IDS 705 Final Project Report

Classification of Fake and Real Faces

And Its Implications for Dating Apps

Team Members:
Kashaf Ali
Beibei Du
Eric Rios Soderman
[in alphabetical order by last name]

Team Number: 9

Abstract

This paper explores the problem of fake profiles on social media, which are often used to deceive and scam users. We specifically focus on the use of AI-generated photos in creating these fake profiles. While some individuals may use fake profiles innocently due to disinterest or shyness, others may have malicious intentions. Our goal is to evaluate the viability and implications of using a ResNet50 model to identify fake facial images and evaluate the limitations of such models. In addition to the stated goals, this paper also investigates the performance of a grayscale image model and compares it to the ResNet50 model for identifying fake facial images. The evaluation will provide insights into the effectiveness of the two models and which one is more suitable for detecting fake facial images. To evaluate the performance of the models, we test them on our dating application spaces to identify potential limitations of our model in the real world world. The RGB ResNet model had an accuracy of 95.39%, and the grayscale one achieved an accuracy of 93.78% on the test set. As such, we learned that there may not be much of an advantage of using grayscale models, although it did reduce False Positives, which gives potential hope for further exploration as to why that may be the case, as it opens the door for questions about lighting and backgrounds. In terms of the application dataset, the model had a low accuracy of 65%. The application set contained images, intentional by design, to fool the model, which were curated by us. This showed that the model is only as good as its data, and it cannot account for images on which it was not trained, such as body shots and tilted faces or side profiles of them. In its current state, this model cannot handle the sheer variety of pictures seen in dating sites unless they are face shots. In terms of bias of the study, although we were not able to understand the underlying distribution of the training data, we were able to glean some distribution of the misclassifications, and White faces were especially prominent in the misclassifications. Lastly, as this space rapidly evolves and as fake imagery becomes ever more convincing, we have learned that very robust and creative modeling approaches will be required to handle the myriad of fake imagery present in the dating sphere, especially the advancements in the area of fake faces like the ones found in CivitAi, as they so effectively fooled our model.

1. Introduction

Fake profiles are a widespread issue in social media, particularly in online dating apps where the intention is to create connections with others and portray a certain persona or image online [1]. Increasing prevalence of Fake profiles is also highlighted in a literature review by Gyan Vihar University researchers, who point out that fake profiles are alarmingly common in online dating and various scams [25, 26]. Moreover, the Federal Trade Commission also reported $547 million record loss from online romance scams in 2021, with the 70 and older age group being affected the most with the highest individual median loss of $9,000 [20].

Social media profiles, in general, serve as templates for conveying user information and tastes, often through pictures. However, with the rise of AI-generated photos, or "deepfakes," concerns about authenticity and deception have increased as users may use these images in their profiles [11]. These false depictions can be accompanied by fake information and pictures, posing significant risks to users who want to use these apps to develop meaningful connections with others. Even now, certain individuals generate fake images for phishing scams, catfishing, and other criminal activities [11]. It is important to realize the creation of fake dating profiles is certainly a significant issue. Thus, we need to recognize that AI-generated (GAN) photos have broader implications that extend beyond their use in dating apps. In this case, researchers need to explore these broader implications in order to fully understand the impact of this technology on our society, including issues such as privacy and bias.

The problem that we are trying to answer is strictly academic, and we have no plans on making a commercial tool from this, as our dataset is not intended for and should not be used for development or improvement of facial recognition technologies [14]. The prevalence of fake profiles and AI-generated photos in social media has raised concerns about authenticity, particularly in dating apps. Therefore, researchers should investigate implications and develop strategies to detect and prevent fake profiles to promote transparency and trust in online interactions in social medias like dating apps [1]. We will implement machine learning techniques, such as image classification using ResNet50 and K-Nearest Neighbors (KNN) classifiers, to train real and fake facial image data, in order to develop a tool that classifies real and fake faces. Our goal is to test our fake and real facial image classification tool on a real and fake image dataset, which will serve as a proxy for the kind of images that are used in a dating profile or other social media applications. Furthermore, we also aim to evaluate the viability and the implications of our ResNet50 model in our application space, compared to our baseline KNN model and a grayscale version of the ResNet50 model as well. In terms of implications, we mean to verify what images our model will correctly and incorrectly classify, in order to attempt to find salient patterns in the predictions, if any exist. With these tools and considerations for our problem space, our objective is to potentially use our model for classifying fake images, and consequently, identifying fake users on dating apps. Additionally, we also hope to understand the limitations of these models in terms of gauging their effectiveness for helping to identify fake profiles.

Further expanding on the limitations, beyond general effectiveness in discriminating real from fake faces, we also consider other challenges inherent to this study. It is limited by the ResNet50 model's restrictions in transfer learning capability, requiring additional fine-tuning for optimal performance in our face classification problem. It is also important to highlight that our model's interpretability is also limited, making it difficult to fully comprehend the reasoning behind certain image classifications solely through quantitative insights, given the nature of neural network models. Hence, we aim to supplement our understanding about these misclassifications through a qualitative analysis as well. Moreover, while we tested our model on a proxy dataset, it is critical to note that the model's

effectiveness with novel types of fake images is uncertain, and its application might be limited to our specific problem space. Ultimately, our goal is to develop a model that can aid strategies for detecting and preventing the use of fake images in dating profiles, while minimizing the potential harm to real users.

1.1 Goal/Objective

The Main Goal: We aim to develop a fake facial image classification tool that aids us in classifying real and fake images, specifically in the context of dating apps, and a model that can potentially be used by these dating app companies to prevent the use of fake images on their platform. Hence, our goal is to also assess the viability of this model in our application space, as well as identify potential biases in the model's classifications based on gender, race, or other factors. Our main research questions, shown below, are related to the accuracy, fairness, and practicality of this fake facial image classification tool.

Main Research Questions:

1. What is the viability of applying our image classification model for fake and real face detection for a dating app company?

2. Will the model that we train misclassify faces due to some form of gendered or racial bias and/or issues with faces that do not have enough facial proximity or carry other types of features (such as wearing glasses)?

3. What are the limitations of both the model selection and the data of the application space?

1.2 Background

In the current world of rapid progress in Artificial Intelligence (AI), the automated generation of poems, lyrics, music, and images has become increasingly popular in recent years. While newly developed and advanced technologies can bring convenience to human lives, there are many drawbacks, especially when ethical concerns come into play. Many researchers have delved into the topic of deep fakes, which we are also interested in, to differentiate between AI-generated faces and actual human faces [2, 3, 4, 5]. This is particularly important in a social context, as many people use fake faces as their profile pictures on dating apps, and fraud can sometimes occur.

Rosa et al. conducted a study on fake faces in the context of Tinder and found that appearance is a significant factor in the swiping decision-making process, along with judgments of moral character may be even more critical [1]. In other words, physical attractiveness serves as a cue for rapid judgments. Thus, the profile picture and personal photos posted on the Tinder profile are decisive, and from a social psychology perspective, they tend to correlate with the cognition of morals. The emergence of fake faces might distort the original intention of dating apps, resulting in harmful social interactions. To solve the potential problem of deception with fake profiles in dating apps, we can directly assess someone's facial features from their dating profiles to make more informed decisions in the swiping process. By accurately classifying images as real or fake, we can potentially increase the success rate of love connections and decrease the prevalence of deception on dating apps. Furthermore, this can ultimately lead to a more positive experience for users and a higher level of trust in the platform.

This leads us to the question: How can we differentiate between real and fake facial images in order to prevent any potential detrimental influence on dating apps and online social interactions, or even the illegal theft of facial privacy? Luckily, there are many scientists who have tried different algorithms to predict whether an image depicts a real face or not in order to seek ways to prevent ethical violations of fake face usage online. For instance, Taeb et al. mentions several deep learning models for detecting deepfake faces in his paper, such as CNNs, RNNs, and GAN-based models [2]. However, in an effort to achieve higher classification accuracy, they decided to implement a custom CNN architecture (VGG-19 model) and DenseNet architecture, which achieves an accuracy of 95% eventually. Although these complex models may offer great results, we wanted to assess if simpler models are also able to offer the same performance.

In past research, various researchers have tackled the problem of deepfake detection using different deep learning models, and they have achieved relatively high accuracy. However, as AI technology continues to advance rapidly, the effectiveness of these algorithms may decline and become outdated by referring to the huge amounts of past research on this topic [4].

In this study, we will train a deep learning model using the ResNet50 architecture to classify fake and real faces. We will evaluate the performance of our model using a held out test set, as well as applying the model to a proxy dataset of images one would likely see in a dating profile. By doing so, we aim to assess the accuracy and effectiveness of our model in detecting fake and real images on dating apps.

2. Methodology

2.1 Data

We combined two datasets, the Flickr dataset for real images [13] and the 1 million fake faces dataset for fake images [15], described below, to construct our training, validation and testing datasets. Combining roughly 70k real and fake images from each of these two datasets allowed us to create a complete dataset of almost 140k images with equal representation of fake and real images. While we provide a detailed description of the data preprocessing and merging steps in our Data Merging and Cleaning subsection, it is important to note that we used this combined dataset to train and test our initial model, and then later also used it to create a grayscale version of the same data for the same ResNet model to understand how our model will perform if we remove the lighting conditions and exposure effects from our images. Since we have the colored and grayscale data as well as the same model architecture for each of them, we will compare results in our model results subsections to see if there are any meaningful differences. It is also important to note that we also used the combined and colored dataset, with the same training, validation and testing splits, to train a K-Nearest Neighbors, which we will mainly use as a benchmark against our ResNet50 model. As will be noted later on, there was no need to extend this application of the KNN for the grayscale combined dataset because the training for the KNN was only carried out on a subset of our dataset due to computational constraints. Further details regarding all three models will be covered in our Models Section.

We separated a test dataset from our original combined dataset, as discussed above. Testing on this dataset will provide us with a baseline to compare results from our other application-relevant test datasets.

1. **Real Images from the Flickr FacesHQ Dataset** [13]

This dataset contains 70,000 high-quality PNG images of human faces at 1024 x 1024 resolution. It was collected by NVIDIA from Flickr and thus, inherits all the biases of that website. According to NVIDIA's study on this dataset, we can surmise (but not quantitatively verify) that this dataset includes considerable variation in terms of age, ethnicity and image background, and also has substantial coverage of accessories such as eyeglasses, sunglasses, hats, etc [14]. The images were already aligned and cropped using dlib, automatic filters were used to prune the data, and Amazon Mechanical Turk was used to exclude paintings and statues from these images to make them ready for use as a training set for facial images.

This dataset is relevant to our project objectives as our aim is to design a model that is able to effectively detect fake images on dating apps and social media profiles, and many of those profiles have similar images of human faces with accessories. Moreover, the dataset is large enough to be used for real images in our training, validation and initial testing stage.

2. **Fake Images from 1 million fake faces [15]**

This dataset contains 1 million images of fake human faces that have been artificially generated using a variety of Generative Adversarial Network models and it was created by Alexander Reben, who is an artist and a MIT-trained roboticist at V7labs. Moreover, each image is of size 1024 x 1024 pixels, and is in the JPG format. As this dataset used the same StyleGAN algorithm and model by NVIDIA to create fake images from NVIDIA's Flickr FacesHQ Dataset, it also exhibits considerable variation in terms of age, ethnicity and image background, and also has faces with accessories such as eyeglasses, sunglasses, hats, etc [14]. NVIDIA's StyleGan algorithm, trained on the NVIDIA's real images dataset scrapped from Flickr, generated this particular dataset of fake faces. We are fully aware that we are making a somewhat strong assumption with the following statement : both the Flickr and the 1 million fake faces are somewhat similar in terms of the types of faces observed in terms of race, headwear, general accessories, and gender. We are not claiming that they have same 1-to-1 mapping of the distributions (i.e. 10,000 Asian faces in the Flickr dataset versus 10,000 Asian faces in the fake faces dataset), but we are assuming that there are similarities (i.e. 10,000 Asian faces in Flickr and maybe 8,000 in the fake faces dataset). Finally, more than anything, we want the model to essentially distinguish a real from a fake face, so we have chosen to combine both datasets for our modeling.

3. **Dataset for the application space**

AI-generated images are a relatively new and novel concept on social media, which makes them intriguing to users online. Users mostly post these images as art or under the disguise of real images on their social media profiles, including dating apps, and we want to extrapolate our model and test it against the best state of the art fake imagery [21]. We used a rendition of a typical dating app, meant to highlight the model's weaknesses in dealing with highly realistic AI-generated faces and faces with fake backgrounds. We used our results to develop possible metrics from our model that actual dating apps companies can use for detecting fake user profiles. We believe this application testing is critical for operationalising our model in the real world.

Considering that there exist various legal and ethical concerns regarding scraping and using profile images from dating apps, we decided to manually curate a combined dataset of different kinds of images, which will act as proxies for the type of images found on dating apps and social media

profiles [27]. It is important to note that social media contains a variety of images, and while our dataset may cover some of these use cases, they are not mutually exclusive and completely exhaustive of the kind of images present in our application space. Although these datasets contain similar images that one may find on a dating or social media profile, these are free to use for our model and without any ethical or legal constraints.

As our goal is to contextualize our results for our application space, we hope to use these datasets to understand how and why our model results will perform better or worse when provided with different kinds of image datasets available within our application space, such as body shots, far away pictures of the face, the introduction of unseen accessories to the face. To reiterate, our goal here is not to achieve complete or high test accuracy on these datasets, but to actually initiate a more academic pursuit of understanding how well our facial classification tool will work with new types of data, or in other words, we hope to understand the practicality and generalisability of future facial classification tools through these insights.

We manually collated this combined dataset of proxy dating app images, and it consisted of the following types of images:

- **High resolution AI generated images**: these were manually selected from Civitai [17], a technology company that specializes in computer vision and machine learning, and features high resolution fake AI-generated images of human faces, which predominantly includes images of Asian women. Although there is a lack of diversity in terms of both gender and race, with more images of women than men and more images of Asians than other races, we believe that this dataset still provides valuable insights into the performance of our model on these particular types of images.

- **Proxy Dating Profile Dataset:** We used our own group members' "potential" dating app pictures. Since we created this dataset, we ensured that this dataset contains images from different angles and accessories. This also includes full body images with backgrounds, and non-facial images to understand how our model will perform on those image types.

2.2 Data Merging and Cleaning

We sampled 140,000 images with 70,000 images each from the Flickr FacesHQ Dataset and the 1 million fake faces Dataset to be used in the training, validation, and initial testing stages of our project, where the first dataset had the real images, and the second had fake ones. As mentioned previously for the latter, we cannot quantify if the representation of races, sexes, people with accessories are proportional to the ones seen in the Flickr dataset, but we do hope that it is similar enough for our modeling purposes. We decided to use 70,000 images from each dataset because our real images dataset had 70,000 images in total. However, in practice due to image corruption present in the data, our real faces dataset numbered 69,999, and our fake faces dataset numbered 69,997, both totaling to 139,996 images. Additionally, we had computational power constraints due to which a dataset equally as big or bigger than this was difficult to train on our systems, taxing cloud and local resources.

It is important to note that while we randomly picked our training, test and validation sets, our aim was to ensure that the split was stratified with roughly the same number of Real and Fake images. Below are our training, test and validation splits:

Real-Fake	Training	Validation	Test
Fake	50,397	5,600	14,000
Real	50,399	5,600	14,000
Total	**100,796**	**11,200**	**28,000**

Table 1. Training, Validation and Test split

Image Preprocessing

We also carried out some additional image preprocessing to ensure that all our image files in the combined dataset were standardized and ready for the modeling stage. We carried out two kinds of preprocessing on our image data:

- Standardizing the image type: To maintain consistency, each image in the form of JPG or PNG was converted to PNG.
- Set the same resolution for all images: To ensure consistency in our data throughout the training, testing, and application process, we decided to standardize the size of our images to a resolution of 256 by 256 from a higher resolution of 1024 x 1024..

This same image preprocessing pipeline was also applied to all images in the additional testing dataset for the application space.

2.3 Exploratory Data Analysis

We conducted the following exploratory data analysis to understand what our dataset looks like and how we can use it in our model.

1. Real Faces Sample is displayed below (12 samples):

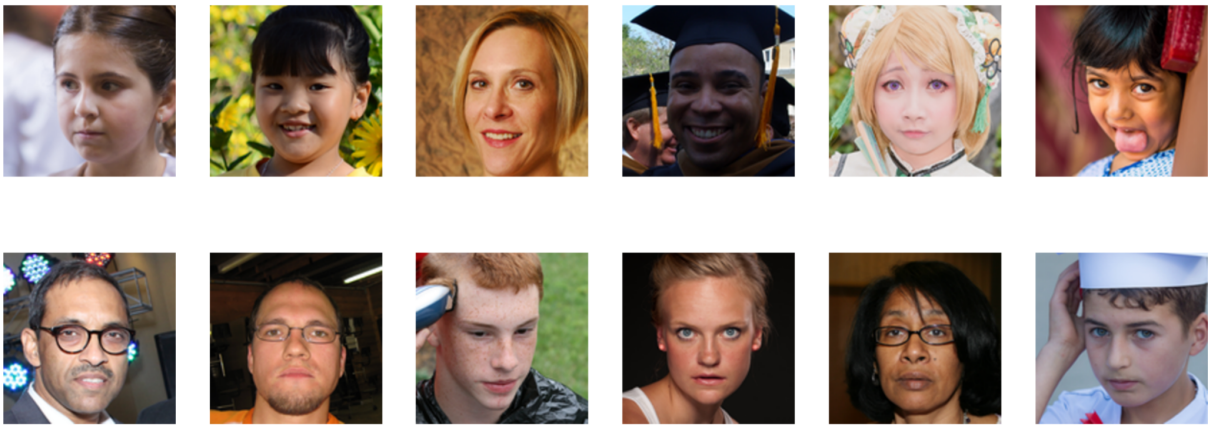

Figure 2. Randomly selected 12 samples from the True faces dataset

2. Fake Faces Sample is displayed below (12 samples)

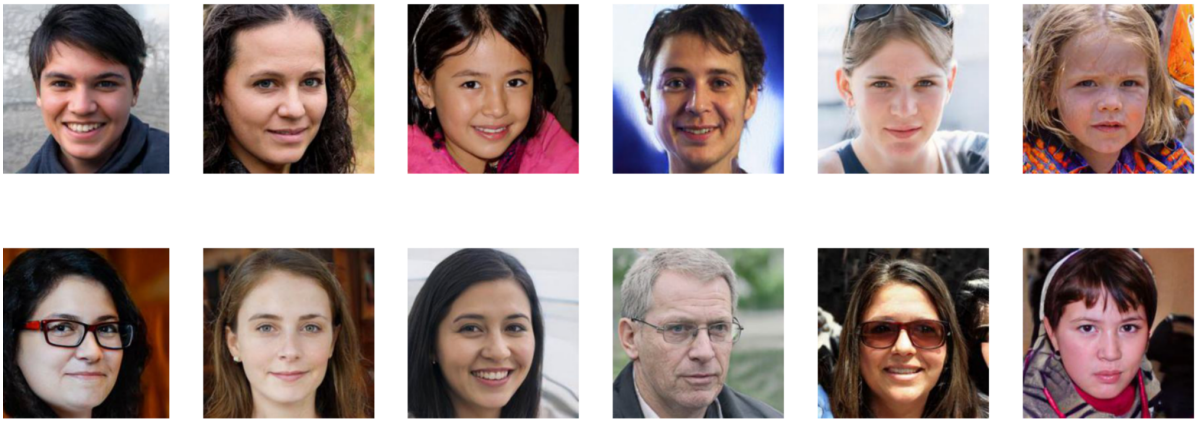

Figure 3. Randomly selected 12 samples from the Fake faces dataset

Although the research conducted by NVIDIA verifies that our datasets have considerable variation in terms of age group, races and gender, and the Generative Adversarial Network models used to create the fake images from these real ones also ensures that both real and fake images are similar in terms of image style, objects, backgrounds and accessories [14], we decided to conduct further data exploration to learn more about these images.

While we conducted a limited, manual review of the real and fake facial images in our dataset, we discovered that the diversity in these datasets can be quite fair at first glance. Additionally, we were not able to find any unusual rotations in the sample images (the dataset was also processed to ensure that [14]). As we believe or theorize that AI-generated fake faces are more likely to present some "errors" that can be detected through some deviation in the pixel values, we were also motivated to conduct a data visualization analysis on RGB pixel values to further explore this. This will provide us with valuable insights into the differences between real and fake faces and guide us to determine the efficacy of pixel values as a detection tool.

By analyzing the mean 3-channel RGB values of each face in the sample data, we initially observed that the real faces occupy a distinct plane compared to the fake faces according to Figure 4. This finding suggests that the differences in mean 3-channel RGB between the two classes, real and fake faces, can potentially be used to distinguish between them. From the plots below, it is evident that the mean pixel values for both real and fake images are sloping upwards, indicating that the mean pixel value increases across the pixel value range. However, for the fake images, even though the mean pixel values are sloping upwards, there are less values in the center, suggesting that the images in the dataset might have a higher mean pixel value for the darker pixels and a lower mean pixel value for the brighter pixels. This is possible in cases where images are underexposed or if the lighting conditions were not consistent across the images in the dataset.

In order to test the validity of our assumption, we did a two-sample t-test on each of the color channels and compared their distributions, as shown in Figure 5. The statistical analysis (two sample t-test), conducted on the repeated sampling of 1000 images for the mean and median color channels, revealed that none of the color channels (red, green, and blue) show a statistically significant difference based on the results obtained. This implies that the mean and median color channel values are consistent and stable, and there is no significant variation or deviation in the color channels across the real and fake image samples. Therefore, it can be inferred that there is no significant difference between rgb values

across the real and fake facial images. However, we are still left wondering how the sampling of 30 images could result in such a drastically different scatter plot on mean RGB values. This has piqued our curiosity, prompting us to explore how a grayscale model, where we remove the lighting and exposure from our images, would differ from the RGB-based model. By conducting this additional analysis, we hope to gain a better understanding of the underlying factors contributing to the variations in the data and further strengthen our conclusions. We will discuss and compare results of both models in our model results subsection.

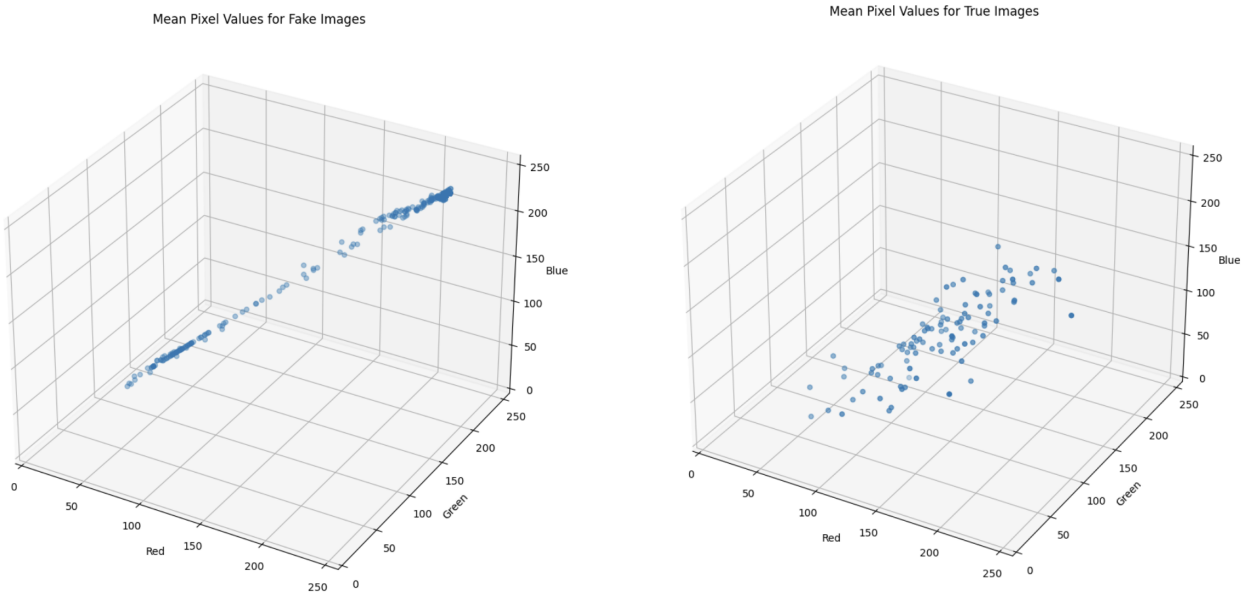

Figure 4. A 3D scatterplot has been plotted based on 30 random samples

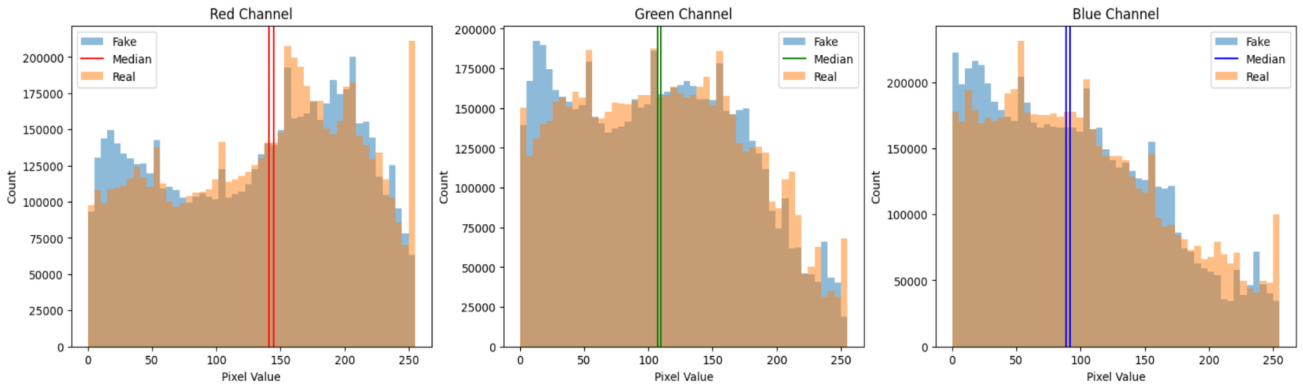

Figure 5. Mean RGB 3 channel Visualization on Real and Fake Faces
(Vertical lines indicate the median rgb channel values)

3. Models

We used a ResNet50 model for our classification problem and a K-Nearest Neighbors classifier as our baseline model to understand our ResNet50 model's performance. Moreover, we employ the same ResNet50 models in two different ways, by training on the RGB dataset and a grayscale version of that dataset, which affords us two variants of the same model architecture and configuration.

3.1 ResNet50

ResNet50 is a deep neural network architecture that is widely used for image recognition and classification tasks. It was introduced by researchers at Microsoft in 2015 [24], and it has since become a standard model in the field of computer vision [19]. We decided to use ResNet50 for our problem because it is a good starting point for image classification problems and offers simpler alternatives to more complex models like ViT and GAN. In addition to offering better performance on image classification tasks, its architecture allows it to also handle complex visual patterns, which is particularly important to us given that the faces in these datasets have headwear, glasses, different skin tones, hair styles, eye colors, nose shapes, facial hair, amongst many others. Additionally, ResNet50 was specifically designed to mitigate the vanishing gradients problem, which is a common issue in training very deep neural networks, where the learning of the weights is minimal. It does so by using residual connections, which allow the model to learn and propagate information more effectively through very deep networks [18].

The original ResNet34 consisted of 34 weighted layers, an efficient technique to add more convolutional layers to a CNN, without encountering the vanishing gradient problem. The ResNet architecture is based on two core design rules: the number of filters per layer is equal, but it depends on the size of the output feature map. The second is that, if the feature map's size is halved, it has double the number of filters to maintain the time complexity of each layer. ResNet50 iterates upon this existing architecture and implements a bottleneck residual block with 1×1 convolutions to effectively reduce the number of parameters and matrix multiplications to speed up training per layer.

As shown in this table, the model starts with the pre-trained ResNet50 model, using the weights from "imagenet". These weights were trained on the data from the ImageNet Large Scale Visual Recognition Challenge, where 3 of the synsets the images contain faces. Therefore, based on this finding, although ImageNet is an object detection problem, we think that the ImageNet weights can serve as a starting point that can be fine-tuned through ResNet50 for our classification purpose, since those weights already apply to a wide range of visual recognition tasks [22]. In fact, several researchers have used ResNet50 for similar facial image classification tasks, such as detecting fake face images or facial expression recognition. For example, in the paper "Deep Learning for Face Anti-Spoofing: A Survey," the authors used ResNet50 to classify real and fake face images for the task of face anti-spoofing [19].

The classic model has a 50-layer convolutional neural network with 48 convolutional layers, one MaxPool layer and one average pool layer. On top of those elements, we opted to also flatten those outputs (known to be 2048 based on the architecture) and add a fully connected layer after that max pool layer. That dense layer received those outputs into its 512 nodes with a ReLu activation function. Then the results of those nodes are fed into one final, output node to fit the binary classification of our problem, where there is a sigmoid activation function. The score that comes from this output is the

confidence score for the positive class. Lastly, the final and total number of trainable parameters in the model is 1,049,601.

Sequential Form	**Output dimensions**
ResNet50	2048 elements - Final Pooling Layer
Flattened	2048
Dense Layer	512 with ReLu
Dense Layer	1 - Final Output, passed through Sigmoid Activation

Table 2. ResNet50 Model Architecture

3.2 Grayscale ResNet50 Model

Our Grayscale ResNet50 model is, for all intents and purposes, identical in terms of configuration to our rgb-based ResNet50 convolutional neural network. Its adaptation is minor in the sense that the images that were fed to the model were preprocessed into grayscale, so the difference between them is the data on which they were trained. In the context of a fake facial image classification problem, this model, like its RGB counterpart, can be used to classify images as either fake or real based on their visual features. In the case of grayscale images, the input channels are reduced to a single channel, as grayscale images have only one color channel. In addition to this, grayscale models are also more robust to noise and other artifacts present in the images, as it is more reliant on the overall visual structure of the image rather than the specific color values, and hence, can useful in cases where image quality is low or there are challenging lighting conditions. Just like the RGB model, the weights are the ImageNet ones, meant to be used as a starting framework for fine tuning, and we are curious to see if removing color information from the features can assist in predictive performance.

Overall, this model can be an effective tool for fake image classification as it learns high-level visual features from a large pre-trained dataset and adapts these features to the specific problem space. It is also useful in contexts where color information is not as relevant or necessary for the problem. For fake facial image classification, color information might not be a significant factor in classifying whether an image is fake or real because many fake images are generated using digital manipulation techniques that can introduce artifacts and inconsistencies that are visible in grayscale [28]. In addition to this, grayscale models are also more robust to noise and other artifacts present in the images, as it is more reliant on the overall visual structure of the image rather than the specific color values, and hence, can useful in cases where image quality is low or there are challenging lighting conditions.

3.3 K-Nearest Neighbors

In our research, we hypothesize that the RGB values of the two groups have some potential distinctions from each other. To investigate this hypothesis, we plan to use the K-Nearest Neighbors (KNN) algorithm [23] to calculate the closest distance to each group and classify the data. KNN is an excellent starting point for our analysis, as it is a simple and intuitive algorithm that can provide a baseline model for our research. By using KNN, we can test the effectiveness of our hypothesis and determine if there are any discernible patterns in the data that can help us to classify the two groups

accurately. Overall, our initial approach using KNN will be an essential step in our analysis and help us to better understand the data.

K-Nearest Neighbors (KNN) is a simple and effective classification algorithm that can serve as a baseline model for comparison with more complex models like ResNet50. By comparing the accuracy of KNN with that of ResNet50, we can gain insight into how much additional benefit is provided by using a more complex model. An advantage of KNN is that it can be easily adapted to handle various types of input data, including image data. In image classification tasks, KNN can be used to classify images based on their pixel values, without the need for complicated feature extraction or dimensionality reduction techniques. In this specific case, we used the same preprocessed split for training, validation and initial testing for our KNN model. Lastly, KNN can be effective in situations where the decision boundary between different classes is highly nonlinear and complex, as it can capture local structures and dependencies in the data that may be missed by more global models.

4. Results

In this results section, we will evaluate the performance of our classification model using various metrics such as accuracy, ROC (Receiver Operating Characteristic) curve, precision-recall (PR) curve, and misclassified images. These metrics will provide us with an overall understanding of the model's performance and highlight areas that need improvement. In addition to analyzing the model's performance, we will also discuss the limitations of the model, which may affect the generalization of the results to new datasets or applications.

4.1 K-Nearest Neighbors Results

Results from our KNN classifier showed that the model had an accuracy of 50.13%, and it did not perform well in classifying real and fake images. As depicted in the ROC curve below, the model has an AUC of 0.58, which indicates that the model's overall performance is slightly better than random guessing, but not very good, and there is a 58% chance that the model will rank a randomly chosen positive example higher than a randomly chosen negative example. Moreover, the Precision-Recall curve suggests that the model has an average precision of 0.56, indicating that it only has a moderate ability to correctly classify positive examples, and there is still room for improvement.

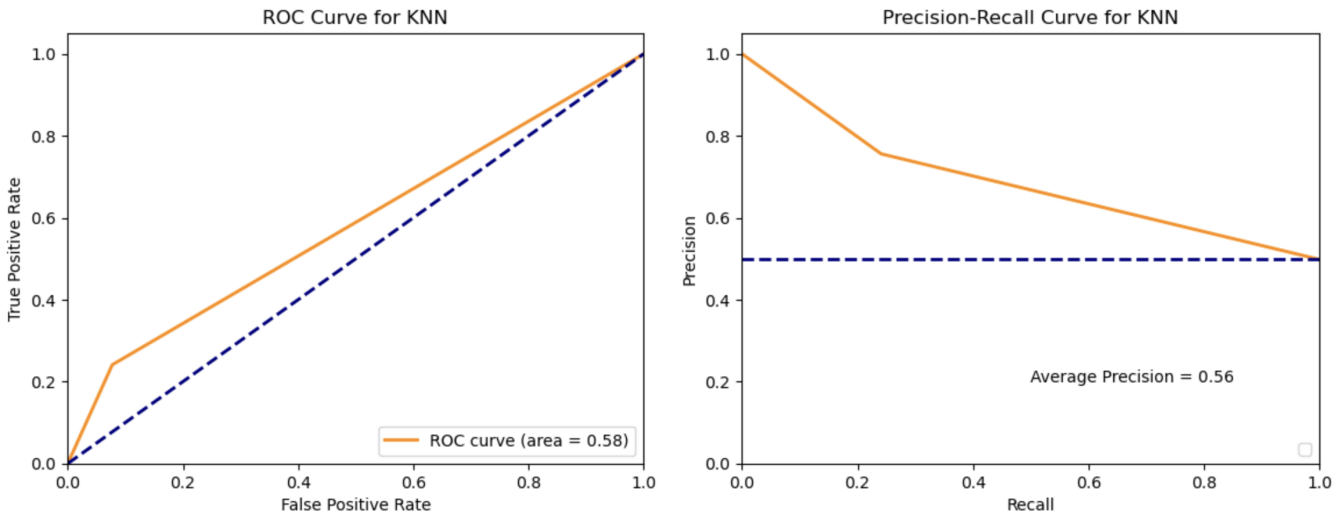

Figure 7. ROC and Precision-Recall Curves for KNN

Although our KNN model served as a simple baseline model to compare with our ResNet model, we were only able to run it on a sample of 20k images from our complete dataset. It was difficult to run our KNN model on a dataset larger than that because of the high computational cost and memory usage associated with it. It is important to note that KNN is a lazy learning algorithm that stores all the training data in memory, which means that as the size of the dataset increases, the memory required to store the data also increases. In addition to this, the time complexity of KNN increases linearly with the size of the dataset, making it impractical for large datasets. As a result, we decided to run our KNN on a subset of our data to substantiate that the ResNet50 model was a better choice for our problem case, as similar subsampling techniques were also suggested in the literature. Hence, we then redirected our time and computational resources to tuning our ResNet50 model.

4.2 RGB-based ResNet50 Model Results

Results from our RGB-based ResNet50 Model show an accuracy score of 95.39%, indicating that the ResNet50 model has excellent discrimination ability and is capable of correctly classifying fake and real images. Moreover, the ROC curve shows an AUC of 0.9921 which means that the model has a low false positive rate of around 0.008. Additionally the Precision-Recall curve also shows an average precision score of 0.9928, indicating that the model has a high precision and recall, and it can classify the positive samples with high accuracy while minimizing false positives.

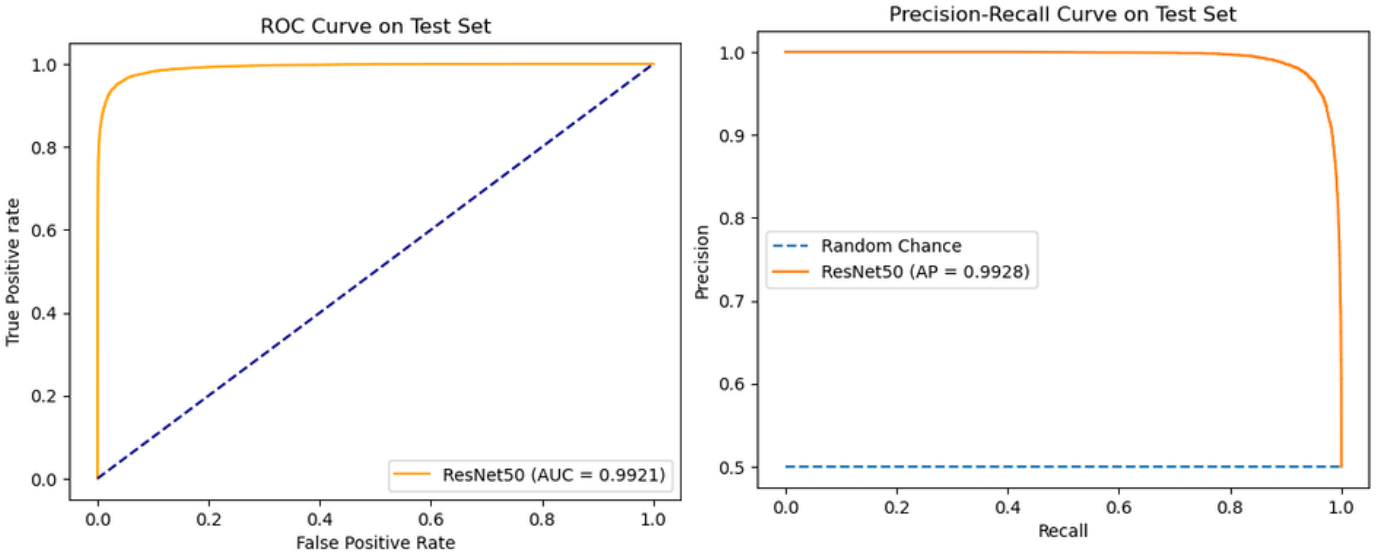

Figure 8. ROC and Precision-Recall Curves for RGB-based ResNet50 Model

Misclassified Results

This model misclassified 1454 images (around 1% of our dataset) with 364 real faces misclassified as fake, i.e. false negatives, and 1090 fake faces misclassified as real, i.e. false positives.

False Negatives - Real Faces misclassified as Fake (18 samples):

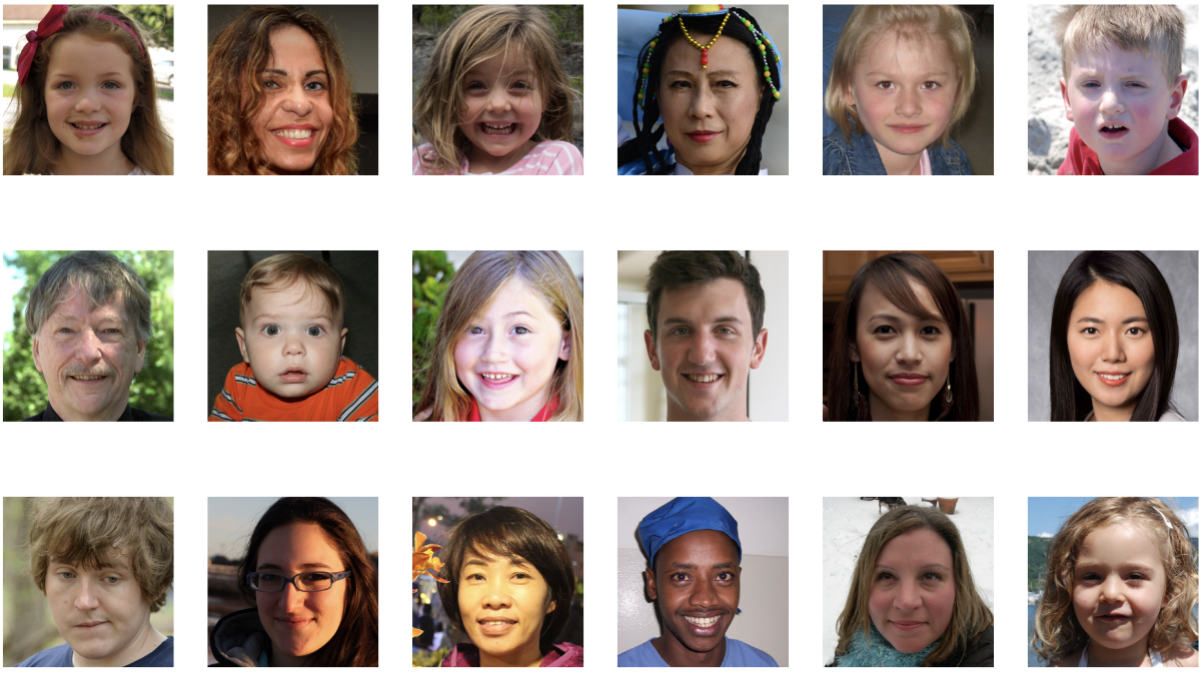

Figure 9. Randomly selected 18 samples from False Negatives

False Positives - Fake Faces misclassified as Real (18 samples):

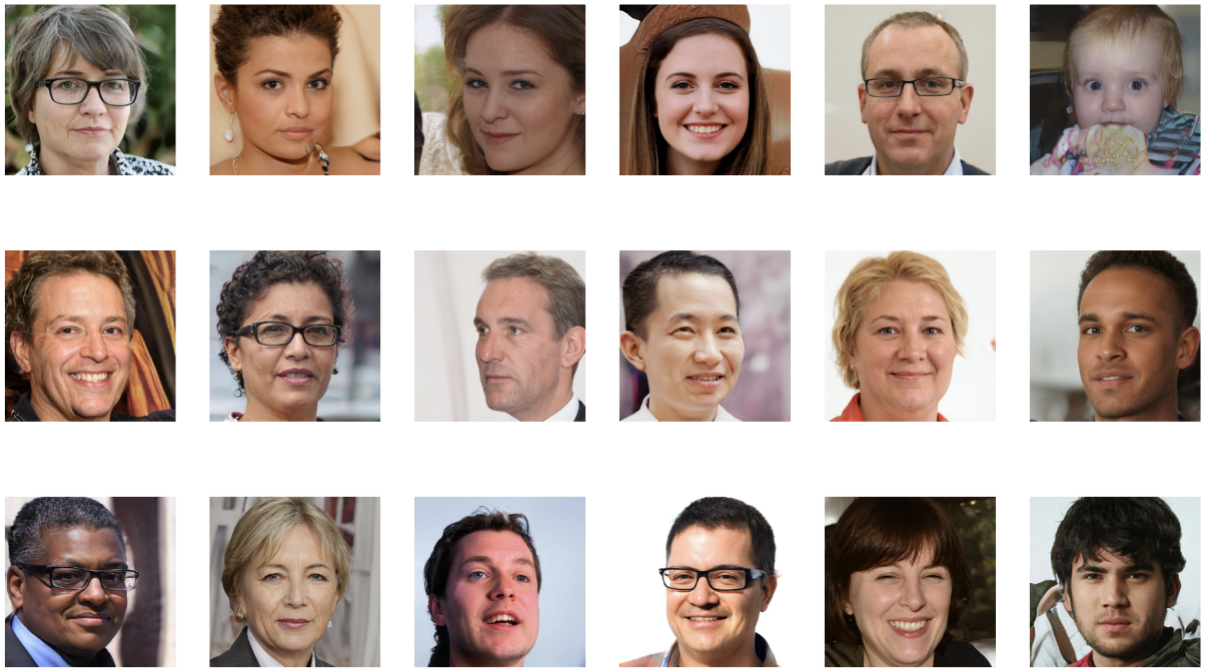

Figure 10. Randomly selected 18 samples from False Positives

Initially, we assumed that a misclassified fake image (false positives) would have different RGB values than real ones (false negatives), as sometimes AI might mistakenly merge details such as ears with hair or clothes. However, after analyzing a sample of the misclassified images, we found that the RGB values did not significantly differ from those of all the data. This means that the differences across the RGB channels are not significantly different among the false positives and false negatives. For the following explanations, we aim to provide some qualitative analysis to supplement our understanding of why some of these misclassifications may be ocurring.

For instance, in ***Figure 10***, in the third row and the last element of this row, we can see that behind the image's background, there are some edges that don't form clear enough to depict any limbs or humans. This is most likely caused by errors in the AI. The skin color of the faces blends in with certain parts of the background, which makes this image appear fake to the human eye. As a result, this image is classified as a false positive.

Likewise, in ***Figure 9***, comparable patterns can be observed in the false negatives. Nonetheless, due to the lack of interpretability of the ResNet50 model, we are unable to access the rationale behind the model's classification of the images as either fake or real. We can’t attribute any particular misclassification precisely for this reason. At best, we can guess as to what may be occurring, such as the tilt of the face, the presence of certain headwear, the lighting and other factors potentially contributing to the misclassification.

In ***Figure 11***, we attempted to perform an RGB value analysis on the misclassified images. We observed that there was no statistically significant difference between the false positives and false negatives. Although our assumption may apply to only a few images, and, in most cases, we are unable to provide a rationale for why the model classified an image as either fake or real.

In order to obtain insights about the racial distributions in our misclassification analysis, we produced ***Table 3*** and ***Table 4*** below. These tables reveal that the majority of the misclassified faces belong to the white racial group, as seen in both misclassified categories. Asians and Hispanics follow right after White. This finding suggests nothing conclusive. By this statement, we clarify that we do not know if there may be a racial bias in the ML algorithm used. If there was, then the unknown distribution of the underlying data would tell us that. Secondly, we do not know if some groups are overrepresented or underrepresented in Nvidia's Flickr and fake faces datasets.

Race	**Count**
White	418
Latino hispanic	91
Middle Eastern	41
Asian	160
Black	44
Indian	15

Table 3. Misclassified Fake Faces Race Distribution

Race	**Count**
White	758
Latino hispanic	214
Middle Eastern	76
Asian	175
Black	45
Indian	12

Table 4. Misclassified Real Faces Race Distribution

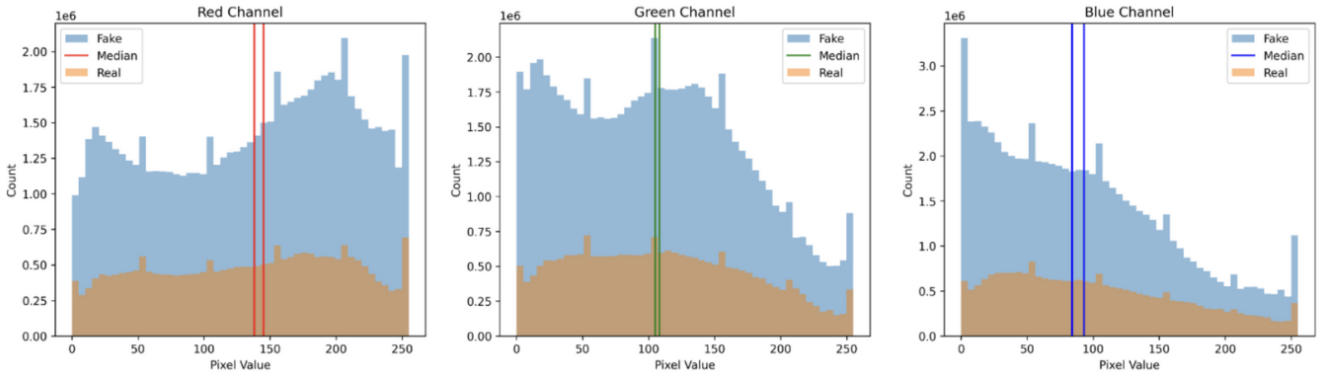

Figure 11. Mean RGB 3 channel Visualization on Misclassified Real and Fake Faces
(Vertical lines indicate the median rgb channel values)

4.3 Grayscale ResNet50 Model Results

The results of the Grayscale model show an accuracy of 93.78%, which is slightly less than the accuracy of the rgb version of our model, indicating that the removal of exposure and lighting effects may have contributed to more fake images to be classified as real. Although, it is possible that a lower accuracy of the grayscale version of the ResNet50 model could be due to lighting and exposure issues in the Fake images, it is not necessarily a definitive indicator of this, and there may be other factors that could have affected the predictions from the grayscale model to be less accurate than the rgb-version. It is important to note that, as the rgb-version of the ResNet50 model is designed to work with color images, which have three color channels (red, green, and blue), it contains more information than grayscale images, which have only one channel representing the intensity of the image. Hence, it could be a reason as to why our grayscale version had a slightly lower accuracy score. Moreover, the ROC curve shows an AUC of 0.9874 which means that the model has a low false positive rate. Additionally the Precision-Recall curve also shows an average precision score of 0.9882, indicating that the model has a high precision and recall, and it can classify the positive samples with high accuracy while minimizing false positives. However, these scores are slightly lower than the rgb-version of the ResNet50 model. As for misclassified imagery, we did not need to reiterate a qualitative analysis of the misclassifications, given that many of the hypotheses that we could state are the same. Quantitatively, we managed to obtain the racial composition of the misclassified images, where we hold the same general conclusions as previously shown. Although, less white fake faces were misclassified in total. We seem to observe a lower False Positive rate for the fake faces.

Race	**Count**
White	255
Latino hispanic	41
Middle Eastern	20
Asian	71

Black	5
Indian	0

Table 5. Misclassified Fake Faces Race Distribution - Grayscale

Race	**Count**
White	810
Latino hispanic	147
Middle Eastern	66
Asian	233
Black	69
Indian	26

Table 6. Misclassified Real Faces Race Distribution - Grayscale

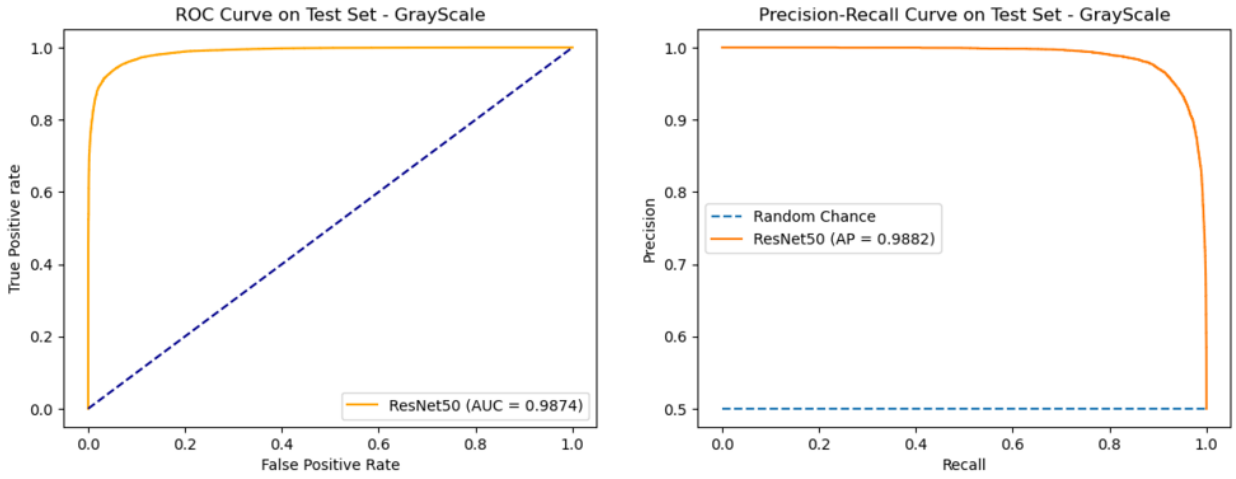

Figure 12. ROC and Precision-Recall Curves for Grayscale-based ResNet50 Model

4.4 Application Space Testing Results

We deployed our model on the complete data and tested it out on an application-based dataset which contains different images relevant to our problem space. We used a mixed dating profile image dataset that includes two different types of fake images: the highly realistic AI-generated Asian faces dataset and faces with fake backgrounds. These types of images present significant challenges to our model as they are designed to deceive the model into classifying them as real. However, by incorporating these types of images into our dataset, we can better evaluate the effectiveness of our model in detecting fake images on dating apps such as Tinder and Bumble, which will help us develop better techniques to detect and prevent the use of fake profile images in online dating for future research.

Specifically in this dataset, we achieved an accuracy of 65% using the rgb-based ResNet50 model, with both precision and recall also at 0.65. Unfortunately, we have an F1 score of 0.394, indicating that the balance between precision and recall is not ideal for our application space, and this will be addressed in our limitation section. Notably, we did not encounter any false negatives in our testing. Testing on the application-space enabled us to identify two main limitations that our model will encounter if it were to be deployed in the real world of dating apps. Firstly, our model was not able to predict any of the highly realistic AI-Generated Faces accurately. Misclassification of these highly realistic AI-generated faces indicate that perhaps these AI-generated faces may not be well-represented in the training data used to train the model, resulting in our model not having learned how to accurately recognize their unique characteristics. Another reason for this misclassification could be that these AI-generated faces were very close to real images and created to be very complex, with many subtle features that are difficult to detect; this, however, highlights a potential threat that similar face detection or classification models may encounter as technological improvements lead to even better AI-generated images. The samples of these misclassified highly realistic AI-generated faces are displayed below.

Figure 13. Randomly selected 3 samples of Misclassified Highly Realistic AI-generated Images

Secondly, our model also misclassified facial images with fake backgrounds as real. The main reason behind this kind of misclassification could potentially be that our model was only trained on facial images, mainly cropped to a thumbnail. Hence, the lack of real background information in our training set made it difficult for our model to distinguish between real and fake backgrounds. As dating and social media profiles may also contain images with backgrounds or body parts in addition to facial information, it might be useful for future tool development in this area to also include background information in the training set. The sample of these types of misclassified images is shown below.

Figure 14. Randomly selected 2 samples of Misclassified Images with Background Information

4. Conclusion

4.1 Conclusion

The rise of fake profiles and fake imagery in dating app scams are not a threat to be taken lightly. We have friends and family who have been scammed and deceived in very unfortunate ways, and they always lost more than just pride, money, and time; they lost hope in building connections.. As such, we felt compelled to attempt to make a dent in this act of deception by trying to assess the viability of our model in classifying fake and real faces. In many respects, we learned an astounding amount of information about the challenges that model deployment will face.

The first thing we learned is that the model is only as good as its data. Our proxy dataset was intentionally different from our training data, and we learned that good generalization should not be expected, especially when the images in the application space are so much more varied, from body shots to tilted shots of the face. In addition, the acceleration of AI imagery is a challenge to our model, and we can imagine that it will be a challenge to all dating app companies. As such, we perfectly understand that our model is not currently viable to handle this problem space. We need more training, more models for comparison, and an even more profound understanding of our problem space, as well as rethinking the ways in which we can bolster our methods. For example, we think combining NLP approaches to read the false profile text data can help us improve our predictions for aiding to identify fake profiles, which is a shared goal with wanting to identify fake faces, as they are both means to the same end.

As far as bias or patterns in the misclassifications, we were not able to extract the underlying distribution of the data, and encountered challenges in obtaining meaningful insights. As such, we are keenly aware that we can only know as much as our data. The little that we did observe helped us understand that, although we can make no definitive claims, there may be distributional differences between racial groups within the datasets, which is something we are excited to study as well.

Lastly, any approach to this space will carry limitations because it is rapidly evolving. The images people post on their profiles are widely different and the improvements in fake imagery are growing. We foresee, based on our experience, that very creative solutions involving the use of multiple models and even more varied datasets will be required to properly handle the challenges of the future, and we are grateful that we ventured with our best foot forward, even if the results were not what we expected.

4.2 Limitation

Although the ResNet50 model is a widely used for image classification, it has the following limitations:

- **Limited transfer learning capability:** As ImageNet is used for getting pre-trained weights of the ResNet50 model, it may not be suitable for other datasets with different image characteristics or object categories. In our specific case, it is important to note that although ImageNet contains a diverse range of objects, and not all of them are faces, it might be difficult for the model to learn features that are specific to face detection. This factor was one

we thought to be potentially evidenced when fake backgrounds were detected as real. Hence, even more fine-tuning might be necessary to achieve better performance.

- **Limited interpretability:** ResNet50, like other deep learning models, is considered a black box with a complex architecture and numerous parameters. In our application space, it is pertinent to not only just classify faces as fake or real, but to also understand in what cases our model was not able to perform well or why a certain fake image might be classified as real. In that case, we had to manually look through the false positive and false negative data to interpret how the model makes its predictions or what features it is using to make those predictions.
- **Generalization to new types of fake images:** With new AI tools, the kinds of fake facial images available online are constantly changing. As our model may be trained on a specific kind of fake images, it may not generalize well to other types of fake facial image data. In our application dataset, we have images that encompass a wide range of faces, including animals with humans in the pictures, very obvious fakes pasted on real bodies,. Consequently, it is uncertain whether our model's performance can be applied to the new dataset in the application domain to an adequate extent that might compromise its external validity.
- **Unknown distribution and computational constraints:** One of the greatest challenges was that we could never ascertain the distribution of races and gender across the dataset of real and fake faces. The algorithm to know this information for the misclassified faces worked because the data was small, but not with 140,000 images due to speed constraints. In addition, not knowing the racial composition of the source datasets significantly hindered our findings on racial bias. We also could not expand the study to account for identifying the amount of faces with glasses and specific types of headwear. Furthermore, preprocessing the data and modeling the data completely required around 4 hours, irrespective of troubleshooting.
- **Limited External Validity**: Our poor proxy performance and even the excellent test performance results on the test set by the ResNet50 model may be attributed to the fact that the Nvidia's fake faces dataset was generated and trained on the Flickr images also scraped from Nvidia. This close relationship between the datasets may explain the performance patterns shown in the results section, where we speculate led us to potentially overfitting on the data. This can impact the generalizability of the model to new and unseen data. We should consider the possibility of overfitting in the evaluation of the model's performance and take steps to mitigate this issue, such as using a more diverse set of fake data or using regularization techniques during training.

4.3 Future Considerations

1. We want to develop a way to assess the impact of the fake image quality based on how effectively it fools the discriminative models, which is an extension of this study.
    a. There are multiple ways to improve deepfake detection, and many of them are currently underdeveloped in research. One potential solution is to implement adversarial ML techniques, which can be used to create more sophisticated deepfakes that are designed to evade detection by existing deepfake detection models. By learning to make better fakes, we may be able to train models on them and improve our detection of fake faces. In response to the trends of adversarial techniques, researchers are currently designing new techniques for adversarial training that can improve the ability of deepfake detection models to defend against these types of attacks. Additional methods are available, depending on the specific needs and requirements.

2. We would like to add more types of models to the study when computing resources becomes less of a limiting factor.
    a. Several previous studies have demonstrated the effectiveness of models such as the Vision Transformer (ViT) and CNN for deepfake detection. However, due to constraints in time and computation, we aim to evaluate additional models for comparison in terms of interpretability and performance metrics. Our goal is to identify the model that achieves the highest level of interpretability and provides the best performance on the given task for future research.

3. We are implementing an unsupervised learning strategy (PCA) to learn what makes a real image as a form of feature extraction, and then use that information for classification.
    a. When feasible, we may consider enlarging the solution to, for example, 1024 by 1024. In this case, PCA may be necessary to extract the key features from the high-dimensional data and identify clusters or common points across different images in the two groups.

References

[1] Olivera-La Rosa A, Arango-Tobón OE, Ingram GPD. Swiping right: face perception in the age of Tinder. Heliyon. 2019 Dec 2;5(12):e02949. doi: 10.1016/j.heliyon.2019.e02949. PMID: 31872122; PMCID: PMC6909076.

[2] Taeb, Maryam, and Hongmei Chi. 2022. "Comparison of Deepfake Detection Techniques through Deep Learning" Journal of Cybersecurity and Privacy 2, no. 1: 89-106. https://doi.org/10.3390/jcp2010007

[3] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[4] Shahzad HF, Rustam F, Flores ES, Luís Vidal Mazón J, de la Torre Diez I, Ashraf I. A Review of Image Processing Techniques for Deepfakes. Sensors (Basel). 2022 Jun 16;22(12):4556. doi: 10.3390/s22124556. PMID: 35746333; PMCID: PMC9230855.

[5] Maher Salman F, Abu-Naser S. Classification of Real and Fake Faces Using Deep Learning. 2022. International Journal of Academic and Engineering Research. Vol. 6, pp. 1-14. 2022 March. ISSN: 2643-9085. https://philarchive.org/archive/SALCOR-3

[6] Merrigan A, Smeaton A. Using a GAN to Generate Adversarial Examples to Facial Image Recognition. 2021 November 30. arXiv:2111.15213v. https://arxiv.org/pdf/2111.15213.pdf

[7] Wang, J., & Li, Z. (2018). Research on face recognition based on CNN. IOP Conference Series: Earth and Environmental Science, 170, 032110. https://doi.org/10.1088/1755-1315/170/3/032110

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, June 3). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv.org. Retrieved March 10, 2023, from https://arxiv.org/abs/2010.11929

[9] M. Coşkun, A. Uçar, Ö. Yildirim and Y. Demir, "Face recognition based on convolutional neural network," 2017 International Conference on Modern Electrical and Energy Systems (MEES), Kremenchuk, Ukraine, 2017, pp. 376-379, doi: 10.1109/MEES.2017.8248937.

[10] Coccomini, D. A., Caldelli, R., Falchi, F., Gennaro, C., & Amato, G. (2022, June 28). Cross-forgery analysis of vision transformers and CNNS for deepfake image detection. arXiv.org. Retrieved March 10, 2023, from https://arxiv.org/abs/2206.13829

[11] Aura. The Unexpected Dangers of Online Dating [11 Scams to Know]. Retrieved March 10, 2023 from https://www.aura.com/learn/dangers-of-online-dating

[12] Datagen, "ResNet-50: The Architecture Explained," Datagen, Retrieved on April 5, 2023. [Online]. Available: https://datagen.tech/guides/computer-vision/resnet-50/

[13] Nvidia Corporation. (2019). FFHQ Dataset. [Online]. Available: https://github.com/NVlabs/ffhq-dataset.

[14] S. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, F. Moreno-Noguer, "GANimation: One-Shot Anima-tion of Facial Expressions using Generative Adversarial Networks," arXiv preprint arXiv:1812.04948, 2018.

[15] V7 Labs. "1 Million Fake Faces." V7 Labs Open Datasets. https://www.v7labs.com/open-datasets/1-million-fake-faces.

[16] Duke University Data Science Program, "Graduates of 2024," Accessed April 20, 2023. https://datascience.duke.edu/people/graduation-year/2024/.

[17] Civitai. "Homepage," Accessed April 20, 2023. https://civitai.com/.

[18] Agarwal, R., AbdAlmageed, W., Wu, Y., & Natarajan, P. (2020). Deep Fake Detection: A Survey of Facial and Body Cues. In Proceedings of the 1st Workshop on Deep Learning for Deepfakes Detection (pp. 1-6). Papers with Code. Retrieved from https://paperswithcode.com/paper/deep-fake-detection-survey-of-facial

[19] Ma, S., Wang, J., Huang, H., & Cui, Y. (2021). Deepfake Detection Based on Pupil Light Reflex and Convolutional Neural Networks. arXiv preprint arXiv:2106.14948.

[20] Federal Trade Commission. (2022, February). Reports of Romance Scams Hit Record Highs in 2021. [Online]. Available: https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/02/reports-romance-scams-hit-record-highs-2021

[21] A. Smith and B. Johnson, "The Rise of Deepfake Technology: Implications for Society and National Security," IEEE Security & Privacy, vol. 18, no. 1, pp. 26-33, Jan.-Feb. 2020.

[22] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," ImageNet, Available: https://www.image-net.org/about.php. [Accessed: Apr. 19, 2023].

[23] N. N. Luong, T. Le, and D. Niyato, "A Comprehensive Survey on Blockchain Energy Systems," arXiv preprint arXiv:1909.11573, 2019. [Online]. Available: https://arxiv.org/pdf/1909.11573.pdf. [Accessed: Apr. 19, 2023].

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv preprint arXiv:1512.03385v1, Dec. 2015. [Online]. Available: https://arxiv.org/abs/1512.03385v1. [Accessed: Apr. 19, 2023].

[25] N. Mir et al., "Fake Identity through Online Dating Applications," 2019 Curtin University's Networked Society Conference, Perth, Australia, 2019, pp. 1-5.

[26] P. Singh and P. Shukla, "Recognition of Fake Profiles in Social Media: A Literature Review," Gyan Vihar University Journals, vol. 1, no. 1, pp. 10-15, Aug. 2019.

[27] J. Smith and K. Lee, "The Role of Profile Pictures in Online Dating: A User Study," IEEE Transactions on Human-Machine Systems, vol. 49, no. 3, pp. 211-219, May 2019.

[28] C. Kanan and G. W. Cottrell, "Color-to-Grayscale: Does the Method Matter in Image Recognition?" in IEEE PLoS ONE, vol. 7, no. 1, pp. e29740, Jan. 2012, doi: 10.1371/journal.pone.0029740.