Fakebook

### **by Amy Rosato,**

### **Chetan shah,**

### **Swagat Kumar Pendayla**

# Final assignment

## Abstract

Artificial Intelligence (AI) is having a massive impact on the way we interact with technology. AI is being used to create deepfakes, which are realistic digital representations of people or objects that can be manipulated to make them look or sound like something they are not. Deepfakes have become a popular tool in the online discourse, allowing users to spread misinformation and sway public opinion (Sample, 2020). As AI technology continues to advance, deepfakes are becoming increasingly difficult to detect. This is leading to greater concern about the potential for deepfakes to be used to manipulate public opinion and spread misinformation (Sample, 2020). It is crucial to understand the potential for deepfakes to be used in online discourse and the potential impact it could have on our society, to help mitigate any foreseen threats against public safety.

## Introduction

As the obsession grows to develop machines that mimic human behavior and intelligence, we are seeing a rise in subcategories of artificial intelligence and machine learning due to continuous research and the ability to collect, store and process large amounts of data (Marr, 2018). There is no denying that AI and ML have sprung into focus and have seen an increase in adoption across all industries in the past few years. As access to processing power and data drive this new trend, the world struggles to play catch-up in understanding how this technology should be used, and how the current society structures effect the impact of these technologies. Companies, industries, and governments are looking to implement policies to help combat the rising amount of people suffering from adoption of AI, as ‘manipulation, bias, social discrimination, violations of privacy’ become the headlines associated with AI (Ebers & Navarro, 2019) but in some instances not quick enough.

In recent years, we have seen how ‘fake news’ and false claims can be shared and accelerated thanks to the combination of social media. The most obvious example is the spread of hatred by Trump which led ultimately to the Capital One building being stormed by Trump supporters (Schroeder, 2021). Since engagement leads to revenue, we have seen many social media technology companies back away from condemning and clamping down on their platforms (Babuta & Sullivan, 2018). When we look to then factor in digital impersonation, where someone could “hijack someone’s identity – voice, face, body”, we start to imagine examples of where content of believable videos are being created of people saying and doing things they never did (Chesney & Citron, 2018). Deepfakes pose an immediate threat to national security and democracy if left unchecked and unregulated.

Identity, harassment, and blackmail are other threats that could arise. There have already been fake sex videos of celebrities created without consent (Chesney & Citron, 2018). Reducing women to sexual objects, harming their reputation and career are some of outputs of unregulated uses of deepfakes. We can see how this technology in the wrong hands of stalkers and abusers could add to the ever-rising problem of abuse in today’s world.

## Artificial intelligence

First, it is worth understanding what Artificial Intelligence (AI) is before diving into the subfields which help us understand what technologies power Deepfakes. AI has been the focus of development since 1956 as we have aimed to create systems that can solve complex problems which required logical reasoning (Dickson, 2021). At the Dartmouth AI conference in 1956 the technology was described as such: “Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (Reese, 2017). AI is an umbrella term which includes a range of approaches to help with problem solving, learning, reasoning, and understanding language (Bullinaria, 2005). Examples of AI are all around us, from voice recognition like Amazon Alexa, to a computer playing a game of chess. There are many subfields of AI some of which to mention are Neural Networks, Robotics and Machine Learning (Bullinaria, 2005).

Artificial intelligence (AI) is a branch of computer science that focuses on developing intelligent machines that can think, reason and act like humans (IBM, 2020). AI-driven machines use algorithms, data-driven models, and predictive analytics to interact with their environment and make decisions. AI enables machines to solve problems and learn from experience without requiring explicit programming (IBM, 2020). AI technology is being used in various industries today, including healthcare, finance, retail, transportation, and manufacturing. AI is also being used for applications such as facial recognition, natural language processing, and robotic process automation. AI is quickly becoming a major part of our lives, and its potential for improving our lives is only beginning to be tapped.

However, there are many concerns associated with AI. These include the potential for AI to make decisions that may be biased or unethical, the risk of introducing unintended biases into AI models, and the potential for AI to automate jobs and displace human workers (Walch, 2019). Additionally, AI systems can be vulnerable to cyberattacks, and AI-driven systems may lack the ability to explain their decisions, meaning that their decisions are difficult to understand and may lack transparency (Walch, 2019). Finally, there is the risk of AI systems becoming autonomous, meaning they may operate without human oversight or input.

## Deep learning

Machine learning (ML) is a subfield of AI and requires data to be trained to predict an outcome. ‘ML algorithms are at their core complex mathematical functions that map observations to outcomes’ (Dickson, 2021). There is a strong argument that machine learning is only as good as the data it receives, as it uses data to learn and recognize patterns to make predictions. User cases of ML can include facial recognition, object recognition, translation, amongst other tasks (Reese, 2017).

Deep Learning is therefore a subset of machine learning where we have created artificial neural networks which mimic our human brains, to learn from large datasets in order to provide an output or perform a particular task (Marr, 2018). This method is very much based on how humans learn from experience. A machine would repeatedly carry out a task, making small tweaks each time to improve the outcome. Unlike ML, which is utilizing mathematical functions on top of a dataset to predict an outcome, deep learning tries to find associations between a set of inputs and outputs through a neural network with multiple hidden layers (Liang, 2018). Liang goes onto explain that information is passed between the network hidden layers which are aligned a weight and bias parameter which can be tuned to provide better performance. The overall evaluation of the network’s performance is done through something called a loss function, and the goal is to ‘minimize loss by adjusting weights and biases of the network’ (Liang, 2018). The name deep learning comes from the multiple network layers called “nodes” and they require vast amount of data to ‘learn’. It is also worth pointing out that because of this, deep learning approaches are susceptible to difficulties such as cost and bias (Reese, 2017).

## deepfakes

One of the most recent developments in AI is Deepfakes, a technology that uses deep learning to generate highly realistic digital images and videos (Sample, 2020). Deepfakes can be used to create false images of people, places, and objects and can be used to manipulate public opinion, spread misinformation, and commit fraud. Deepfakes rely on generative adversarial networks, which are AI algorithms that are trained on a large dataset of images in order to generate realistic images that can be used to create realistic videos (Spring Wise, 2020).

GANs, or Generative Adversarial Networks, are a type of deep learning technique used to create “deepfake” videos and images. Generative adversarial networks (GANs) were a by-product of a computer scientist called Ian Goodfellow, where GANs can create convincing images using an algorithm (Spring Wise, 2020). Deepfakes are generated using two neural networks that work against each other. The first is a generator, which is responsible for creating the fake content (Lakshmanan, Gillard, & Gorner, 2021). This is done by taking in data such as images, videos, and audio and creating a new, fake version. The second is a discriminator, which is responsible for distinguishing between the real and the fake content (Lakshmanan, Gillard, & Gorner, 2021). The two networks are trained on the same data set and compete against each other. The generator’s goal is to create convincing output that can fool the discriminator, while the discriminator’s goal is to identify which output is real and which is fake. The two networks learn from each other as they compete, allowing them to improve and become more sophisticated over time (Lakshmanan, Gillard, & Gorner, 2021).

Although initially, most of the outcomes of this technology seem to be negative, there are some positives we must consider. When we look specifically at art creation, editing and user cases within the production world, GAN technology can help save these industries money, as photos and videos do not need to be reshot if something is wrong in a scene or image. Adverts, photographs, and videos can all be reused and updated to represent a present date or reference without the need of reshooting (Spring Wise, 2020). Not only does this contribute positively to sustainability efforts but also to driving cost savings and efficiencies. Outside of entertainment purposes there are arguments for using GAN technology for Research and training user cases. For example, GAN algorithms could be “used to create “fake” brain scans based on actual patient data. These fake scans are then used to train algorithms to spot tumors in real images” (Spring Wise, 2020).

The cons to this technology seem to currently outweigh the pros, but this may change as the world adapts and catches up to the technology. As it stands today, the lack of legislation around this technology means we are dependent on technology companies enforcing their own ethics policy. Companies like Microsoft will not synthesize any voices or images of individuals who have passed and did not explicitly write in their will that their voice and image could be used. Whereas companies such as Metaphysic AI (Metaphysic AI, 2022) have actively used deepfake technology of Elvis on the recent series of America’s got Talent. Without any real guardrails in place, we can see dangerous situations of scamming and “fake news” being spread. During covid, the number of online scams that were happening rose drastically, as scammers took advantage of a new situation, we were all adapting to (Simmons & Quinton, 2021). We can all recall a time where we have had a scammer call us trying to cheat us out of our money, imagine if that scammer then had a face and voice of someone you knew and trusted. When we focus on the damage and hatred that fake news can cause to groups across the world, imagine if this new messaging was seen to be promoted by someone who was viewed as a respectable individual, but it was deepfake technology. Where we have celebrities having to declare “this is an Ad” when promoting something on their social channels, similar needs to be done when companies are utilizing GAN technologies. I also believe we need to develop and enhance our ability to be able to tell if an image or video is utilizing deepfakes. Just as effort is being put into to develop GAN algorithms, a counter needs to be developed to help identify use of deepfakes and protect people globally.

## our approach

To help prevent the misuse of GANs, computer vision models can be used to detect deepfakes. These models use machine learning algorithms to analyze images and videos and identify subtle differences between real and fake content. The models use a variety of techniques such as facial recognition, motion detection, and video analysis to detect deepfakes.

We will be building a computer vision model that will be able to tell the difference between an AI generated image compared to that of an image of a real person. We will be utilizing two training sets, one of real people and one of ‘fake’ people. These datasets can be found here: [Real and Fake Face Detection | Kaggle](https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection)

* Real people dataset: [Real and Fake Face Detection | Kaggle](https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection)
* Fake people dataset: [Real and Fake Face Detection | Kaggle](https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection)

(Kaggle, 2018)

In order to detect Deepfakes, several preprocessing techniques need to be implemented. These include data augmentation, which involves adding more data to the training dataset in order to increase the accuracy of the model; resizing, reshaping. Preprocessing techniques such as image resizing, reshaping, color space conversion, and normalization can be used to improve the accuracy of the model (Kumar, 2021). From here, we will split the dataset into training and test data 80:20 and apply a pretrained model to our combined and augmented dataset and measure the accuracy and loss of the model. We will use transfer-learning to finetune the pre-trained model on our dataset. This means initializing the model with pre-trained weights on a large dataset, and then training it on our smaller dataset (Yadav, 2022). This can help the model learn to recognize features relevant to our specific task.

Once we are happy with the model’s performance, we will then look to make predictions with this model by feeding the model new images it has not seen before. These images can be uploaded from our own computer and can contain pictures of ourselves. We can also go to [www.thesepeopledontexist.com](http://www.thesepeopledontexist.com) and grab generated AI images of people to test our model.

We also have put together a simple DASH app where you can upload any picture and get a result from the model to display under the image if the image is a real person or an AI generated one. This application is very much dependent on the model itself, but we hope to demonstrate how a more user-friendly GUI could be used for anyone to upload content and get an understanding if it is real or fake.

## the model

When we look at the dataset we will be utilizing to build and train our model, the combined number of images with real and fake is total to 2041. Even with data augmentation we will have a relatively small dataset to train a model from scratch. For this problem we have thus chosen to utilize a pre-trained model. DenseNet is a popular convolutional neural network architecture that is commonly used for image classification tasks. It is well-known for its ability to achieve high accuracy on a wide range of image classification tasks and is often used as a starting point for training models that can recognize real images from fake ones. One of the ways it is so effective on classification takes is because ‘each layer of the network is receiving a “collective knowledge” from all the preceding layers’ (Tsang, 2018). This allows for each channel to be more compact making it more computationally efficient.

One reason why DenseNet is well-suited for this task is because it is a deep learning model, which means that it can learn to recognize complex patterns in data by learning a hierarchy of features at different levels of abstraction (Tsang, 2018). This is particularly useful for tasks like image classification, where the model needs to be able to recognize a wide range of different objects and scenes in order to perform accurately.

Another reason why DenseNet is a good choice for this task is because it is a pretrained model, which means that it has already been trained on a large dataset of images (Yadav, 2022). This allows the model to learn a set of general-purpose visual features that are useful for a wide range of image classification tasks, including the task of identifying real images from fake ones. By starting with a pretrained DenseNet model, we can save a lot of time and computational resources when training our own model, and we can often achieve good performance using relatively little training data.

Therefore, after splitting the data into train and test at a ratio of 80:20, we went ahead building our model. We were classifying all images between two classes – ‘real’ or ‘fake’ and looking at model accuracy and loss to help determine the success of the model. We used the defacto ‘ImageNet’ for the weight as this has been trained on images in 1000 categories, which makes it a great baseline for any image classification model (Stack Overflow, 2020). To also avoid overfitting by training the whole network with this pre-trained model we set the ‘se\_model.trainable = false’. We utilized the **ImageDataGenerator** function to perform real-time data augmentation over the images in a loop.

Early stopping is a technique in which training is stopped before the model has fully converged. This can prevent the model from overfitting by stopping training before the model has had a chance to learn patterns that are specific to the training data and do not generalize to new data. We have used this in our model creation to help stop the model stop training once it stops improving.

Creating the model, we have used the batch normalization technique that ‘mitigates the effect of unstable gradients within the neural networks’ (Alake, 2020). It helps standardize and normalize the input which helps normalize the input images to help the performance of the model. We have then used dropout layers after each input to help drop a fraction of the input unit to help regularize the dataset and model and help avoid overfitting (Sharma, 2020).

In machine learning, an activation function is a function that takes in the weighted sum of all inputs from the previous layer, applies a mathematical operation to them, and produces an output (Praharsha, 2022). This output is then passed to the next layer in the network as input. The rectified linear unit (ReLU) activation function is a piecewise linear function that outputs the input if it is positive, and 0 if it is negative (Praharsha, 2022). This function has been widely used in deep learning models because it has been found to work well in practice. The sigmoid activation function is a s-shaped curve that takes in any input value and outputs a value between 0 and 1 (Ovalle, 2020). This function is often used in models where the output should be interpreted as a probability, such as in binary classification tasks (Ovalle, 2020).

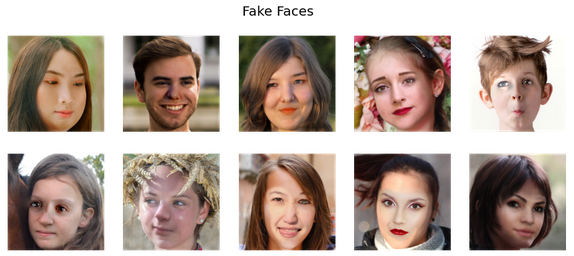
Both ReLU and sigmoid activation functions are commonly used in deep learning models because they can help to introduce non-linearity into the model (Shrivastav, 2022). This can be useful in modeling complex, real-world phenomena that cannot be accurately represented using a linear model. ReLU activation functions are often used in the initial layers of a deep learning model because they can help to introduce non-linearity into the model and allow it to learn more complex representations of the data, this is what we have implemented in the model. This is because ReLU functions can take on a wide range of values, whereas sigmoid functions are always bounded between 0 and 1 (Praharsha, 2022).

In contrast, sigmoid activation functions are often used in the final layer of a model when the output should be interpreted as a probability. This is because sigmoid functions have a smooth, s-shaped curve that allows them to output values between 0 and 1, which can be interpreted as probabilities (Shrivastav, 2022). Again, we have followed this best practice when building our model out.

## results

We started with few real and fake images, below is the sample of them.



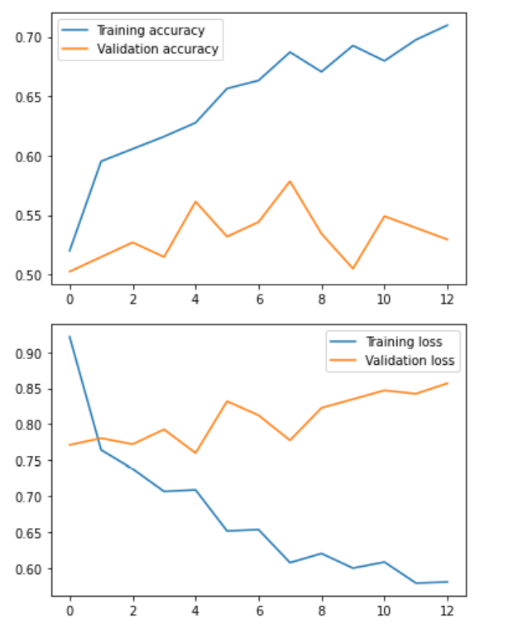


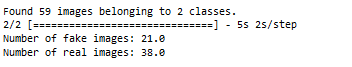
We preprocessed all of the images into dataset e.g., Grayscale, resize, reshape and augmentation. Below are the results after preprocessing them:



Later, we fed the data into our model (we used Densenet121 as our pretrained model). We divided data into Train and Validation sets and trained model.

As you can see from the results of our model, our model performs fairly well on our training data, achieving an accuracy rate of above 70%. However, when we then apply the model on the validation dataset our model performs significantly less accurately. We are always expected to see a lower level of accuracy on the validation dataset, due to the model being comfortable with the training dataset, however, there seems to be a significant drop off in accuracy.

Lastly, we introduced a complete new set of images (mix of real and fake) and predicted test set. Model was able to recognize fake vs real persons as shown below:



There are a few possible reasons why our training accuracy might be higher than our validation accuracy. One possible reason is overfitting. Overfitting occurs when a model performs well on the training data but does not generalize well to new, unseen data. In other words, the model has learned patterns in the training data that do not generalize to the validation data. This can cause the model to perform poorly on the validation data, even though it performs well on the training data.

Another possible reason is that the validation data is not representative of the real-world data that the model will be used on. If the validation data is not a good representation of the data that the model will be used on, then the validation accuracy will not be a good indicator of how well the model will perform in the real world.

In general, it is important to evaluate a model on data that is representative of the real-world data that the model will be used on. This will help ensure that the model is not overfitting and that its performance on the validation data is a good indicator of its performance in the real world.

Another thing to consider, is we preprocessed the image dataset before creating a model. We then took this dataset and ran it through an **ImageDataGenerator**. An **ImageDataGenerator** is a utility in the Keras library that can be used to preprocess images and generate augmented versions of the images on the fly during training. This can be useful in a number of situations, such as when you have a small dataset and want to use data augmentation to improve the model's performance, or when you want to standardize the input images to make them easier for the model to process. Standardizing the images can be beneficial because it can help to ensure that the input images have a consistent scale and distribution. This can make it easier for the model to learn useful patterns from the data and make more accurate predictions.

However, if you have already standardized the images and you use an **ImageDataGenerator** to standardize them again, this could potentially affect the accuracy of the model. This is because the **ImageDataGenerator** may apply different transformations to the images than the ones you used to standardize them, which could change the distribution of the input data in ways that the model is not expecting. This could be an explanation for the poor accuracy and may be something we want to look into changing when moving forward with the enhancements to the model.

The fact the validation loss is much higher than that of the training loss again supports the argument we are potentially seeing the causes of overfitting within our model. Some common techniques for addressing overfitting include:

* Collecting more data: One of the most effective ways to prevent overfitting is to collect more data. This can help the model to learn more accurate and generalizable patterns from the data, which can make it less likely to overfit.
* Regularization: Regularization is a technique that can help to prevent overfitting by penalizing the model for having too many parameters. By adding a regularization term to the loss function, you can encourage the model to use only the most relevant parameters and discard the ones that are not useful.
* Early stopping: Early stopping is a technique in which training is stopped before the model has fully converged. This can prevent the model from overfitting by stopping training before the model has had a chance to learn patterns that are specific to the training data and do not generalize to new data.
* Dropout: Dropout is a technique in which random subsets of the model's parameters are set to zero during training. This can help to prevent overfitting by forcing the model to learn multiple, independent representations of the data, rather than relying on a single, overfit representation.

Enhancements may include experimenting with the dropout rate, since we have already utilized a rate of 0.2. We also might also want to look to collecting more data to help improve our model’s accuracy and also implementing regularization after each layer in the network. Another experiment we could do in later versions is to include feature extraction. Feature extraction involves extracting the most important features from the data in order to improve the accuracy of the model. We will continue to experiment with the above and keep early stopping within the model.

## Ethical considerations

Ethics is an important aspect of AI because it deals with the values and principles that should guide the development and use of AI technology. The field of AI ethics is concerned with issues such as fairness, transparency, accountability, and the potential impacts of AI on individuals and society (Office of Artificial Intelligence, 2019). AI ethics seeks to ensure that AI technology is developed and used in a responsible and ethical manner, and that it respects the values and rights of individuals. This is particularly important given the rapidly advancing capabilities of AI technology and its increasing presence in many aspects of our lives.

There are many companies that are taking a responsible and ethical approach to the development and use of AI technology. Some examples of companies that are known for their strong focus on AI ethics include Google, Microsoft, and IBM (IBM Cloud Education, 2021). These companies have all established guidelines and principles for the ethical development and use of AI and have committed to using AI technology in a way that is transparent, accountable, and respects the rights and interests of individuals. In addition, there are many organizations and research groups that are dedicated to promoting ethical AI, such as the Partnership on AI (Partnership on AI, 2022), the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2022), and the Center for Human-Compatible AI at the University of California, Berkeley (Center for Human-compatible artifical intelligence, 2022).

Ethics is important when creating a computer vision model because the way the model is designed and used can have significant consequences for individuals and society (Indeed, 2021). Computer vision models can be used to make decisions that affect people's lives, such as whether to grant someone a loan or parole, and it is important to ensure that these decisions are fair and unbiased. In addition, the use of computer vision technology raises privacy concerns, as it can be used to collect and analyze large amounts of personal data without people's knowledge or consent. Ensuring that computer vision models are designed and used in an ethical manner is essential for protecting people's rights and interests, and for building trust in the technology.

There are several good approaches to ensuring that computer vision models are designed and used in an ethical manner. Some examples of these approaches include:

1. Ensuring that the data used to train computer vision models is diverse, representative, and accurate. This can help prevent bias and discrimination in the model's predictions and decisions (Najibi, 2020).
2. Designing computer vision models with interpretability and explainability in mind. This can help ensure that the model's decisions are transparent and understandable and can help prevent unintended consequences (M, 2022).
3. Establishing clear guidelines and principles for the ethical development and use of computer vision models. This can help ensure that the technology is used in a responsible and accountable manner (M, 2022).
4. Engaging with stakeholders, such as researchers, policymakers, and the public, to discuss the ethical implications of computer vision technology and to identify potential risks and challenges. This can help promote dialogue and collaboration in the field of AI ethics.

Overall, ensuring the ethical use of computer vision technology requires a combination of technical solutions, transparent and accountable processes, and ongoing dialogue and collaboration.

We will take all ethical considerations and advice into consideration when modeling and enhancing our model. Our initial datasets we have sourced we would need to look into the grounds on how this data was collected, especially when looking at the pictures of real people. We will look to make sure all demographics, genders and races are equally represented. We will also look to put in a place a constant review of the model and we can utilize tools in the field provided by technology organizations like Microsoft, to test our models for bias, fairness, and transparency and to also test the model in simulated environments so we can catch anything that may cause harm to society before it hits production.

## conclusions

In conclusion, we understand the need to research and invest time in creating models to help classify between deepfake images and videos. This is due to the risk of fake news and misinformation being spread throughout the public, which can lead to the break-down in trust with technology but also the break-down in potentially the political systems that uphold democracy.

We have seen in recent years positive user cases of this technology being utilized. Normally this is within the media and entertainment space, where images can be quickly created and edited without the need for reshoots. We also have seen companies entering talent shows such as America’s got talent, to showcase how this technology could be used to bring old celebrities back to life. However, more and more technology organizations are writing their AI ethics approach and guidelines to include more guardrails which stops anyone from impersonating others without formally written consent.

Although ethics is a hot topic within AI, there seems to be a delayed implementation of rules and guidelines as new technology is released onto the world before the legal systems can catch-up. Therefore, it is down to individual companies to decide where they stand in the utilization of this technology.

In regard to the model we have created, for a first go at creating the Fakebook model to help classify between real images of people and fake images of people, we believe it was a good step in the right direction. As mentioned above, we can see the model is not where we would want it to be in terms of accuracy, but there is still room for experimentation to help continuously work towards a better accuracy rate. We think one of those approaches would be to try the DenseNet121 model on the preprocessed dataset without applying an **ImageDataGenerator**. If this failed to produce an improved accuracy rate, we would then look to either experimenting with the dropout rate or potentially adding in layers to the model itself.

The DASH app created could be improved aesthetically, but the ease-of-use the GUI provides, as it allows for individuals to upload pictures to check if the image is of a real or fake person, is again a great use of embedding the model into something the public can utilize.

Going forward, after enhancements have been made to the model to help improve accuracy, we would like to extend the capability of classification out from just images to videos. We can all picture examples of where we might come across content of celebrities saying or doing something they weren’t. In many user cases today, the number of celebrities being deepfaked into pornographic scenes has been on the rise since this technology has come onto the market. This is damaging to someone’s image but also promotes degrading views of women in today’s society (CARE, 2022). Being able to then display if a video is AI generated would be beneficial for the protection of Individuals and again helps promote trust with technology and the spread of misinformation.

# References

Alake, R. (2020, April 23). *Batch Normalization In Neural Networks (Code Included)*. Retrieved from Medium: https://towardsdatascience.com/batch-normalization-in-neural-networks-code-d7c9b88da9f5

Babuta, A., & Sullivan, J. (2018, April 17). *https://rusi.org/explore-our-research/publications/commentary/how-can-social-media-companies-stop-spread-fake-news*. Retrieved from Rusi: https://rusi.org/explore-our-research/publications/commentary/how-can-social-media-companies-stop-spread-fake-news

Bullinaria, J. (2005). *The Roots, Goals and sub-fields of AI.* Retrieved from https://www.cs.bham.ac.uk/~jxb/IAI/w2.pdf#:~:text=AI%20now%20consists%20many%20sub-fields%2C%20using%20a%20variety,learning%2C%20time%20series%20version%20prediction%2C%20space%20learning%20classification

CARE. (2022, September 29). *‘Deepfake’ pornography is ruining women’s lives*. Retrieved from CARE: https://care.org.uk/news/2022/09/deepfake-pornography-is-ruining-womens-lives

Center for Human-compatible artifical intelligence. (2022). *Center for Human-compatible artifical intelligence*. Retrieved from Center for Human-compatible artifical intelligence: https://humancompatible.ai/

Chesney, R., & Citron, D. (2018, February 21). *Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?* Retrieved from Lawfare: https://www.lawfareblog.com/deepfakes-looming-crisis-national-security-democracy-and-privacy

Dickson, B. (2021, January 18). *Understanding the AI alignment problem*. Retrieved from Tech Talks: https://bdtechtalks.com/2021/01/18/ai-alignment-problem-brian-christian/

Ebers, M., & Navarro, S. (2019). Regulating AI and Robotics: Ethical and Legal Challenges. In *Algorithms and Law,* (pp. 1-51). Cambridge: Cambridge university press.

IBM. (2020, June 3). *What is artificial intelligence*. Retrieved from IBM: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

IBM Cloud Education. (2021, March 18). *AI Ethics*. Retrieved from IBM: https://www.ibm.com/cloud/learn/ai-ethics

IEEE. (2022). *THE IEEE GLOBAL INITIATIVE ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS*. Retrieved from IEEE: https://standards.ieee.org/industry-connections/ec/autonomous-systems/

Indeed. (2021, August 30). *What Is Social Responsibility and Why Is It Important?* Retrieved from Indeed: https://in.indeed.com/career-advice/career-development/what-is-social-responsibility

Kaggle. (2018). *Real and Fake Face Detection*. Retrieved from Kaggle: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection

Kumar, B. (2021, February 10). *Image Preprocessing - Why is it Necessary*. Retrieved from Medium: https://medium.com/spidernitt/image-preprocessing-why-is-it-necessary-8895b8b08c1d

Lakshmanan, V., Gillard, R., & Gorner, M. (2021). *Practical Machine Learning for Computer Vision.* Sebastopol: O'Reilly.

Liang, J. (2018, October 19). *An introduction to Deep Learning*. Retrieved from Towards Data Science: https://towardsdatascience.com/an-introduction-to-deep-learning-af63448c122c

M, K. (2022, February 14). *How to avoid bias in computer vision models*. Retrieved from Roboflow: https://blog.roboflow.com/how-to-avoid-bias-in-computer-vision-models/

Marr, B. (2018, October 1). *What is deep learning AI? A simple guide with 8 practical examples*. Retrieved from Forbes: https://learn-us-east-1-prod-fleet02-xythos.content.blackboardcdn.com/5c2103143e6a3/6787448?X-Blackboard-Expiration=1656514800000&X-Blackboard-Signature=2riwrw%2BEhBBflaCD3IIn%2BUPqcnPLBF5PXzxyztQ2Av4%3D&X-Blackboard-Client-Id=962616&response-cache-contro

Metaphysic AI. (2022). *Metaphysic AI*. Retrieved from Metaphysic AI: https://metaphysic.ai/

Najibi, A. (2020, October 24). *Racial Discriminiation in Face Recognition Technology*. Retrieved from Harvard University: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Office of Artificial Intelligence. (2019, June 10). *Understanding artificial intelligence ethics and safety*. Retrieved from UK Gov: https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety

Ovalle, D. (2020, August 12). *The purpose of activation functions in the neural network*. Retrieved from Medium: https://medium.com/@1087/the-purpose-of-activation-functions-in-the-neural-network-be08f81b7c32

Partnership on AI. (2022). *Partnership on AI*. Retrieved from Partnership on AI: https://partnershiponai.org/

Praharsha, V. (2022). *ReLU (Rectified Linear Unit) Activation Function*. Retrieved from Open Genus: https://iq.opengenus.org/relu-activation/

Reese, H. (2017, January 17). *Understanding the differences between AI, Machine Learning, and deep learning*. Retrieved from TechRepublic: https://deeplearning.lipingyang.org/wp-content/uploads/2016/11/Understanding-the-differences-between-AI-machine-learning-and-deep-learning-TechRepublic.pdf

Sample, I. (2020, January 13). *What are deepfakes and how can you spot them?* Retrieved from The Guardian: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

Schroeder, J. (2021, January 12). *Did social media echo chambers lead to the US Capitol riot?* Retrieved from Linkedin: https://www.linkedin.com/pulse/did-social-media-echo-chambers-lead-us-capitol-riot-jenna-schroeder/

Sharma, P. (2020, October 25). *Keras Dropout Layer Explained for Beginners*. Retrieved from Machine Learning Knowledge: https://machinelearningknowledge.ai/keras-dropout-layer-explained-for-beginners/#:~:text=Dropout%20Layer%20is%20one%20of%20the%20most%20popular,accuracy%20on%20the%20test%20data%20or%20unseen%20data.

Shrivastav, A. (2022). *Sigmoid Activation (logistic) in Neural Networks*. Retrieved from Open Genus: https://iq.opengenus.org/sigmoid-logistic-activation/

Simmons, D., & Quinton, M. (2021, March 24). *Covid fraud: £34.5m stolen in pandemic scams*. Retrieved from BBC: https://www.bbc.co.uk/news/technology-56499886

Spring Wise. (2020, October 9). *PROS AND CONS: DEEPFAKE TECHNOLOGY AND AI AVATARS*. Retrieved from Spring wise: https://www.springwise.com/pros-cons/deepfake-technology-ai-avatars/#:~:text=Deepfakes%3A%20Pros%201%20Art%20Star%20Wars%20fans%20are,4%20Identity%20protection%20...%205%20Financial%20saving%20

Stack Overflow. (2020, July). *Why are Keras models instantiated with imagenet weights only?* Retrieved from Stack Overflow: https://stackoverflow.com/questions/62739893/why-are-keras-models-instantiated-with-imagenet-weights-only

Tsang, S.-h. (2018, November 25). *Review: DenseNet — Dense Convolutional Network (Image Classification)*. Retrieved from Towards Data Science: https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803

Walch, K. (2019, December 19). *Ethical Concerns with AI*. Retrieved from Forbes: https://www.forbes.com/sites/cognitiveworld/2020/12/29/ethical-concerns-of-ai/?sh=72e219a723a8

Yadav, B. (2022, May 8). *Image Classification using pre-trained DenseNet model in PyTorch | by Bhimraj Yadav*. Retrieved from Medium: https://bhimrazy.medium.com/image-classification-using-pre-trained-densenet-model-in-pytorch-c1d7fba627db