted Positive
Actual Negative	13	1
Actual Positive	0	18

3. Discussion

The CNN model has demonstrated impressive performance in predicting between real and artificial (FAKE) images based on the metrics shown above. This achievement is significant as it gives a solution to the real-world problem, also, it gives room for advancement for digital content verification systems.

Conclusion and future work

In conclusion, the findings from this study demonstrates that building an effective AI model to accurately detect AI generated images is possible and feasible. The detection of AI-generated images is highly accurate, which allows for future system improvements to verify the validity of image content.

Moreover, the project overall demonstrates that the deep learning approach when dealing with images is the most effective way to build a machine learning system.

To expand upon this study, future investigations that focus on examining AI-generated digital content beyond images is needed, specifically in video media. Increasing phenomenon such as deepfakes poses challenges but gives opportunities for research in this area (Li et al. 2018). Therefore, exploring advanced machine learning approaches such as deep learning models to detect fake videos would be a valuable contribution in this field.

References

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S. and Asari, V. K., 2019. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics 2019, Vol. 8, Page 292* [online], 8 (3), 292. Available from: https://www.mdpi.com/2079-9292/8/3/292/htm [Accessed 10 Apr 2024].

Bird, J. J. and Lotfi, A., n.d. CIFAKE: IMAGE CLASSIFICATION AND EXPLAINABLE IDENTIFICATION OF AI-GENERATED SYNTHETIC IMAGES. [online]. Available from: https://huggingface.co/CompVis/stable-diffusion-v1-4 [Accessed 23 Apr 2024].

Chesterman, S., 2024. Good models borrow, great models steal: intellectual property rights and generative AI. *Policy and Society*.

David Rozado, 2023. Danger in the Machine: The Perils of Political and Demographic Biases Embedded in AI Systems | Manhattan Institute. [online]. Available from: https://media4.manhattan-institute.org/sites/default/files/the-perils-of-political-and-demographic-biases-embedded-in-ai-systems.pdf [Accessed 12 Mar 2024].

Davis, J. and Goadrich, M., 2006. The relationship between precision-recall and ROC curves. *ACM International Conference Proceeding Series* [online], 148, 233–240. Available from: https://dl.acm.org/doi/10.1145/1143844.1143874 [Accessed 15 Apr 2024].

Gupta, J., Pathak, S. and Kumar, G., 2022. Deep Learning (CNN) and Transfer Learning: A Review. *Journal of Physics: Conference Series* [online], 2273 (1), 012029. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/2273/1/012029 [Accessed 10 Apr 2024].

Hasabo, I., 2020. *Image Classification using Machine Learning and Deep Learning | by Islam Hasabo | The Startup | Medium* [online]. The Startup. Available from: https://medium.com/swlh/image-classification-using-machine-learning-and-deep-learning-2b18bfe4693f [Accessed 8 Apr 2024].

Huang, S., Huang, M. and Lyu, Y., 2019. A novel approach for sand liquefaction prediction via local mean-based pseudo nearest neighbor algorithm and its engineering application. *Advanced Engineering Informatics*, 41.

Javaheri, S. H., Sepehri, M. M. and Teimourpour, B., 2013. Response Modeling in Direct Marketing. A Data Mining-Based Approach for Target Selection. *Data Mining Applications with R*, 153–180.

Li, X. and Guo, Y., 2013. Adaptive Active Learning for Image Classification.

Li, Y., Chang, M.-C. and Lyu, S., 2018. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. [online]. Available from: https://arxiv.org/abs/1806.02877v2 [Accessed 18 Apr 2024].

Lu, D. and Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* [online], 28 (5), 823–870. Available from: https://www.tandfonline.com/doi/abs/10.1080/01431160600746456 [Accessed 2 May 2024].

Mammone, A., Turchi, M. and Cristianini, N., 2009. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics* [online], 1 (3), 283–289. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/wics.49 [Accessed 8 Apr 2024].

Rosipal, R., Trejo, L. J. and Matthews, B., 2003. Kernel PLS-SVC for Linear and Nonlinear Classification. *Proceedings of the Twentieth International Conference on Machine Learning*.

Schmarje, L., Santarossa, M., Schröder, S. M. and Koch, R., 2021. A Survey on Semi-, Self- and Unsupervised Learning for Image Classification. *IEEE Access*, 9, 82146–82168.

Uzun, L., 2023. ChatGPT and Academic Integrity Concerns: Detecting Artificial Intelligence Generated Content. *Language Education and Technology* [online], 3 (1), 45–54. Available from: http://www.langedutech.com/letjournal/index.php/let/article/view/49 [Accessed 12 Mar 2024].

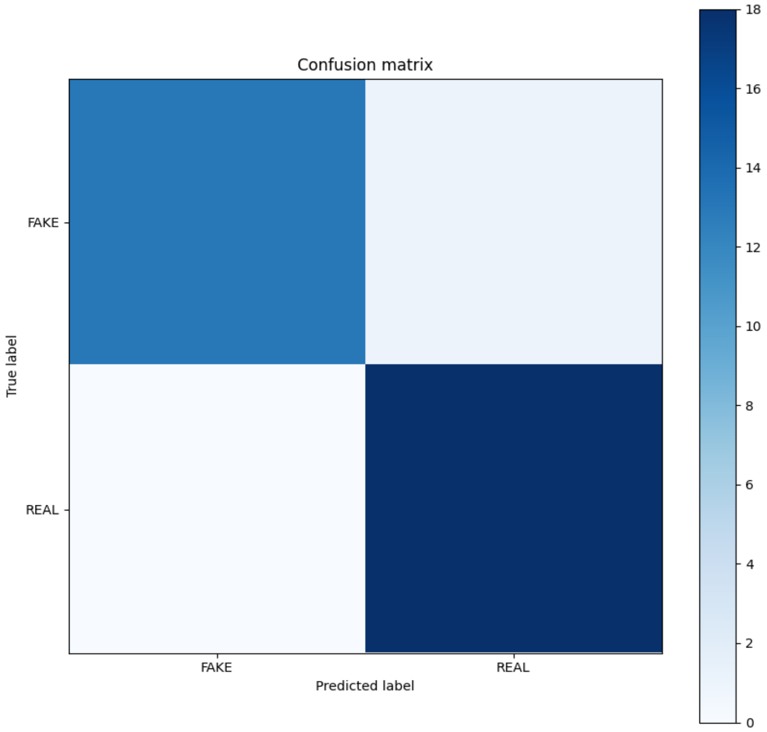
Wang, P., Fan, E. and Wang, P., 2021. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141, 61–67.

Yang, A., 2023. *How can you tell if a photo is AI generated? Here are some tips.* [online]. Available from: https://www.nationalgeographic.com/premium/article/how-can-you-tell-if-a-photo-is-ai-generated-here-are-some-tips [Accessed 6 Apr 2024].

Ye, J., Chow, J. H., Chen, J. and Zheng, Z., 2009. Stochastic gradient boosted distributed decision trees. *International Conference on Information and Knowledge Management, Proceedings* [online], 2061–2064. Available from: https://dl.acm.org/doi/10.1145/1645953.1646301 [Accessed 8 Apr 2024].

Appendix

Picture 1:


Confusion matrix
True label
FAKE
REAL
Predicted label
FAKE
REAL