IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 December 2023, accepted 17 January 2024, date of publication 19 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3356122

RESEARCH ARTICLE

CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images

JORDAN J. BIRD AND AHMAD LOTFI, (Senior Member, IEEE)
Department of Computer Science, Nottingham Trent University, NG1 4FQ Nottingham, U.K.
Corresponding author: Jordan J. Bird (jordan.bird@ntu.ac.uk)

ABSTRACT Recent advances in synthetic data have enabled the generation of images with such high quality that human beings cannot distinguish the difference between real-life photographs and Artificial Intelligence (AI) generated images. Given the critical necessity of data reliability and authentication, this article proposes to enhance our ability to recognise AI-generated images through computer vision. Initially, a synthetic dataset is generated that mirrors the ten classes of the already available CIFAR-10 dataset with latent diffusion, providing a contrasting set of images for comparison to real photographs. The model is capable of generating complex visual attributes, such as photorealistic reflections in water. The two sets of data present as a binary classification problem with regard to whether the photograph is real or generated by AI. This study then proposes the use of a Convolutional Neural Network (CNN) to classify the images into two categories; Real or Fake. Following hyperparameter tuning and the training of 36 individual network topologies, the optimal approach could correctly classify the images with 92.98% accuracy. Finally, this study implements explainable AI via Gradient Class Activation Mapping to explore which features within the images are useful for classification. Interpretation reveals interesting concepts within the image, in particular, noting that the actual entity itself does not hold useful information for classification; instead, the model focuses on small visual imperfections in the background of the images. The complete dataset engineered for this study, referred to as the CIFAKE dataset, is made publicly available to the research community for future work.

INDEX TERMS AI-generated images, generative AI, image classification, latent diffusion.

I. INTRODUCTION

The field of synthetic image generation by Artificial Intelligence (AI) has developed rapidly in recent years, and the ability to detect AI-generated photos has also become a critical necessity to ensure the authenticity of image data. Within recent memory, generative technology often produced images with major visual defects that were noticeable to the human eye, but now we are faced with the possibility of AI models generating high-fidelity and photorealistic images in a matter of seconds. The AI-generated images are now at the quality level needed to compete with humans and win art competitions [1].

Latent Diffusion Models (LDMs), a type of generative model, have emerged as a powerful tool to generate synthetic imagery [2]. These recent developments have caused a paradigm shift in our understanding of creativity, authenticity, and truth. This has led to a situation where consumer-level technology is available that could quite easily be used for the violation of privacy and to commit fraud. These philosophical and societal implications are at the forefront of the current state of the art, raising fundamental questions about the nature of trustworthiness and reality. Recent technological advances have enabled the generation of images with such high quality that human beings cannot tell the difference between a real-life photograph and an image that is no more than a hallucination of an artificial neural network's weights and biases.

Generative imagery that is indistinguishable from photographic data raises questions both ontological, those which concern the nature of being, and epistemological, surrounding the theories of methods, validity, and scope. Ontologically, given that humans cannot tell the difference between images from cameras and those generated by AI models such as an Artificial Neural Network, in terms of digital information, *what is real and what is not?* The

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

epistemological reality is that there are serious questions surrounding the reliability of human knowledge and the ethical implications that surround the misuse of these types of technology. The implications suggest that we are in growing need of a system that can aid us in the recognition of real images versus those generated by AI.

This study explores the potential of using computer vision to enhance our newfound inability to recognise the difference between real photographs and those that are AI-generated. Given that there are many years worth of photographic datasets available for image classification, these provide examples for a model of real images. Following the generation of a synthetic equivalent to such data, we will then explore the output of the model before finally implementing methods of differentiation between the two types of image.

There are several scientific contributions with multidisciplinary and social implications that arise from this study. First, a dataset, called CIFAKE, is generated with latent diffusion and released to the research community. The CIFAKE dataset provides a contrasting set of real and fake photographs and contains 120,000 images (60,000 images from the existing CIFAR-10 dataset (Collection of images that are commonly used to train machine learning and computer vision algorithms available from: https://www.cs.toronto.edu/ kriz/-cifar.html) and 60,000 images generated for this study), making it a valuable resource for researchers in the field. Second, this study proposes a method to improve our waning human ability to recognise AI-generated images through computer vision, using the CIFAKE dataset for classification. Finally, this study proposes the use of Explainable AI (XAI) to further our understanding of the complex processes involved in synthetic image recognition, as well as visualisation of the important features within those images. These scientific contributions provide important steps forward in addressing the modern challenges posed by rapid developments of modern technology and have important implications for ensuring the authenticity and trustworthiness of data.

The remainder of this article is as follows; the state-of-the-art research background is initially explored in Section II with a discussion of relevant related studies in the field. Following this, the methodology followed by this study is detailed in Section III, which provides the technical implementation and the method followed for the binary classification of real versus AI-generated imagery. The results of these experiments are presented with discussion in Section IV before this work is finally concluded, and future work is proposed in Section V.

II. BACKGROUND

The ability to distinguish between real imagery and those generated by machine learning models is important for a number of reasons. Identification of real data provides confirmation on the authenticity and originality of the image; for example, a fine-tuned Stable Diffusion Model (SDM) could be used to generate a synthetic photograph of an individual committing a crime or vice versa, providing false evidence of an alibi for a person who was, in reality, otherwise elsewhere. Misinformation and *fake news* is a significant modern problem, and machine-generated images could be used to manipulate public opinion [3], [4]. Situations where synthetic imagery is used in fake news can promote its false credibility and have serious consequences [5]. Cybersecurity is another major concern, with research noting that synthetically generated human faces can be used in false acceptance attacks and have the potential to gain unauthorised access to digital systems [6], [7]. In [8], it was observed that synthetically generated signatures could overcome signature verification systems with ease.

Latent Diffusion Models are a new approach for generating images, which use attention mechanisms and a U-Net to reverse the process of Gaussian noise and, ultimately, use text conditioning to generate novel images from random noise. Details on the methodological implementation of LDM can be found in Section III. The approach is rapidly developing but is young, and thus literature on the subject is currently scarce. The models are a new approach in the field of generative models; thus, the literature is young, and few applications have been explored. Examples of notable models include Dall-E by OpenAI [9], Imagen from Google [10], and the open source equivalent, SDM from StabilityAI [2]. These models have pushed the boundaries of image quality, both in realism and arguably in artistic ability. This has led to much debate about the professional, social, ethical, and legal considerations of technology [1].

The majority of research in the field is cutting-edge and is presented as preprints and recent theses. In [11], researchers proposed to train SDM on medical imaging data, achieving higher-quality images that could potentially lead to increased model abilities through data augmentations. It is worth mentioning that in [12] and [13], diffusion models were found to have the ability to generate audio and images. In 2021, the results of Yi et al. [14] suggested that diffusion models were highly capable of generating realistic artworks, fooling human subjects into believing that the works were created by human beings. Given this, researchers have noted that diffusion models have a promising capacity for co-creating with human artists [15].

DE-FAKE, proposed by Sha et al. [16], shows that images generated by various latent diffusion approaches may contain digital fingerprints to suggest they are synthetic. Although visual glitches are increasingly rare given the advances in model quality, it may be possible that computer vision approaches will detect these attributes within images that the human eye cannot. The Fourier transforms presented in [17] show visual examples of these digital fingerprints.

When discussing the topic of vision, the results in [18] suggest that optical flow techniques could detect synthetic human faces within the FaceForensics dataset with around 81.61% accuracy. Extending to the temporal domain, [19] proposes recurrence in AI-generated video recognition achieving 97.1% accuracy over 80 frames due to minor visual glitches at the pixel scale. In Wang et al. [20],

EfficientNets and Vision Transformers are proposed within a system that can detect forged images by adversarial models at an F1 score of 0.88 and AUC of 0.95, competing with the state of the art on the DeepFake Detection Challenge dataset while remaining efficient. In the aforementioned study, a Convolutional Neural Network was used to extract features, similarly to the approach proposed in this study, prior to processing using attention-based approaches.

Similarly, convolutional and temporal techniques were proposed in [21] to achieve 66.26% to 91.21% accuracy in a mixed set of synthetic data detection datasets. Chrominance components *CbCr* within a digital image were noted in [22] as a promising route for the detection of minor pixel disparities that are sometimes present within synthetic images.

Human recognition of manipulation within images is waning as a direct result of image generation methods improving. A study by Nightingale et al. [23] in 2017 suggested that humans have difficulty recognising photographs that have been edited using image processing techniques. Since this study, there has been nearly five years of rapid development in the field to date.

Reviewing the relevant literature has highlighted rapid developments within AI-generated imagery and the challenges today posed with respect to its detection. Generative models have enabled the generation of high-fidelity, photorealistic images within a matter of seconds that humans often cannot distinguish between when compared to reality. This conclusion sets the stage for the studies presented by this work and argues the need to fill the knowledge gap when it comes to the availability of examples of synthetic data.

III. METHOD

This section describes the methods followed by this study in terms of their technical implementation and application for the detection of synthetic images. This section first describes the collection of data for the real data, and then the methodology followed to generate the synthetic equivalent for comparison. Sections III-A and III-B will describe how 60,000 images are collected and 60,000 images are synthetically generated, respectively. This forms the overall dataset of 120,000 images. Section III-C will then describe the machine learning model engineering approach which aims to recognise the authenticity of the images, before Section III-D notes the approach for Explainable AI to interpret model predictions.

A. REAL DATA COLLECTION

For the class label ''REAL'', interpreted as a positive class value ''1'', data is collected from the CIFAR-10 dataset [24]. It is a dataset of $60,000$, 32×32 RGB images of real subjects divided into ten classes. Classes within the data set are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. There are $6,000$ images per class. For each class, 5,000 images are used for training and $1,000$ for testing, i.e., a testing dataset of 16.6%. In this study, all images from the training dataset are used for the training of positive class ''REAL''. Therefore, $50,000$ are used for training and $10,000$ for testing.

Samples of images within the CIFAR-10 dataset that form the ''REAL'' class can be observed in Figure 1.

B. SYNTHETIC DATA GENERATION

The synthetic images generated for this study use CompVis SD (https://huggingface.co/CompVis/stable-diffusion-v1-4), an open source LDM. The goal is to model the diffusion of image data through a latent space given a textual context. If noise, such as that of a Gaussian distribution, is iteratively added to an image, the image ultimately becomes noise and all prior visual information is lost. To generalise, the reverse of this process is to, therefore, generate a synthetic image from noise. The method of reverse diffusion can be put simply as, given an image x at timestep t, x_t, output the prediction of x_{i-1} through the prediction of noise and subsequent removal by classical means.

A noisy image x_t is generated from the original x_0 by the following:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, \tag{1}$$

where noise is ε, and the adjustment according to the time step t is $\bar{\alpha}$. The method of this study is to make use of the reverse process of 50 noising steps, which from x_{50} will ultimately form x_0, the synthetic image. The neural network ε_θ thus minimises the following loss function:

$$Loss = E_{t,x_0,\varepsilon}\left[||\varepsilon - \varepsilon_\theta(x_t, t)||^2\right]. \tag{2}$$

Further technical details on the approach can be obtained from [2].The model chosen for this approach is Stable Diffusion 1.4, which is trained in the LAION2B-en, LAION-high-resolution and LAION-aesthetics v2.5 + datasets (https://laion.ai/blog/laion-5b/). The aforementioned datasets are a cleaned subset of the original LAION-5B dataset [25], which contains 5.85 billion text-image pairs.

SDM is used to generate a synthetic equivalent to the CIFAR-10 dataset which contains $6,000$ images of ten classes. The classes are airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. Following observations from the CIFAR-10 dataset, this study implements prompt modifiers to increase diversity of the synthetic dataset, which can be observed in Table 1. As in the real data set, $50,000$ images are used for training data and $10,000$ for testing data, provided with a class label to indicate that the image is not real.

C. IMAGE CLASSIFICATION

Image classification is an algorithm that predicts a class label given an input image. The learnt features are extracted from the image and processed in order to provide an output, in this case, whether or not the image is real or synthetic. This subsection describes the selected approach to classification.

In this study, the Convolutional Neural Network (CNN) [26], [27], [28] is employed to learn from the input

FIGURE 1. Examples of images from the CIFAR-10 image classification dataset [24].

TABLE 1. Latent diffusion prompt modifiers for generating the 10-class synthetic dataset. All prompts are preceded by "a photograph of {a/an}" and modifiers are used equally for the 6000 images.

Class Label	Prompt Modifiers
Airplane	:- aircraft, airplane, fighter, flying, jet, plane
Automobile	:- family, new, sports, vintage
Bird	:- flying, in a tree, indoors, on water, outdoors, walking
Cat	:- indoors, outdoors, walking, running, eating, jumping, sleeping, sitting
Deer	:- herd, in a field, in the forest, outdoors, running, wildlife photography
Dog	:- indoors, outdoors, walking, running, eating, jumping, sleeping, sitting
Frog	:- European, in the forest, on a tree, on the ground, swimming, tropical, wildlife photography
Horse	:- herd, in a field, in the forest, outdoors, running, wildlife photography
Ship	:- at sea, boat, cargo, cruise, on the water, river, sailboat, tug
Truck	:- 18-wheeler, car transport, fire, garbage, heavy goods, lorry, mining, tanker, tow

images. It is the concatenation of two main networks with intermediate operations. These are the convolutional layers and the fully connected layers. The initial convolutional network within the overall model is the CNN, which can be operationally generalised for an image of dimensions x and a filter matrix w as follows:

$$(x * w)(i, j) = \sum_{m=1}^{M} \sum_{n=1}^{N} x(i + m - 1, j + n - 1) w(m, n), \quad (3)$$

where (i, j) is the output for the feature map, and (m, n) represents the location of the filter w. The output is derived by applying convolutional operations to the input x with each of the filters (which are learnable) and applying an activation function f, which, in the context of this study, is the Rectified Linear Unit (ReLu) $f(x) = \max(0, x)$.

For an image of $(height, width)$ dimensions and filters depending on the filter kernel of $(height_{kernel})$ and $(width_{kernel})$ with a $stride = 1$ and no padding for simplicity, the output would have dimensions:

$$(height - height_{kernel} + 1, width - width_{kernel} + 1). \quad (4)$$

Then, a pooling operation is performed to reduce the spatial dimensions and flatten the output so it can be entered into densely connected layers. For $L = HWD$ (height, width, and dimensions), the flattened one-dimensional output vector is simply $x = [x_1, x_2, \ldots, x_L]$. The output vector y is ultimately the output from the dense layer(s) as $y = f(W_L + b)$, for the weight matrix W and the bias b. The activation function f in this study, as in CNN, is the ReLu activation function $f(x) = \max(0, x)$.

The goal of the network in this study is to classify whether the image is a real photograph or an image generated by a LDM, and thus is a problem of binary classification. Therefore, the output of the network is a single neuron with the S-shaped Sigmoid activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

The "FAKE" class is 0 and the "REAL" class is 1, therefore, a value closer to either of the two values represents the likelihood of that class. Although this aids in learning, because it is differentiable, the values are rounded to the closest value for inference.

Although the goal of the network is to use backpropagation to reduce binary cross-entropy loss, this study also notes an extended number of classification metrics. These are the Precision, which is a measure of how many of the predictive positive cases are positive, a metric which allows for the analysis of false-positives:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}. \quad (6)$$

The Recall which is a measure of how many positive cases are correctly predicted, which enables analysis of false-negative predictions:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \quad (7)$$

This measure is particularly important in this case, as it is in fraud detection, since a false negative would falsely accuse the author of generating their image with AI. Finally, the F-1 score is considered:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (8)$$

which is a unified metric of precision and recall.

The dataset that forms the classification is the collection of real images and the equivalent synthetic images generated, detailed in Sections III-A and III-B, respectively. 100,000 images are used for training (50,000 real images and 50,000 synthetic images), and 20,000 are used for testing (10,000 real and 10,000 synthetic).

Initially, CNN architectures are benchmarked as a lone feature extractor. That is, the filters of $\{16, 32, 64, 128\}$ are benchmarked in layers of $\{1, 2, 3\}$, flattened, and

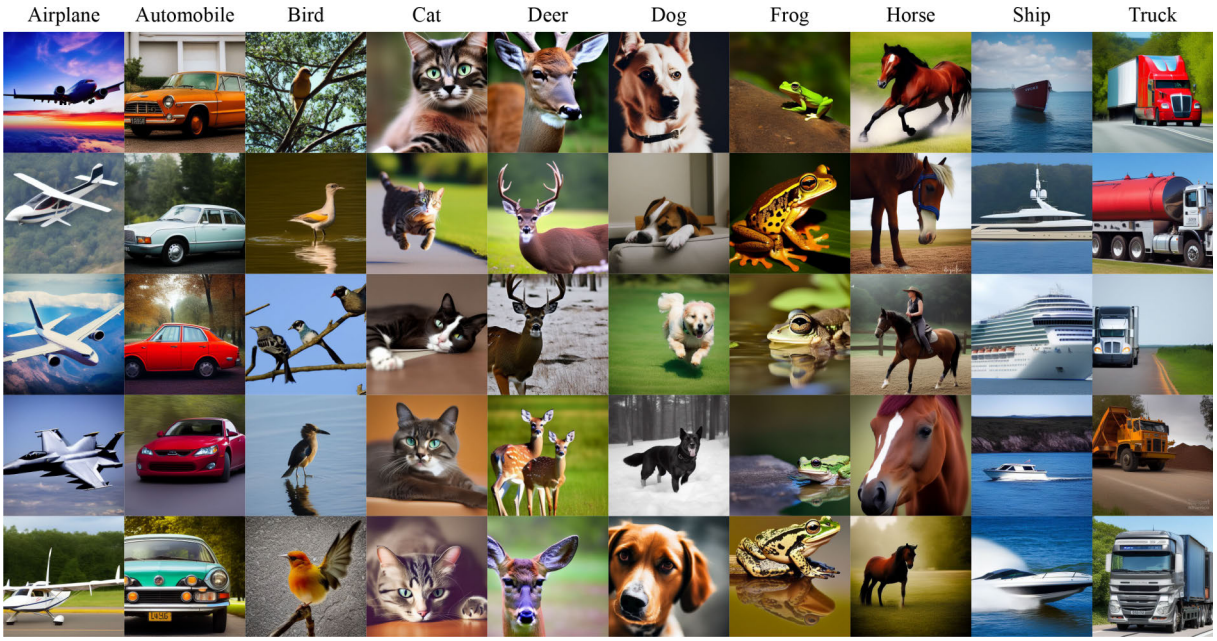

FIGURE 2. Examples of AI-generated images within the dataset contributed by this study, selected at random with regards to their real CIFAR-10 equivalent labels.

connected directly to the output neurone. The topology of the highest performing feature extractor is then used to compare the highest performing dense network featuring $\{32, 64, 128, 256, \ldots, 4096\}$ rectified linear units in layers of $\{1, 2, 3\}$. These 36 artificial neural networks are then compared with regard to classification metrics to derive the topology that performs best.

D. EXPLAINABLE AI

While deep learning approaches often lead to impressive predictive ability, many algorithms are black boxes that provide no reasoning for their classification. The aim of Explainable AI (XAI) is to extract meaning from algorithms and provide readable interpretations of why a prediction or decision is being made [29]. Regarding the experiments in this work, the CNN simply predicts that an image *is* real or synthetic, and then XAI is used to provide interpretations as to *why* the image is real or synthetic.

Given that the literature shows that humans have a major difficulty in recognising synthetic imagery, it is important to display and visualise minor defects within the image that could suggest that it is not real.

The method selected for explainable AI (XAI) and interpretation is Gradient Class Activation Mapping (Grad-CAM) [30]. Grad-CAM interprets the gradients of the predicted class along with the CNN feature maps, which can therefore be spatially localised with respect to the original input (the image) and produce a heatmap. This is generated by the Recitified Linear Unit (ReLU) function as:

$$L^{c}_{Grad-CAM} = ReLU(\sum_{k} \alpha_k A^k), \tag{9}$$

where α_k is the global average pooling $\frac{1}{Z}\sum_i \sum_j \frac{\partial y_c}{\partial A^k_{i,j}}$ of spatial locations Z, and $\frac{\partial y_c}{\partial A^k_{i,j}}$ are the gradients of the model.

The approach is used for interpretation in the final step of this study, given the random data selected from the two classes. Due to the nature of heatmapping, the results of the algorithm are visually interpreted with discussion.

E. EXPERIMENTAL HARDWARE AND SOFTWARE

The neural networks used for the detection of AI-generated images were engineered with the TensorFlow library [31]. All TensorFlow seeds were set to 1 for replicability. The Latent Diffusion model used for the generation of synthetic data was Stable Diffusion version 1.4 [2]; Random seed vectors were denoised for a total of 50 steps to form images and the Euler Ancestral scheduler was used. Synthetic images were rendered at a resolution of 512px before resizing to 32px by bilinear interpolation to match the resolution of CIFAR-10.

All algorithms in this study were executed using a single Nvidia RTX 3080Ti GPU, which has 10,240 CUDA cores, a clock speed of 1.67 GHz, and 12GB GDDR6X VRAM.

IV. RESULTS AND OBSERVATIONS

This section presents examples of the dataset followed by the findings of the planned computer vision experiments. The dataset is also released to the public research community for use in future studies, given the important implications of detecting AI-generated imagery.[1]

[1]The Dataset can be downloaded from: https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images

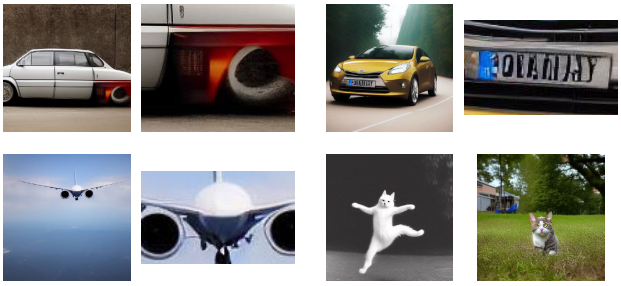

FIGURE 3. Examples of visual defects found within the synthetic image dataset.

TABLE 2. Observed classification accuracy metrics for feature extraction networks.

Filters	Layers		
	1	*2*	*3*
16	90.06	91.46	91.63
32	90.38	**92.93**	92.54
64	90.94	92.71	92.38
128	91.39	92.98	92.07

A. DATASET EXPLORATION

Random samples of images used in this study and within the dataset provided can be observed in Figure 2. Five images are presented for each class label, and all of the images within this figure are synthetic, which have been generated by the SDM. Note within this sample that the images are high-quality and, for the most part, seem to be difficult to discern as synthetic by the human eye. Synthetic photographs are representative of their counterparts from reality and feature complex attributes such as depth of field, reflections, and motion blur.

It can also be observed that there are visual imperfections within some of the images. Figure 3 shows a number of examples of the win of the dataset in which the model has output images with visual glitches. Given that the LAION dataset provides physical descriptions of the image content, little to no information on text is provided, and thus it can be seen that the model produces shapes similar to alphabetic characters. Also observed here is a lack of important detail, such as the case of a jet aircraft that has no cockpit window. It seems that this image has been produced by combining the knowledge of jet aircraft (in particular, the engines) along with the concept of an Unmanned Aerial Vehicle's chassis. Finally, there are also some cases of anatomical errors for living creatures, seen in these examples through the cat's limbs and eyes.

Complex visual concepts are present within much of the dataset, with examples shown in Figure 4. Observe that the ripples in the water and reflections of the entities are highly realistic and match what would be expected within a photograph. In addition to complex lighting, there is also evidence of depth of field and photographic framing.

B. CLASSIFICATION RESULTS

In this subsection, we present the results for the computer vision experiments for image classification. The problem faced by the CNN is that of binary classification, whether or not the image is a real photograph or the output of an LDM.

TABLE 3. Observed validation loss for the filters within the convolutional neural network.

Filters	Layers		
	1	**2**	**3**
16	0.254	0.222	0.21
32	0.237	0.18	0.193
64	0.226	0.196	0.219
128	0.234	0.221	0.259

TABLE 4. Observed validation precision for the filters within the convolutional neural network.

Filters	Layers		
	1	*2*	*3*
16	0.903	0.941	0.921
32	0.878	0.923	0.937
64	0.908	0.947	0.936
128	0.92	0.948	0.94

TABLE 5. Observed validation recall for the filters within the convolutional neural network.

Filters	Layers		
	1	*2*	*3*
16	0.897	0.885	0.911
32	0.938	0.936	0.912
64	0.92	0.904	0.91
128	0.906	0.909	0.898

TABLE 6. Observed validation F1-Score for the filters within the convolutional neural network.

Filters	Layers		
	1	*2*	*3*
16	0.9	0.912	0.916
32	0.907	0.93	0.924
64	0.91	0.925	0.923
128	0.913	0.928	0.919

The validation accuracy of the results and the loss metrics for the feature extractors can be found in Tables 2 and 3, respectively. All feature extractors scored relatively well without the need for dense layers to process feature maps, with an average classification accuracy of 91.79%. The lowest loss feature extractor was found to use two layers of 32 filters, which led to an overall classification accuracy of 92.93% and a binary cross-entropy loss of 0.18. The highest accuracy model, two layers of 128 filters, scored 92.98% with a loss of 0.221.

Extended validation metrics are presented in Tables 4, 5, and 6, which detail validation precision, recall, and F1 scores, respectively. The F1 score, which is a unification of precision and recall, had a mean value of 0.929 with the highest being 0.936. A small standard deviation of 0.003 was observed.

Following these experiments, the lowest-loss feature extractor is selected for further engineering of the network topology. This was the model that had two layers of 32 convolutional filters.

FIGURE 4. A selection of AI-generated images within the dataset. Examples of complex visual attributes generated by the latent diffusion model that include realistic water and reflections.

TABLE 7. Observed validation accuracy for the dense layers within the convolutional neural network.

Neurons	Layers		
	1	*2*	*3*
32	93.2	92.84	92.96
64	93.55	92.73	93.26
128	92.99	93.29	93.18
256	92.97	92.88	92.88
512	93.05	92.58	93.33
1024	92.9	92.91	92.75
2048	92.78	92.76	92.7
4096	92.62	92.52	92.88

TABLE 8. Observed validation loss for the dense layers within the convolutional neural network.

Neurons	Layers		
	1	*2*	*3*
32	0.186	0.182	0.187
64	0.182	0.193	0.177
128	0.187	0.183	0.178
256	0.187	0.192	0.194
512	0.188	0.193	0.184
1024	0.199	0.194	0.192
2048	0.194	0.2	0.219
4096	0.234	0.204	0.19

The results of the general network engineering are presented in Tables 7 and 8, which contain the validation accuracy and loss, respectively. The lowest loss observed was 0.177 binary cross-entropy when the CNN was followed by three layers of 64 rectified linear units. The highest accuracy, on the other hand, was 93.55%, which was achieved by implementing a single layer of 64 neurons.

Additional validation metrics for precision, recall, and F-1 score are also provided in Tables 9, 10, and 11, respectively. Similarly to the prior experiments, the standard deviation of F1-scores was relatively low at 0.003. The highest F-1 score was the network that used a single dense layer of 64 rectified linear units, with a value of 0.936. The aforementioned highest F1 score model is graphically detailed in Figure 5 to provide a visual example of the network topology.

TABLE 9. Observed validation precision for the dense layers within the convolutional neural network.

Neurons	Layers		
	1	*2*	*3*
32	0.932	0.916	0.929
64	0.925	0.92	0.93
128	0.948	0.942	0.935
256	0.939	0.926	0.931
512	0.944	0.924	0.946
1024	0.933	0.94	0.939
2048	0.942	0.922	0.929
4096	0.926	0.914	0.923

TABLE 10. Observed validation recall for the dense layers within the convolutional neural network.

Neurons	Layers		
	1	*2*	*3*
32	0.932	0.943	0.93
64	0.948	0.936	0.935
128	0.91	0.923	0.973
256	0.919	0.932	0.926
512	0.915	0.928	0.919
1024	0.925	0.916	0.915
2048	0.912	0.934	0.925
4096	0.926	0.939	0.936

TABLE 11. Observed validation F1-Score for the dense layers within the convolutional neural network.

Neurons	Layers		
	1	*2*	*3*
32	0.932	0.929	0.93
64	0.936	0.928	0.933
128	0.928	0.932	0.932
256	0.929	0.929	0.929
512	0.929	0.926	0.932
1024	0.929	0.928	0.927
2048	0.927	0.928	0.927
4096	0.926	0.926	0.929

Figure 6 shows examples of the interpretation of predictions via Grad-CAM. Brighter pixels in the image represent

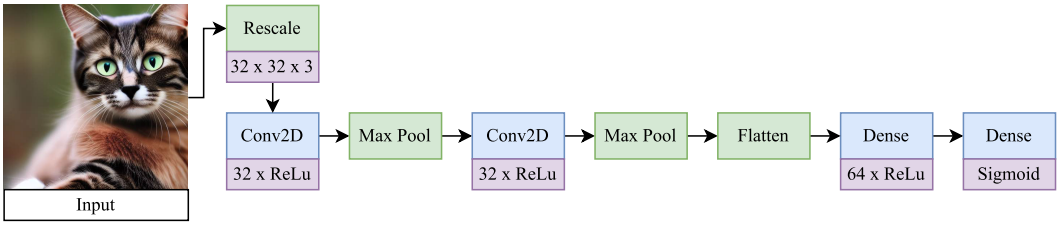

FIGURE 5. An example of one of the final model architectures following hyperparameter search for the classification of real or AI-generated images.

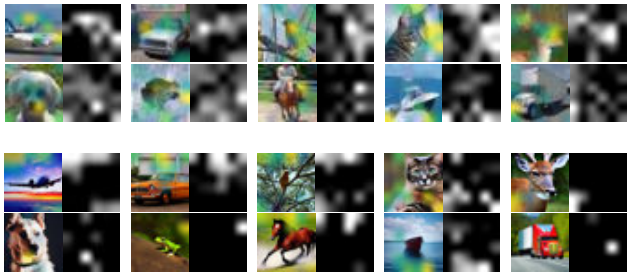

FIGURE 6. Gradient class activation maps (Grad-CAM) overlays and raw heatmaps for prediction interpretation. Top examples show real images and bottom examples show AI-generated images. Brighter pixels represent features contributing to the output class label.

areas that contribute more to the decision of the CNN. It can be observed that there is a significantly different distribution of features given the binary classification problem. Firstly, the classification of real images can be interpreted as a more holistic approach in which the majority of the contents of the image are useful for recognition. However, the classification of synthetic images is somewhat more atomistic and sparse. Note that for the recognition of AI-generated imagery, activation occurs in select parts of the image that are more likely to present visual glitches that are difficult to recognise with the human eye. An example of this can be seen for the image of the frog, where an out-of-focus bokeh is the only attribute that suggests the image is not real. For the truck, only the radiator grill pattern is considered useful for classification.

The XAI approach also shows an interesting mechanic in a more general sense. Given the examples of airplane, bird, frog, horse, and ship, note that the object within the image has little to no class activation overlay whatsoever. This suggests that the actual focus of the image itself, the entity, contains almost no useful features for synthetic image recognition. This suggests that the model is often available to produce a near-perfect representation of the entity.

V. CONCLUSION AND FUTURE WORK

This study has proposed a method to improve our waning ability to recognise AI-generated images through the use of Computer Vision and to provide insight into predictions with visual cues. To achieve this, this study proposed the generation of a synthetic dataset with Latent Diffusion, recognition with Convolutional Neural Networks, and interpretation through Gradient Class Activation Mapping. The results showed that the synthetic images were high quality and featured complex visual attributes, and that binary classification could be achieved with around 92.98% accuracy. Grad-CAM interpretation revealed interesting concepts within the images that were useful for predictions.

In addition to the method proposed in this study, a significant contribution is made through the release of the CIFAKE dataset. The dataset contains a total of 120, 000 images (60, 000 real images from CIFAR-10 and 60,000 synthetic images generated for this study). The CIFAKE dataset provides the research community with a valuable resource for future work on the social problems faced by AI-generated imagery. The dataset provides a significant expansion of the resource availability for the development and testing of applied computer vision approaches to this problem.

The reality of AI generating images that are indistinguishable from real-life photographic images raises fundamental questions about the limits of human perception, and thus this study proposed to enhance that ability by *fighting fire with fire*. The proposed approach addresses the challenges of ensuring the authenticity and trustworthiness of visual data.

Future work could involve exploring other techniques to classify the provided dataset. For example, the implementation of attention-based approaches is a promising new field that could provide increased ability and an alternative method of explainable AI. Furthermore, with even further improvements to synthetic imagery in the future, it is important to consider updating the dataset with images generated by these approaches. Furthermore, considering generating images from other domains, such as human faces and clinical scans, would provide additional datasets for this type of study and expand the applicability of our proposed approach to other fields of research.

Finally, in conclusion, this study provides contributions to the ongoing implications of AI-generated images. The proposed approach supports important implications of ensuring data authenticity and trustworthiness, providing not only a system that can recognise synthetic images, but also data and interpretation. The public release of the CIFAKE dataset generated within this study provides a valuable resource for interdisciplinary research.

VI. AVAILABILITY OF DATA AND MATERIALS

The datasets generated and analysed during the current study are available in the CIFAKE repository, https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images.

REFERENCES

[1] K. Roose, "An AI-generated picture won an art prize. Artists aren't happy," *New York Times*, vol. 2, p. 2022, Sep. 2022.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.
[3] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends Cogn. Sci.*, vol. 25, no. 5, pp. 388–402, May 2021.
[4] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, Dec. 2022.
[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, "On the use of Benford's law to detect GAN-generated images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5495–5502.
[6] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.
[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system," *KSII Trans. Internet Inf. Syst.*, vol. 15, no. 3, pp. 1100–1118, Mar. 2021.
[8] J. J. Bird, A. Naser, and A. Lotfi, "Writer-independent signature verification; evaluation of robotic and generative adversarial attacks," *Inf. Sci.*, vol. 633, pp. 170–181, Jul. 2023.
[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.
[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022, *arXiv:2205.11487*.
[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," 2022, *arXiv:2210.04133*.
[12] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," 2023, *arXiv:2301.11757*.
[13] F. Schneider, "ArchiSound: Audio generation with diffusion," M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.
[14] D. Yi, C. Guo, and T. Bai, "Exploring painting synthesis with diffusion models," in *Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, Jul. 2021, pp. 332–335.
[15] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, "ArtVerse: A paradigm for parallel human–machine collaborative painting creation in Metaverses," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 4, pp. 2200–2208, Apr. 2023.
[16] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models," 2022, *arXiv:2210.06998*.
[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," 2022, *arXiv:2211.00680*.
[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.
[19] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-modal multi-scale transformers for Deepfake detection," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 615–623.
[21] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A hybrid CNN-LSTM model for video Deepfake detection by leveraging optical flow features," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.
[22] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," *Signal Process.*, vol. 174, Sep. 2020, Art. no. 107616.
[23] S. J. Nightingale, K. A. Wade, and D. G. Watson, "Can people identify original and manipulated photos of real-world scenes?" *Cognit. Res., Princ. Implications*, vol. 2, no. 1, pp. 1–21, Dec. 2017.
[24] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.
[25] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5B: An open large-scale dataset for training next generation image-text models," 2022, *arXiv:2210.08402*.
[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[28] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.
[29] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, Dec. 2019, Art. no. eaay7120.
[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[31] M. Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.google.com/search?q=httpsSoftware+available+from+tensorflow.org&rlz=1C1GCEU_enIN1087IN1087&oq=httpsSoftware+available+from+tensorflow.org&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIKCAEQABiABBiiBDIKCAIQABiiBBiJBTIKCAMQABiABBiiBDIKCAQQABiABBiiBDIKCAUQABiABBiiBNIBCDY0MDZqMGo5qAIAsAIA&sourceid=chrome&ie=UTF-8

JORDAN J. BIRD received the B.Sc. and Ph.D. degrees in computer science from Aston University, U.K. He is currently a Senior Lecturer of computer science with Nottingham Trent University, U.K. He received significant external grant funding toward his research projects, which involve applications of artificial intelligence in the real-world. His research interests include artificial intelligence (AI), focusing on human–robot interaction (HRI), machine learning (ML), deep learning, transfer learning, and data augmentation. His professional academic contributions include roles as a technical program committee member and the deep learning session chair of several international conferences.

AHMAD LOTFI (Senior Member, IEEE) received the B.Sc. degree in control systems from the Isfahan University of Technology, Iran, the M.Tech. degree in control systems from the Indian Institute of Technology Delhi, India, and the Ph.D. degree in learning fuzzy systems from The University of Queensland, Australia, in 1995. He is currently a Professor of computational intelligence with Nottingham Trent University, where he is leading the research group in computational intelligence and applications. He has authored and coauthored over 200 scientific papers in the area of computational intelligence, the Internet of Things, abnormal behavior recognition, and ambient intelligence in highly prestigious journals and international conferences. He received external funding from Innovate UK, EU, and industrial companies to support his research. His research interest includes the identification of progressive changes in behavior of older individuals suffering from dementia or any other cognitive impairments. Accurate identification of progressive changes through utilization of unobtrusive sensor network or robotics platform will enable carers (formal and informal) to intervene when deemed necessary. Research collaboration is established with world-leading researchers. He has involved in collaboration with many healthcare commercial organizations and end-users. He has been invited as an expert evaluator and a panel member of many European and international research programs.