HIGH SCHOOL EDITION
Journal of Student Research

Analyzing and Improving Existing Neural Network-Based Approaches to Identify AI Generated Images

Sanika Rewatkar

Maggie L. Walker Governor's School, USA

ABSTRACT

Images produced by AI software, particularly human faces, have the potential to spread misinformation, thus detection tools are essential to tell facts from fiction. However, current neural network-based tools misclassify images with extreme photography conditions, facial obstructions, or post-processing. This project aimed to address the limitations of existing detection models, especially regarding poorly-photographed or edited images. A MobileNet-v2-based neural network to identify synthetic images was analyzed for possible shortcomings. After obtaining its baseline accuracy, it was stress tested on validation data with brightness, contrast, hue, and saturation edited by set increments. The model performed poorly on processed images and faces with glasses or shadows. Possible limitations were insufficient data augmentation, the Global Average Pooling Layer, too few epochs trained, and inadequate regularization strength or dropout rate. The model was then modified to address these shortcomings by altering the brightness, contrast, hue, and saturation of training data, adding Gaussian noise to training images, implementing a Global Max Pooling Layer, augmenting the number of epochs trained, and changing the regularization strength and dropout rate. The modified model was evaluated on augmented and non-augmented images. Overall, the model improved when Gaussian noise was added with standard deviation 0.1, it was trained for 15 epochs, and there were no dropout layers. This was because the Gaussian noise layer approximated degraded image quality, and the model was too simple to learn image features when trained for fewer epochs or when neurons were disabled. This research could be expanded by testing multiple modifications at once.

Introduction

AI-generated image content is gradually becoming more realistic and difficult to tell apart from content created by humans due to the improvements in leading generative software. Generative Adversarial Networks (GANs) such as StyleGAN2, as seen on the website thispersondoesnotexist.com, can generate increasingly realistic human faces from being trained on an extensive dataset of faces. Recent diffusion models such as DALL-E 2, Midjourney, and Stable Diffusion are capable of creating both photorealistic images and humanlike artwork by processing text prompts. Due to its similarities to authentic content, generated images and videos have the potential to cause major harm by spreading false information online. “Deepfake” video imitations of politicians can be used to mislead citizens about policy decisions. News articles can feature AI-generated content to spread misinformation or even be authored by nonexistent personas using AI-produced avatars. Proper detection tools are essential to identifying synthetic images and preventing the spread of misinformation.

An example of AI-generated imagery being used to further a false narrative is the 2020 case of biased articles regarding the Middle East posted by 19 fake journalists, some of whom used AI-generated profile photos, across various news websites. In June of 2020, the Daily Beast published an article investigating the authors of numerous opinion pieces across news outlets such as the *Washington Examiner*, *The Jerusalem Post*, and the *South China Morning Post* expressing views lauding the United Arab Emirates and heavily criticizing Qatar. They discovered that the writers of these pieces were fabricated personas rather than actual journalists, and some avatars used across news media were produced by GAN software. The fictitious journalists were suspected to be part of a pro-UAE propaganda network attempting to sway opinions in favor of the UAE and against Qatar. Some articles by the network were shared by politicians such as the French senator Nathalie Goulet and people adjacent to politics such as the co-founder of the group Students for Trump. As soon as the network was brought to light, many of the news outlets removed the personas' op-eds and issued disclaimers about the content [1]. While the pro-UAE propaganda network was uncovered in 2020, numerous recent propaganda campaigns, such as the 2022 deepfake video of the Ukrainian President Volodymyr Zelenskyy calling for surrender placed on a Ukrainian news website by hackers [2] and recent videos in favor of a military coup in Burkina Faso featuring an AI avatar by the text-to-speech video generator Synthesia [3], use AI-generated content to spread their views.

The fictional journalists' work and profile pictures exemplify the difficulties in identifying AI-generated imagery. Images produced by AI software often cannot be identified by reverse image searches, and human audiences often struggle to differentiate them from genuine photographs. While generated images and faces can often contain artifacts that characterize them as AI-generated, such as non-elliptical pupils or uneven catchlights in eyes [4], they are often unnoticeable by an untrained observer. Neural networks are typically used to identify synthetic images using details that humans cannot detect, but their effectiveness is hindered by limitations in their architecture and training data along with the constant advancement of generative AI software. Such tools also often incorrectly classify both real and synthetic images that have been edited, post-processed, or photographed with grainy or poor conditions.

This project aimed to determine in what ways could the limitations of existing detection models, especially in accounting for poorly-photographed or edited images, be addressed to create a model with improved effectiveness. A current neural network-based detection tool, a MobileNet-v2-based model created by Pooja Kabber, Scott Lai, Song Young Oh, and Emma Wang and uploaded to the GitHub repository "Detecting AI-Generated Fake Images" [5], was analyzed to detect possible limitations to its classification accuracy. It was then stress tested on validation data that had been post-processed by editing their brightness, contrast, hue, and saturation by set increments. Using knowledge of its discovered shortcomings, the model's code was modified to address them and increase its classification accuracy. The modified model's accuracy was finally evaluated on both augmented and non-augmented validation data. Conclusions were drawn regarding which modifications resulted in an improved classification accuracy and why such modifications were effective.

Related Work

To combat misinformation, many strategies have been used to detect AI-generated content. While humans are effective at recognizing fake images from earlier generative software, they struggle to distinguish images created by more advanced software. In an experiment to test humans' ability to recognize synthetic faces, participants were given 30 images of faces from a wider database and asked to rate each image based on their confidence it was generated. 15 images were real, 5 were produced by the rudimentary PGGAN, 5 were produced by the slightly more advanced StyleGAN, and 5 were produced by the most advanced model, StyleGAN 2. On average, participants accurately classified the real faces 56% of the time, the PGGAN images 80% of the time, the StyleGAN images roughly 50% of the time, and the StyleGAN2 images 26% of the time. Participants used distortions in the background, the presence of visible artifacts, and inconsistencies across facial features to identify faces as synthetic, all of which are most frequent in faces created by earlier models [6]. As images produced by recent AI software typically do not contain such characteristic artifacts, human observers struggle to identify them as AI-generated.

In a similar experiment, human performance at identifying AI-produced images was compared to that of a classification model. 50 human participants were asked to classify 100 images either found from 500px and Google Images or produced by Midjourney as either real or AI-generated. A convolutional neural network

(CNN), was trained and validated real images and text prompts from CC3M and StyleGAN3-Train and AI-generated images created to mirror the real images by StyleGAN3, Deepfloyd IF, and Stable Diffusion. The neural network classified 86.3% of the images in its validation datasets correctly, whereas human participants only classified 61.3% of the 100 images correctly on average [7]. This proves that neural network-based classification models, most often deep CNNs, are more effective than humans at identifying generated image content. Detection models are trained on datasets of both real and synthetic images beforehand, and CNNs are able to identify features common in AI-generated images that a human observer can miss. Such models typically involve a pretrained CNN, such as ResNet50 or ConvNext-S, as a backbone for the model, since its weights and architecture can be modified to accomplish the task at hand.

Often, neural network-based detection models rely upon certain aspects or artifacts of an image to determine whether it is synthetic. For instance, human faces produced by GANs typically contain abnormalities in the eyes. Several neural network-based detection approaches have been trained to focus on specific eye inconsistencies to identify GAN-produced faces. Faces produced by early GAN software possessed inconsistent highlights and eye colors, so generative models can compare both eyes to see if they are the same. Even current GAN software creates faces whose pupils are non-elliptical, so a detection model can determine if the pupils' boundaries fit an elliptical shape [4]. Detection models can also focus upon one part of the image without explicitly being programmed to. In the case of a detection CNN, researchers sought to use Gradient Class Activation Mapping (Grad-CAM) to determine which features the model used to make a classification. The Grad-CAM heatmaps revealed that the model made decisions based on artifacts in the background rather than at the main focus on the image [8]. Such approaches explain what criteria are used to make a classification and what artifacts are typical of AI image content.

While these models can reach high levels of effectiveness at identifying AI-generated images, they have limitations that can hinder their performance. A feature- or artifact-based approach to identify AI-generated images can be circumvented by editing or processing the image through adversarial attacks. Models that compare highlights in a face's eyes effectively classify images produced by earlier GAN software, but more recent GAN-generated faces do not contain inconsistent highlights. Highlights can also differ naturally if the face is angled, thus a detection model may misclassify an angled face. A CNN that makes classifications based on background artifacts [8] is likely to misclassify AI-generated images with artifacts removed or a cropped background. Neural-networks in particular can be limited by the training and validation data used. CNNs often misclassify images with poor lighting and strong shadows [4], indicating a lack of varied training images and data augmentation. CNNs also incorrectly classify compressed or blurred images, though data augmentation that randomly compresses or blurs training data improves accuracy [9]. Training datasets typically only contain images created by one type of generator and are unable to classify images produced by a different generator. For example, a ResNet50-based classification model achieved 99.9% accuracy when detecting images produced by Stable Diffusion, which produced images in its training dataset, but its accuracy dropped to 54.9% when identifying images synthesized by Midjourney [10]. Such limitations cause detection tools to obtain false positives and false negatives on more varied content. In order to ensure increased general accuracy of AI-generated image detectors, the limitations of classification software must be resolved.

Datasets

The datasets used for the analysis of the initial model and training and validation of the modified model were one dataset of real faces and one dataset of synthetic faces. The dataset of real faces was "Flickr-Faces-HQ" (FFHQ), which contained 70,000 images of which 21,000 were used by the initial model [5] and 1,000 by the new model. The dataset of synthetic faces was "1M AI generated images 128x128", which contained 1,120,885 AI-generated faces of which 1422 were used by the new model. The initial model used a dataset of DALL-E 2-

produced faces similar to the real faces from FFHQ. However, “1M AI generated images 128x128” was substituted as a replacement, as it was unavailable. The FFHQ dataset was developed to create a benchmark for StyleGAN’s performance and was sourced from the image hosting website Flickr. It contained faces across various ethnicities and subjects wearing head coverings such as hats and accessories such as glasses. Image backgrounds were often detailed with scenery such as trees or buildings, though a few images included a solid color background. The dataset itself had already been cleaned by its creators to ensure that all images featured were of human faces [11]. Because Flickr typically hosts high-quality photographs, the majority of the images in this dataset were well-lit, and subjects were typically looking directly at the camera or from a 3/4ths angle.

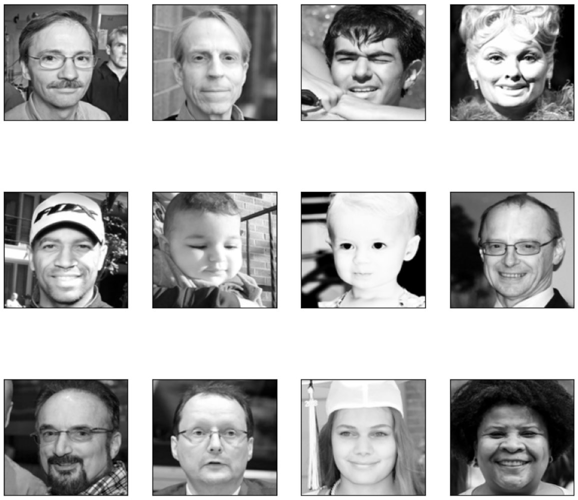


Figure 1. 12 Images of the FFHQ Dataset

The “1M AI generated images 128x128” dataset contained faces produced by NVIDIA StyleGAN [12]. As StyleGAN uses Flickr-Face-HQ to gauge the realism of its generated human faces, images produced by it provide a useful comparison to the FFHQ images to identify AI-generated content. Due to StyleGAN’s utilization of FFHQ as a benchmark, the AI-produced faces spanned a similar variety of ethnicities, and generated faces contained both glasses and head coverings. The dataset did not require cleaning to ensure that it only contains human faces. The images’ backgrounds were also often detailed and contain scenery or other faces. However, as StyleGAN was an earlier, less sophisticated generative AI model, faces produced by it often contained artifacts or background distortions.

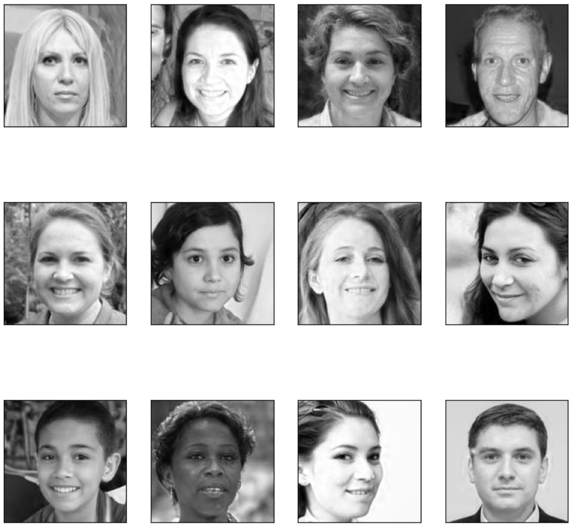


Figure 2. 12 Images of the “1M AI generated images 128x128” Dataset

Analysis of Initial Model

The detection model analyzed to identify its limitations and improve its performance at identifying post-processed images was the MobileNet-v2-based detection model uploaded to the Github repository “Detecting AI-Generated Fake Images” at https://github.com/nogibjj/Detecting-AI-Generated-Fake-Images/tree/main. The model, along with another ResNet50-based model, was created by the Duke University graduate students Pooja Kabber, Scott Lai, Song Young Oh, and Emma Wang. The MobileNet-v2-based model was selected because it achieved a lower classification accuracy than the ResNet50-based model, as the former obtained an accuracy of 0.80 compared to the latter’s accuracy of 0.89. As shown in Figure 3, it consisted of a data augmentation layer, the pretrained CNN MobileNet-v2, a Global Average Pooling Layer, a dense layer with 256 neurons, a dense layer with 128 neurons, and a dense layer with 64 neurons. The final three dense layers contained L2 kernel regularization with a strength of 0.001 and were each followed by a dropout layer with a dropout rate of 0.5 [5].

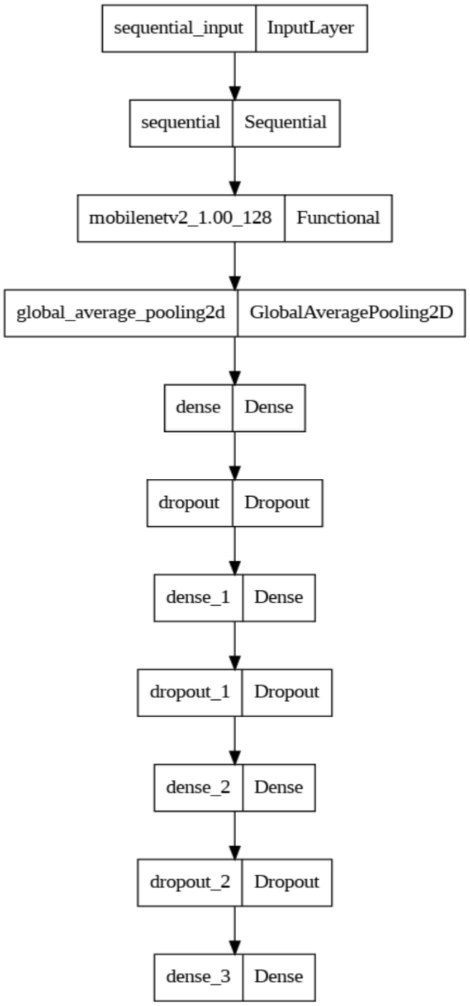


Figure 3. Architecture of the MobileNet-v2-Based Model [5]

MobileNet-v2, the model's base, was a 53-layer CNN that had been pretrained on the ImageNet dataset and recognizes the features of common objects and human faces [13]. It was modeled after a ResNet structure, which allowed CNNs that are both deep and effective to be built. Deep CNNs run the risk of the "vanishing gradient" problem: as the number of layers increases, the gradient passed to previous layers becomes much smaller, and the model is unable to alter its weights and biases. Typical ResNets contain residual blocks, which allow the model to skip over layers and increase in depth without risking a vanishing gradient [14]. As opposed to residual blocks containing two wide convolutional layers connected by a narrower layer, MobileNet-v2 contained inverted residuals connecting two narrower convolutional layers with a wide layer. The inverted residuals included fewer parameters than residual blocks [15], making the resulting model more efficient to train.

The data augmentation layer preceding the backbone randomly flipped the training and validation images horizontally, randomly rotated the images from -1.2566 to 1.2566 radians, and randomly zoomed into the images by a value of 0.2. This approximated different camera angles and focal lengths present in photographed faces. The Global Average Pooling layer after the backbone existed in lieu of several continuous convolutional and pooling layers. It averaged all the pixels in feature maps from MobileNet-v2 to create a flattened vector, preventing overfitting [16]. The dense layers' L2 kernel regularization used the squares of neurons' weights to add penalties to the layers' kernels and generalize the model [17]. Subsequent dropout layers essentially disabled 50% of all neurons to prevent the model from depending upon certain neurons to make a classification. Overall, the architecture of the Mobile-Net-v2-based model allowed for a neural network structure that is both deep and effective at recognizing AI-produced faces. Its safeguards against overfitting accounted for variety across the training and validation data.

Assessment of Initial Model

To establish a baseline accuracy on its new dataset, the model was retrained for 10 epochs on 1,000 images from the FFHQ dataset and 1,422 images from the "1M AI generated images 128x128" dataset with a train-test split of 0.2. 1932 images were used for training, and 484 images were used for validation. It obtained a training accuracy of 0.7260 and a validation accuracy of 0.7355, as shown in Figure 4. It misclassified faces with glasses, strong shadows, intense light conditions, or shut eyes, indicated in Figure 5. Incorrectly classified faces with glasses or closed eyes demonstrated that the model analyzed a face's eyes to determine whether it was AI-generated.

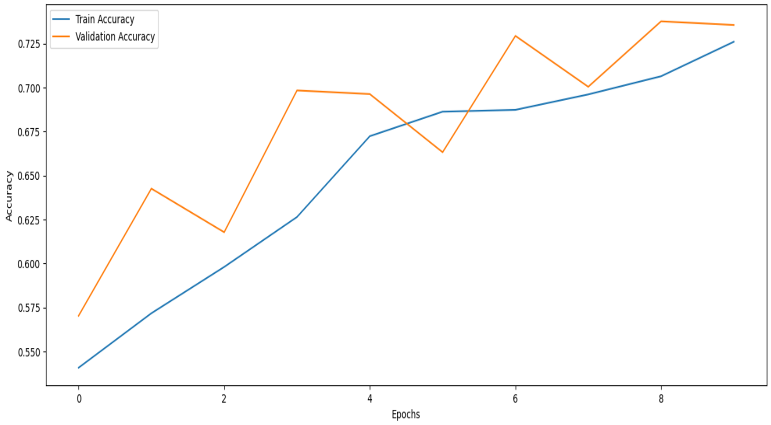


Figure 4. Initial MobileNet-v2-Based Model Accuracy Versus Epochs Trained

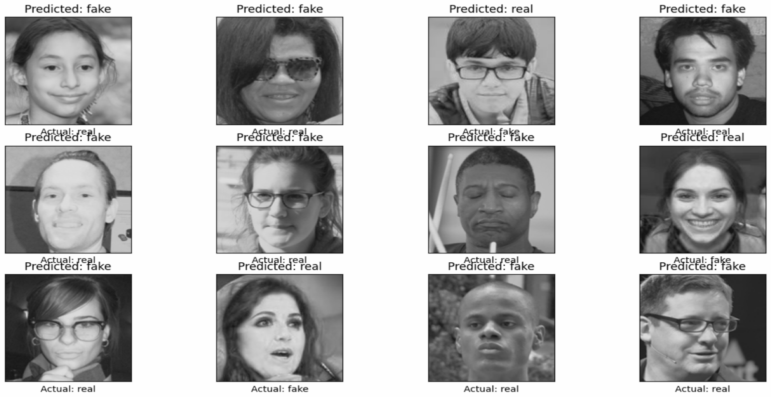
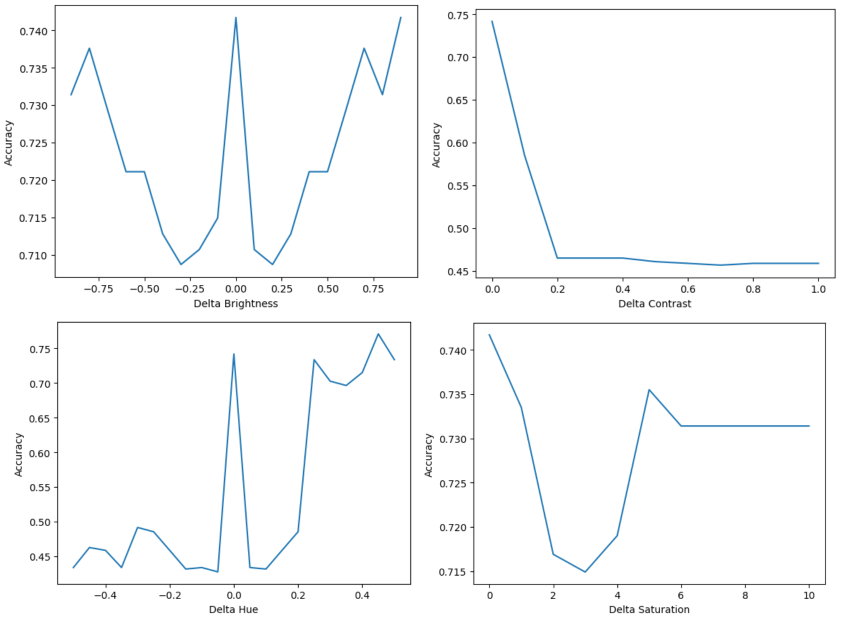


Figure 5. Incorrectly Classified Images

The model was then stress tested on augmented images to assess its approximate accuracy on poorly-taken photographs. Each face in the validation dataset was modified by changing its brightness, contrast, hue, or saturation by a set delta. Brightness was adjusted by deltas between and including -0.9 and 0.9. Contrast was incremented by values between and including 0.0 and 1.0. Hue was augmented by deltas between and including -0.50 and 0.50. Saturation was adjusted by values from 0.0 to 10.0.

In the majority of cases, the model performed poorer on augmented validation data than on non-augmented validation data. As shown in Figure 6, augmentations to delta brightness resulted in symmetrical trends in accuracy across the axis of delta = 0.0. Accuracy would increase as the absolute value of delta approached infinity and decrease as the absolute value of delta approached 0.0. As delta contrast increased, indicated in Figure 7, accuracy decreased. Accuracy decreased by the greatest magnitude from 0.7417 with a delta of 0.0 to 0.4649 with a delta of 0.2 and remained relatively constant for all greater deltas. For negative values of delta hue, as seen in Figure 8, accuracy decreased as the delta approached negative infinity and reached a maximum with a delta of 0.0. Accuracy then decreased to 0.4318 with a delta of 0.10 but increased to 0.7707 with a delta of 0.45. Delta saturation, depicted in Figure 9, initially caused a sharp decrease in accuracy from 0.7417 with a delta of 0.0 to 0.7149 with a delta of 3.0. Accuracy then increased to 0.7355 with a delta of 5.0 and remained constant at 0.7315 for all greater deltas.



Figures 6-9. Results of Stress Testing on Model Accuracy

Possible Limitations and Modifications

In brief, the model achieved a lower accuracy relative to what was obtained in the original project, and its accuracy decreased further on augmented data. The decrease compared to the initial model was caused by the limited amount of data the model was retrained on and the fact that the AI-generated images were produced by StyleGAN rather than DALL-E 2. Its universally poor performance on images with edited or extreme conditions was likely because training and validation images from the FFHQ dataset, the DALL-E 2-produced dataset, and the SyleGAN-produced dataset were fairly high-quality and well-lit. Its existing data augmentation accounted only for varying angles and focal length from the face rather than degraded quality. The model could have also been an overfit and too specific to its training dataset to accurately classify validation images.

One way to improve the model's performance would be by including additional data augmentation to the training data, such as randomly altering the brightness, contrast, hue, or saturation within the ranges used during stress testing. A Gaussian Noise layer with a set standard deviation could also be implemented to approximate a photograph with degraded quality. Other possible modifications to increase general accuracy include altering the model's architecture and how long it was trained. As 10 epochs may not have been long enough for the model to properly identify the features of an edited face, accuracy could improve if it were trained for a greater number of epochs. Additionally, as Global Average Pooling averaged all learned feature maps, the model may not have been able to identify features of distorted image data. Global Max Pooling, however, selects only the pixels with the highest values from all feature maps and could identify prominent features of images that are edited or poorly-lit. The dropout rate of the model could also be modified, as a rate of 0.5 could have disabled too many neurons or resulted in an overfit. Finally, a regulation strength that was too low potentially resulted in a model that was not generalized enough for augmented data, while a regularization strength that was too high could have caused the model to become too simple.

Implementation of Modifications

The finalized modifications to the initial model were augmenting the training data by randomly altering its brightness, contrast, hue, and saturation and implementing a Gaussian noise layer with varied standard deviation, replacing the Global Average Pooling Layer with a Global Max Pooling Layer, changing the number of epochs for which the model is trained, incrementally altering the model's dropout rate, and altering the model's regularization strength. Separate copies of the model's code were created for each modification so that only one could be put in place at a time. The model was retrained for every modification. Accuracy was measured across validation dataset images that had no augmentations, brightness altered by a random value within the bounds [-1.0, 1.0], contrast altered by a random value within the bounds [0.0, 1.0), hue altered by a random value within the bounds [-0.5, 0.5], and saturation altered by a random value within the bounds [0.0, 10.0). Only one augmentation was applied to the validation dataset at a time after the modified model was trained. Results were then compared to the accuracies obtained during stress testing.

First, the training data were augmented to mirror the alterations implemented during stress testing. By exposing the model to images with varying luminance and chrominance, it would ideally become more effective at classifying both augmented and non-augmented images. The brightness of each image in the training dataset was augmented by a random delta within the bounds [-1.0, 1.0]. The contrast of each image in the training dataset was augmented by a random delta within the bounds [0.0, 1.0). The hue of each image in the training dataset was augmented by a random delta within the bounds [-0.5, 0.5]. The saturation of each image in the training dataset was augmented by a random delta within the bounds [0.0, 10.0). All augmentations resulted in a decreased validation accuracy on non-augmented images. Moreover, no improvement from the baseline values, as seen in the "None" row of Table 1, was observed for altered images when training data was augmented. However, classification accuracy on data with edited contrast increased when training data was altered.

Table 1. Validation Accuracy As a Result of Augmentations to Training Data

	Augmentation on Validation Data				
Augmentation on Training Data	None	Brightness	Contrast	Hue	Saturation
None	0.7562	0.7562	0.4938	0.7500	0.4835
Brightness	0.7128	0.7169	0.6467	0.6260	0.6198

Contrast	0.5847	0.5847	0.5661	0.5682	0.5702
Hue	0.6983	0.6880	0.6281	0.6426	0.6198
Saturation	0.7087	0.7211	0.5930	0.6343	0.5909

Afterwards, a Gaussian noise layer was added to the model's existing data augmentation layer to replicate facial obstructions and poor image quality. Its standard deviation was set to values of 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, and 1.0. A trend did not appear to exist between standard deviation and validation accuracy. Universal increases occurred with standard deviations of 0.005, 0.1, and 0.5. Universal decreases were noted with standard deviations of 0.0005, 0.01, and 0.05. The model achieved its highest accuracy across all categories with a standard deviation of 0.1, as seen in Table 2. Validation accuracy was 0.7686 on unaltered images, 0.7603 on images with augmented brightness, 0.7066 on images with augmented contrast, 0.6591 on images with augmented hue, and 0.6364 on images with augmented saturation.

Table 2. Validation Accuracy As a Result of Gaussian Noise Standard Deviation

Standard Deviation	Augmentation to Validation Data				
	None	Brightness	Contrast	Hue	Saturation
0.0001	0.7397	0.7355	0.6550	0.6260	0.6446
0.0005	0.4773	0.5041	0.4793	0.4545	0.4814
0.0010	0.7417	0.7397	0.6219	0.6260	0.5579
0.0050	0.7479	0.7521	0.6343	0.6260	0.5785
0.0100	0.5682	0.5496	0.5413	0.5083	0.5372
0.0500	0.5413	0.5248	0.5537	0.5165	0.4814
0.1000	0.7686	0.7603	0.7066	0.6591	0.6364
0.5000	0.7459	0.7397	0.6674	0.6674	0.6302
1.0000	0.7376	0.7376	0.6302	0.6384	0.6157

A Global Max Pooling layer was then put in the place of the Global Average Pooling layer to recognize the most prominent features of heavily edited images. As a result, the model obtained an accuracy of 0.6343 on non-augmented validation data, 0.6384 on validation data with augmented brightness, 0.5930 on validation data with augmented contrast, 0.5888 on validation data with augmented hue, and 0.5847 on validation data with augmented saturation. This modification caused a marked decrease in accuracy across all validation images except for those with augmented contrast, as depicted in Table 3, where accuracy increased instead.

Table 3. Validation Accuracy As a Result of Implementing a Global Max Pooling Layer

Augmentation on Validation Data	Accuracy
None	0.6343
Brightness	0.6384
Contrast	0.5930
Hue	0.5888
Saturation	0.5847

The model was then retrained for 5, 10, 15, 20, 25, and 30 epochs to determine whether the model was an underfit or overfit when trained for 10 epochs. Accuracy across all categories was the greatest when the model was trained for 15 epochs, as shown in Table 4. The model obtained an accuracy of 0.7603 on non-augmented validation data, 0.7541 on validation data with augmented brightness, 0.6033 on validation data with augmented contrast, 0.6075 on validation data with augmented hue, and 0.5393 on validation data with augmented saturation. Accuracy began to decrease across all categories as the number of epochs trained increased.

Table 4. Validation Accuracy As a Result of Epochs Trained

	Augmentation on Validation Data				
Epochs	None	Brightness	Contrast	Hue	Saturation
5	0.7087	0.7087	0.6322	0.6281	0.6446
10	0.7562	0.7562	0.4938	0.7500	0.4835
15	0.7603	0.7541	0.6033	0.6074	0.5393
20	0.6405	0.5971	0.6012	0.5537	0.6178
25	0.5165	0.5310	0.4649	0.4628	0.4566
30	0.4566	0.4545	0.4421	0.4277	0.4421

Its dropout rate was altered but kept uniform across each layer to ascertain if a dropout rate of 0.5 was sufficient for the model to make adequate classifications on edited images. Rates of 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 were tested. In general, as the dropout rate increased, accuracy across all categories decreased. Accuracy peaked at a rate of 0.0, or no dropout layers, and at a rate of 0.5, or the original dropout rate. The greatest accuracy was achieved with no dropout layers, as depicted in Table 5. Accuracy on validation images with no augmentation was 0.7748, accuracy on validation images with augmented brightness was 0.7665, accuracy on validation images with augmented contrast was 0.6818, accuracy on validation images with augmented hue was 0.5682, and accuracy on validation images with augmented saturation was 0.5041. As the dropout rate was greater than or equal to 0.7, accuracy across all categories remained constant at 0.5661.

Table 5. Validation Accuracy As a Result of Augmentations to Training Data

	Augmentation on Testing Data (Maximum)				
Dropout Rate	None	Brightness	Contrast	Hue	Saturation
0.0	0.7748	0.7665	0.6818	0.5682	0.5041
0.1	0.5888	0.5620	0.5372	0.5103	0.4525
0.2	0.5310	0.4979	0.4917	0.5124	0.5041
0.3	0.5723	0.5124	0.5248	0.5537	0.5682
0.4	0.5289	0.5165	0.4814	0.5496	0.5062
0.5	0.7562	0.7562	0.4938	0.7500	0.4835
0.6	0.5620	0.5744	0.5145	0.5372	0.5269
0.7	0.5661	0.5661	0.5661	0.5661	0.5661
0.8	0.5661	0.5661	0.5661	0.5661	0.5661
0.9	0.5661	0.5661	0.5661	0.5661	0.5661

Finally, the L2 kernel regularization strength was changed to determine if a greater regularization strength would cause the model to be more generalized for edited images. Regularization strengths of 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5 were assessed. No alterations to regularization strength resulted in a universal increase in accuracy, as seen in Table 6, and a peak in accuracy only occurred with the original regularization strength of 0.001. As regularization strength increased, accuracy decreased for non-augmented validation data. It remained constant at 0.5661 across all validation data for all regularization strengths greater than or equal to 0.05.

Table 6. Validation Accuracy As a Result of Regularization Strength

Reg Strength	Augmentation to Validation Data				
	None	Brightness	Contrast	Hue	Saturation
0.0001	0.7004	0.6942	0.5991	0.6178	0.6012
0.0005	0.5826	0.5496	0.4979	0.4979	0.5248

0.0010	0.7562	0.7562	0.4938	0.7500	0.4835
0.0050	0.5807	0.5496	0.5785	0.5950	0.5826
0.0100	0.5826	0.5744	0.5909	0.5826	0.5640
0.0500	0.5661	0.5661	0.5661	0.5661	0.5661
0.1000	0.5661	0.5661	0.5661	0.5661	0.5661
0.5000	0.5661	0.5661	0.5661	0.5661	0.5661

Assessment of Modifications

Overall, the only modifications that resulted in an improvement to the model's validation accuracy were implementing a Gaussian noise layer with a standard deviation of 0.1, training the model for 15 epochs, and removing any dropout layers. The effectiveness of the Gaussian noise layer was most likely because it simulated facial obstructions, preventing the model from misclassifying faces with glasses or shadows. It also mimicked modifications to an image's brightness and contrast. The lack of a trend in accuracy as a result of Gaussian noise standard deviation was due to the random augmentation of validation data. The increase in accuracy at 15 epochs demonstrated that 10 epochs were not sufficient for the model to learn image features, particularly when the images were post-processed. With no dropout layers, the model had an adequate amount of neurons and connections to recognize features without becoming too simple.

Modifications that did not increase the model's validation accuracy either resulted in an overfit, caused the model to become too simple, or did not adequately account for augmented validation data. The lack of improvement when training data was augmented was because both training and testing images were altered by random values, thus the model struggled to make accurate classifications. The reduced validation accuracy on non-augmented data was due to the width of the bounds from which random augmentation deltas were selected, as non-augmented validation images did not possess conditions as extreme as those that may have appeared during training. The Global Max Pooling layer's limited effectiveness was because prominent features of images with altered brightness or saturation were not distinguishable, as all pixels approached similar values when extreme changes were made to images. However, improvement across validation images with augmented contrast was achieved because the most prominent pixels were emphasized by increases in contrast. The decrease in accuracy when the model was trained for greater than 15 epochs suggests that increasing the number of epochs trained leads to an overfit. Both increases to the model's dropout rate and regularization strength resulted in the model plateauing because it was too simple to recognize features.

Challenges and Future Work

Though the modified model achieved a validation accuracy greater than the baseline determined through stress testing, the majority of modifications resulted in a decrease or no change in validation accuracy. In many cases, no trend was present between a certain modification and the resulting accuracy. Moreover, the modified model did not achieve a validation accuracy of 0.80 or higher despite the original MobileNet-v2-based model obtaining an accuracy of 0.80. These issues were caused by either the modified model's limited amount of training and validation data or the random augmentation to the validation data. The model was retrained on only 1932 images and validated on 484 images, thus the relatively limited amount of training and validation data potentially hindered the model's accuracy and caused inconsistent results. If the experiments were repeated, the model would have been trained and validated on 21,000 images from FFHQ and 21,000 StyleGAN-generated images, similar to the initial model. Likewise, assessing all modifications against randomly augmented validation data may have resulted in inconsistent or misleading results, as each modification was essentially evaluated on different data. Images in the validation dataset should have been augmented by a set value in a manner similar to stress testing.

By far the most prominent factor preventing the model's accuracy from increasing to 0.80 or higher was the fact that only one modification was put in place at a time. A single type of data augmentation, layer, or value changed can only result in improvements up to a certain extent. Thus, to extend this research, the modifications most effective at improving validation accuracy should be implemented either all at once or in various combinations to assess whether this improves validation accuracy. This research could also be extended by exploring the MobileNet-v2-based-model's effectiveness at classifying images with different types of post-processing, such as by altering JPEG quality or adding a Gaussian blur. Furthermore, as the AI-generated images in all versions of the model were produced by only one generator, the model's accuracy on images from a different generator could be assessed. Alternatively, synthetic images from the training and validation datasets could have come from two different generators.

Conclusion

In summary, the Mobile-Net-v2-based model was the most accurate at recognizing both augmented and non-augmented AI-generated images when a Gaussian noise data augmentation layer with a standard deviation of 0.1 was added, it was trained for 15 epochs, and it contained no dropout layers. These modifications were effective because Gaussian noise resembled both augmented and natural shadows or obstructions on faces. Additionally, when the model was trained for longer or contained no dropout layers, it could learn image features for classification without becoming an overfit. The model's overall accuracy was hindered by the limited amount of training and validation data and the benchmark established by random augmentations to validation data. The fact that one modification was made to the model at a time likely limited its accuracy, hence future alterations could activate multiple modifications at once. This project emphasized that a CNN-based detection tool must be adequately complex to handle varying image data without being an overfit. However, training on augmented data does not necessarily guarantee improved classification accuracy on augmented and non-augmented images. The improvements caused by implementing a Gaussian noise layer proved that it could simulate conditions causing the model to misclassify images.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Rawnsley, A. (2020, July 6). Right-Wing Media Outlets Duped by a Middle East Propaganda Campaign. *The Daily Beast*. http://www.thedailybeast.com/right-wing-media-outlets-duped-by-a-middle-east-propaganda-campaign

[2] Allyn, B. (2022, March 16). Deepfake Video of Zelenskyy Could Be "Tip of the Iceberg" in Info War, Experts Warn. *NPR*. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

[3] Galer, S. S. (2023, January 27). *Someone Made AI Videos of "Americans" Backing a Military Coup in West Africa*. Vice. http://www.vice.com/en/article/v7vw3a/ai-generated-video-burkino-faso-coup

[4] Wang, X., Guo, H., Hu, S., Chang, M.-C., & Lyu, S. (2023, May 4). *GAN-generated Faces Detection: A Survey and New Perspectives*. ArXiv.org. https://doi.org/10.48550/arXiv.2202.07145

[5] Kabber, P., Lai, S., Oh, S. Y., & Wang, E. (2023). *Detecting AI-Generated Fake Images*. https://github.com/nogibjj/Detecting-AI-Generated-Fake-Images/blob/main/final_report/IDS705%20Final%20Report.pdf

[6] Lago, F., Pasquini, C., Böhme, R., Dumont, H., Goffaux, V., & Boato, G. (2022). More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner]. *IEEE Signal Processing Magazine*, *39*(1), 109–116. https://doi.org/10.1109/msp.2021.3120982

[7] Lu, Z., Huang, D., Bai, L., Liu, X., Qu, J., & Ouyang, W. (2023, April 25). *Seeing is not always believing: A Quantitative Study on Human Perception of AI-Generated Images*. ArXiv.org. https://doi.org/10.48550/arXiv.2304.13023

[8] Bird, J. J., & Lotfi, A. (2023). *CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images*. https://doi.org/10.48550/arxiv.2303.14126

[9] Hulzebosch, N., Ibrahimi, S., & Worring, M. (2020, June 1). *Detecting CNN-Generated Facial Images in Real-World Scenarios*. IEEE Xplore. https://doi.org/10.1109/CVPRW50498.2020.00329

[10] Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., & Wang, Y. (2023). *GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image*. https://doi.org/10.48550/arxiv.2306.08571

[11] Karras, T., Laine, S., & Aila, T. (2018). *A Style-Based Generator Architecture for Generative Adversarial Networks*. ArXiv.org. https://arxiv.org/abs/1812.04948

[12] Dullaz. (2021). *1M AI generated faces 128x128*. Kaggle. https://www.kaggle.com/datasets/dullaz/1m-ai-generated-faces-128x128

[13] *MobileNet-v2 convolutional neural network - MATLAB mobilenetv2*. (n.d.). MathWorks. https://www.mathworks.com/help/deeplearning/ref/mobilenetv2.html

[14] *PyTorch ResNet*. (n.d.). Run:ai. https://www.run.ai/guides/deep-learning-for-computer-vision/pytorch-resnet

[15] Pröve, P.-L. (2018, April 11). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. Medium. https://medium.com/m/global-identity-

2?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fmobilenetv2-inverted-residuals-and-linear-bottlenecks-8a4362f4ffd5

[16] Olu-Ipinlaye, O. (2022, September 30). *Global Pooling in Convolutional Neural Networks*. Paperspace Blog. https://blog.paperspace.com/global-pooling-in-convolutional-neural-networks/

[17] Keras Team. (n.d.). *Keras documentation: Layer weight regularizers*. Keras. https://keras.io/api/layers/regularizers/