Identifying StyleGAN images

Neil Foxcroft

orcid.org/0000-0002-8389-8826

Thesis submitted for the degree *Bachelors Honours* in Computer Science and Information Technology at the **NWU**

Supervisor: Dr. R. Serfontein

2021

School of Computer Science and Information Systems

North-West University

Student number: 28418077

Acknowledgements

I would like to thank the following persons for their support and guidance throughout this project that enabled me to complete it successfully.

Dr. Rudi Serfontein, Thank you for your insight and support in my research project and always being there to answer any questions I had. Your guidance enabled me to complete the project successfully and professionally.

Prof Tiny Du Toit, Thank you for the knowledge you shared with me regarding anything related to artificial intelligence. As a master in your craft it was helpful to learn from you. Thank you for making available your GPU that helped me train a portion of my artefact on.

Manre' van Zyl, Thank you for your help and support and the collaboration in the Labs throughout the year.

Affaan Muhammad, Thank you for your insights and input and the support provided throughout the year.

Abstract

The creation of a Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN) introduced the world to the possibility that images can be generated in such a way that it is difficult for a human to detect that these images are not real. A study conducted on StyleGAN and Convolutional Neural Networks helped in the identification of these images. A was created artefact that demonstrate the proposed neural network solution method. The created CNN proved that this method of detection resulted in promising results. To improve the detection accuracy further the hyperparameter optimization framework Optuna was implemented on the neural network and dataset to increase the model's accuracy drastically. Hyperparameter optimization is a powerful tool in machine learning that can structure a neural network architecture in such a way that massive gains are made in the networks predictions. To allow for easy interaction a Front-end web application formed part of the artefact and the Flask framework was used for the model implementation and the user interface. Uploading images to the application will provide users with feedback whether or not a specific image was generated using StyleGAN or if the image is that of a real human being.

Keywords: Convelutional Neural Network, Flask, Hyperparameter Optimization, Machine Learning, Neural Networks, Optuna, StyleGAN

Opsomming

Die skepping van ’n stylgebaseerde skeppende argitektuur vir generatiewe teëstanderige netwerke (StyleGAN) het die wêreld bekendgestel aan die moontlikheid dat beelde op so ’n manier gegenereer kan word dat dit moeilik is vir ’n mens om vas te stel dat hierdie beelde nie werklik regte beelde is nie. ’n Studie wat op StyleGAN en Konvolusiese Neurale Netwerke gedoen is, het gehelp met die identifisering van hierdie beelde. ’n Artefak was geskep wat die voorgestelde neurale netwerk oplossing metode demonstreer. Die geskepte KNN het bewys dat hierdie metode van identifikasie belowende resultate tot gevolg gehad het. Om die identifikasie akkuraatheid verder te verbeter, is die hiperparameter optimeringsraamwerk Optuna op die neurale netwerk en datastel geïmplementeer om die model se akkuraatheid drasties te verhoog. Hiperparameter-optimering is ’n kragtige instrument in masjienleer wat ’n neurale netwerkargitektuur op so ’n manier kan struktureer dat massiewe winste gemaak word in die netwerkvoorspellings. Om maklike interaksie moontlik te maak was daar ’n webtoepassing deel van die artefak ontwikkel en die Flask-raamwerk is gebruik vir die modelimplementering en die gebruikerskoppelvlak. Die oplaai van beelde na die toepassing sal gebruikers terugvoer gee of ’n spesifieke beeld met StyleGAN gegenereer is en as die beeld dié van ’n regte mens is.

Sleutelwoorde: Flask, Hiperparameter-optimering, Konvolusiese Neurale Netwerk, Masjienleer, Neurale Netwerke, Optuna, StyleGAN

Table of Contents

List of Figures

List of Tables

Table of abbreviations

A table containing a list of abbreviations in the order of appearance, that will be used throughout text.

StyleGAN	a Style-Based Generator Architecture for Generative Adversarial Networks
GAN	Generative Adversarial Network
AI	Artificial Intelligence
DSRM	Design Science Research Methodology
SWOT	Strengths, Weaknesses, Opportunities and Threats
PaaS	Platform as a Service
IaaS	Infrastructure as a Service
VS Code	Visual Studio Code
px	pixles
URL	Uniform Resource Locator

Chapter 1

Introduction

The creation of a Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN) and similar technologies introduced the world to the possibility that images can be generated in such a way that it is difficult for a human to detect that these images are not real and instead generated by a neural network. With this proposed research project, the StyleGAN technology will be studied to determine how an approach can be developed to identify artificially generated images to enable the detection of these falsified identities that used the StyleGAN technology. An artefact will demonstrate the proposed solution methods function and convey the successful method in a practical environment with the use of StyleGAN generated images and images retrieved from the Flickr Faces dataset to evaluate the method. The following section is the proposal for the research project and will form the first chapter of this proposed project. In this Chapter, the background regarding the StyleGAN problem and a brief introduction to the StyleGAN technology will be explained. The Aims and Objectives will be set out to provide a clear goal and reference to the success of the final artefact. A Project management framework will be chosen and with the use of a Gantt chart will the timeline be structured to aid in the development of the artefact and completion of the project.

1.1 Project Description

StyleGAN is an open-source Generative Adversarial Network (GAN) that can be used to generate faces of people that do not exist (such as those shown on thispersondoesnotexist.com). This means that fraudsters can use StyleGAN generated faces that normally would pass a visual inspection conducted by a human inspector as part of false identities. The detection of such images with the use of artificial intelligence will be useful because of the factors that currently lead to misidentification.

1.2 Project Background

In this modern world with humanity currently in its 4th industrial revolution, the additions of innovative technologies require original approaches to implement and maintain these technologies. The change from the digital age to the automation age is accelerated by breakthroughs in the fields of Artificial Intelligence (AI) and security (Skilton and Hovsepian, 2017).

One of the big advances in AI is the creation of StyleGAN, this new approach to a GAN allowed for more control in an image than its predecessors (Karras et al., 2019). StyleGAN uses the principles of a GAN to create new images derived from input that specifies what "styles" need to be included in the image. According to Karras et al. (2019), a style is defined in this context as a set of parameters that modifies the input of the image to result in different outputs. If the input received is that the image needs to be in the style of a person with glasses, red hair and must be female, StyleGAN will then generate that image based on images used of that similar styles in the initial training of the model. The resulting image will thus be that of the "styles" required in the input. While this functionality can be utilized for various positive use cases – this breakthrough also creates various challenges and setbacks in the field of security, more specifically the aspects of facial recognition and identity verification as malicious use of this GAN might aid in the creation of fraudulent identities (Mitra et al., 2021). Figure 1.1 demonstrates how an image can be altered with the use of StyleGAN by specifying what style changes should be made on that image.

The styles applied in figure 1.1 demonstrate the capabilities of StyleGAN and further illustrate the possible difficulties of detecting that these images were generated with a GAN (Karras et al., 2019). Possible misidentification of StyleGAN images is a reality that needs to be addressed. Humans in the role of identifying artificially generated human faces may be susceptible to external factors hindering their capabilities and increasing the rate of error in which they identify fraudulent images (Fysh and Bindemann, 2018). Fysh and Bindemann (2018) also noted that in the specific use case of passport officers that were tested on passport images captured on the same day against the "traveller" presenting that image, that the officers made substantial errors in a controlled environment when comparing the picture identity to that of the traveller. These results enforced their original statement that humans struggle with unfamiliar face identification.

Opportunities for fraud escalated with the recent boom in online applications. This bigger threat for fraud is exaggerated in the mobile market due to the high volume and demographic variety of current mobile users (Mitra et al., 2021). Most applications require user accounts and the unique user's identity is commonly the specific value the application requires.

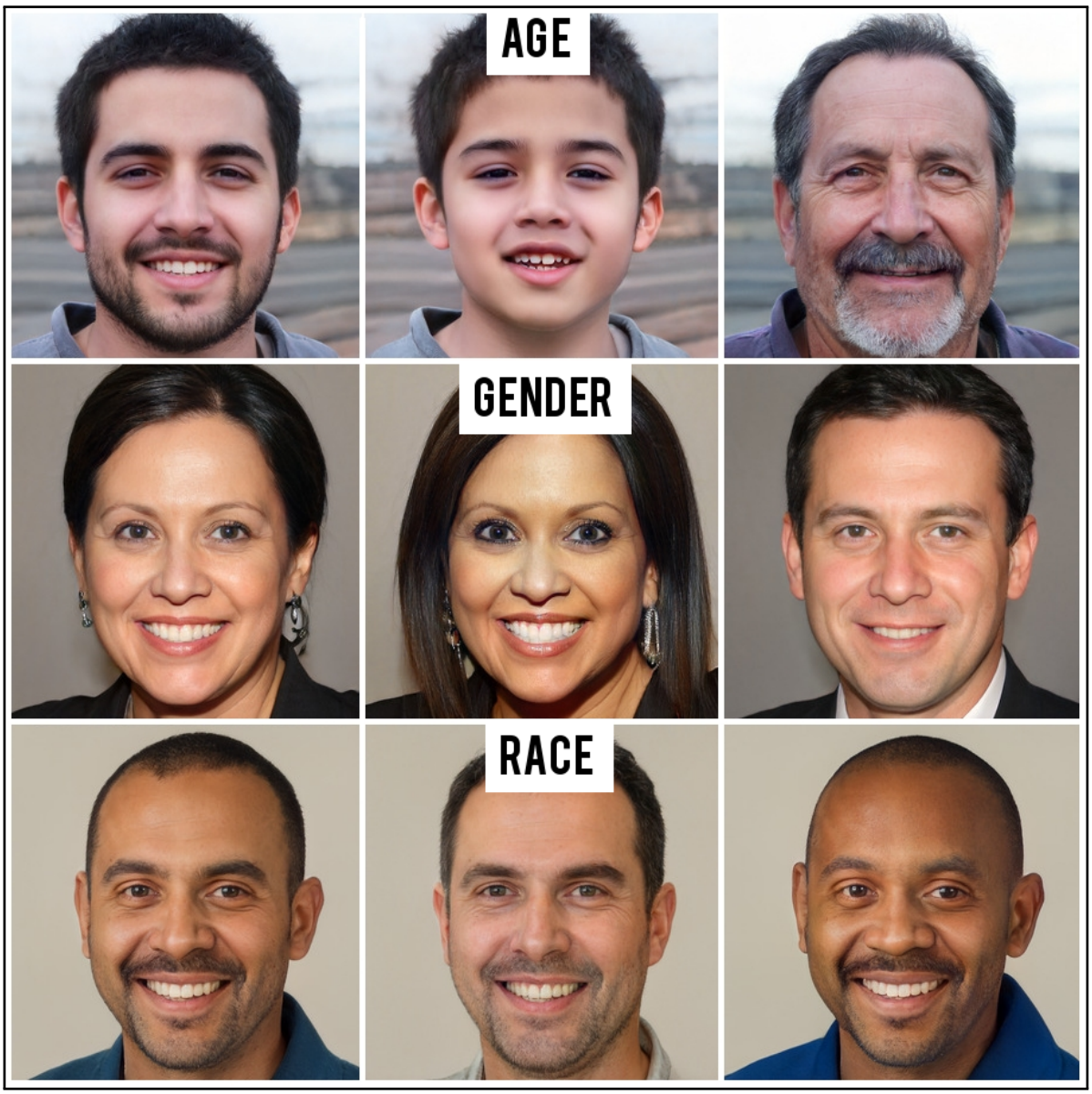


Figure 1.1: Applied "styles" on images using StyleGAN that demonstrates styles and the resulting changes adapted from Karras et al. (2019)

Examples of this in practice is online dating applications namely Bumble, Tinder and Hinge. These free to use applications allow multiple users to connect and interact where interaction mostly starts because of the user's interaction with the images displayed on a profile. This emphasis on the gallery of a user profile creates a unique opportunity where StyleGAN can help cybercriminals fake identities within these apps. Fraudulent use of StyleGAN in this environment can directly lead to an increase in catfishing incidents. Catfishing is the act of deceiving an individual with the use of a specific fictional identity and persona to gain assets that are usually in the form of financial gains or personal information (Chandler et al., 2016). These apps usually have various systems in place to combat such fraudulent actions by requiring account verification.

The verification process requests that a user submit one or multiple images where they pose in specific ways. StyleGAN can successfully counter these forms of verification as an image can be generated wherein a person perform specific poses (Karras et al., 2019). Example images of StyleGAN can be seen in Figure 1.2. The diversity in StyleGAN images is another aspect of the GAN that ensures it can create very accurate images of humans that can go by undetected.

Figure 1.2: Images from the StyleGAN dataset Karras et al. (2019)

Because of StyleGAN's ease of use and availability more fraudsters used it to create false accounts on popular social media platforms since its creation. Facebook took down an undisclosed number of accounts in December 2019 that had reportedly made use of StyleGAN to generate realistic profile pictures for establishing false identities on their site (Gallagher and Calabrese, 2019). This emphasized the possible threat StyleGAN poses to the security of individuals and the IT industry. Because of the threat, StyleGAN poses

the need for a method to detect these generated images is identified. StyleGAN was released to the public in December 2018 with all its packages and source code. This novel approach to GAN's that was developed by Nvidia demonstrated its possible capabilities by the accompanying portraits of convincing human faces. The big leap forward in this technology was the realism brought to the GAN's generated faces dataset that is close to real human beings and not easily detected by humans (Fleishman, 2019; Karras et al., 2019).

StyleGAN was popularised because of a former engineer at Uber, Phillip Wang's website thispersondoesnotexist.com that was released in February 2019. Phillip's main goal when publishing this website was to educate the public on how GAN's work and the dangers that they might pose to the average user. Phillip achieved this by specifically emphasizing StyleGAN and its realistic human face generation capabilities Fleishman (2019). The goal of making more people aware of StyleGAN and possible fake identities was also promoted by another website. Two members of the University of Washington created the website whichfaceisreal.com Fleishman (2019). This website allows users to select an image between a StyleGAN image and a verified human image. This online tool helped users realise that the differences between computer-generated images and real images are only decreasing with the advancement of technology.

The big advancement in StyleGAN that increased the need for this proposed project was the release of StyleGAN2 in February 2020 (Karras et al., 2020). Following the release of version 2, there were improvements in the generation process of these images with this updated version of StyleGAN. StyleGAN2 saw that images were no longer subverted with artefacts or traces left on the image because of the processing method used in the previous iterations (Karras et al., 2020). These traces were easy to identify and clearly showed out of place in the context of the image. With the removal of the traces, usually, in the form of drop-like spots, the difficulty in detecting the generated images increased and similarly the need to detect StyleGAN generated images used maliciously in security verification processes.

By looking at how the technology has been used since its release, the possible use-cases for GAN generated images and the always growing cybercrime industry the possible detection of these images is identified as a crucial function in the 4th industrial revolution. With these identified factors will the proposed project aim to solve the problem of detecting StyleGAN images by using an artificial intelligence approach to solve the problem.

1.3 Research Question

With the identified need for detection of StyleGAN images and the discussed security implications that the invention of StyleGAN and similar methods introduced the proposed project aims to detect these images with the use of a trained neural network. The main research question that this proposed project aims to answer is: How can StyleGAN generated images be detected?

1.4 Aims and Objectives

1.4.1 Aims

The main purpose of the proposed project is to develop a method that can detect fraudulent human faces created by StyleGAN with relative accuracy. Various techniques and approaches to the detection of GAN generated images will be researched and the simplest implemented approach that can still detect these types of images with relative surety will be selected for the artefact.

1.4.2 Objectives

The success of the project will be weighed against the completion of the secondary objectives that have been identified as listed below.

- Perform a literature study on GAN's and specifically analyse the architecture and function of StyleGAN to understand the technology.
- Develop an approach to the successful identification of generated images.
- Develop an artefact that will use the selected method to detect a fake identity.

The successful completion of the above-identified objectives will aid the researcher in satisfying the aim of the proposed research project. The success achieved in the creation of the method of detection will dtermine if the chosen method could consistenctly and in high frequency detected StyleGAN images in a pool of real and fake images. Only if the detection of the images is relatively high in the bounds of the project scope will the chosen method be implemented in the final artefact of this propsosed project.

1.5 Procedures and Methods

This section will describe the research paradigm and the methodologies that will be used to complete the proposed project. By examining the chosen paradigm, methodology and artefact life cycle in an academic viewpoint with specific reference to information systems design and development will the most applicable approaches in these sections be identified and selected. The data that will be captured in the development phase and reviewed in the testing phase will be discussed. Data capture in this proposed project will be the output metrics of the methods of detection performance and the retrieval of the datasets from their respective hosted repositories.

1.5.1 Paradigm

Positivism is a research paradigm that is focused on the world view that "factually accurate" knowledge is gained through the observations made by the observer. In positivistic studies, the research is confined to only the collection of data and the interpretation of this data to gain knowledge and insight into the problem. Positivism requires the researcher to only make quantifiable observations that can directly lead to statistical analysis. With the use of this paradigm, the researcher must reject intuitive knowledge because it cannot be justified by sensory experiences and is thus subjective to the researchers own interpersonal influences. Positivism as a philosophy is justifiable by empiricist views that knowledge stems from a human experience (Collins, 2018).

The positivistic research paradigm is suitable for the proposed project because with the creation of the artefact, the data collected will be examined and an unobjective interpretation of the data is necessary. The data collected will be the results of the artefact's successful identification of StyleGAN generated images, and therefore a positivistic research paradigm will ensure the best method of detection is selected in the context of this proposed projects scope.

1.5.2 Methodologies

Methodologies determine the structure of completion for a specific project. With this proposed project the researcher will study the technology that enables StyleGAN to generate images. To aid in the research process and the development of the artefact the Design Science research methodology (DSRM) will be used throughout the proposed project (Peffers et al., 2007).

DSRM is an information systems specific methodology that focuses on research and iterative design (Peffers et al., 2007). Because the researcher is studying the field and technologies in which they want to solve the specific problem the background knowledge of the problem will be explored parallel to the design of the problem solution. DSRM will enable iterative design and development throughout the completion of the proposed project.

1.5.3 Artefact Life Cycle

With the use of the DSRM, the researcher will amend the project with the acquisition of new knowledge. The artefact will be developed while research is conducted on the technologies required, thus the artefact development will similarly be conducted in increments. The implementation of the Agile methodology for the development of the artefact will be suitable as Agile accommodates changes in artefact scope, planning and incremental deployments of preliminary artefacts (Weiyin et al., 2011).

The artefact that will be developed will aim to detect StyleGAN generated images between a data set of real images of human faces and StyleGAN generated images. The artefact can be hosted online if needed as this increase its compatibility and deployment reach across multiple platforms. Because of these factors, Agile will be the most suitable methodology for artefact development.

1.5.4 Data Capture

The data captured in this project will be collected and displayed to the user in the web application to aid in the demonstration of the artefact, and method of detections success in the finalization phase of the project. Data that will be captured is the number of images used in the training of the models, the number of images that were used throughout the identification testing phase and the accuracy of the chosen identification method. These statistics will help the researcher determine the success of the chosen method of detection.

1.6 Project Management and Project Plan

1.6.1 Scope

The research of the technologies implemented in StyleGAN and the specific architecture of StyleGAN is important for the design of a suitable approach for the detection of these images. The scope of this research is the identification of images, the use of artificially generated human faces, security concerns when determining identities based on profile images and the research of neural network's that will be a possible technology used for the solution to the problem.

Because of the time of StyleGAN3 and 2 individual releases and the initialization of the proposed project the scope of this project will only focus on the detection of StyleGAN1 generated images.

1.6.2 Limitations

Possible limitations that can be faced during the proposed project will hold back advancements that the researcher aims to complete in this study. The limitations thus need to be identified and addressed to ensure the completion of the proposed project within the specified period. The proposed project will make use of neural network's and image processing technologies to answer the research question. These technologies require large computing capabilities for the training of neural network models. This will be addressed by using cloud services instead of traditional hardware. Cloud services provide a more cost-effective approach to large computing needs.

1.6.3 Risks

Possible risks that can be identified that this proposed project is susceptible to include the risk that the scope of the project is not adhered to. The scope defines the boundaries in which this proposed project will take place, therefore careful adherence to the scope will ensure that the research is relevant to the initial project that was proposed. The responsibility of staying within the scope of the project lies in the researcher proposing this project.

1.6.4 Strengths, Weaknesses, Opportunities and Threats

The Strengths, Weaknesses, Opportunities and Threats (SWOT) of this proposed project can be evaluated using a SWOT-analysis table. The risks and limitations mentioned previously are also mentioned when conducting a SWOT analysis. Figure 1.3 shows the SWOT analysis of the proposed project.

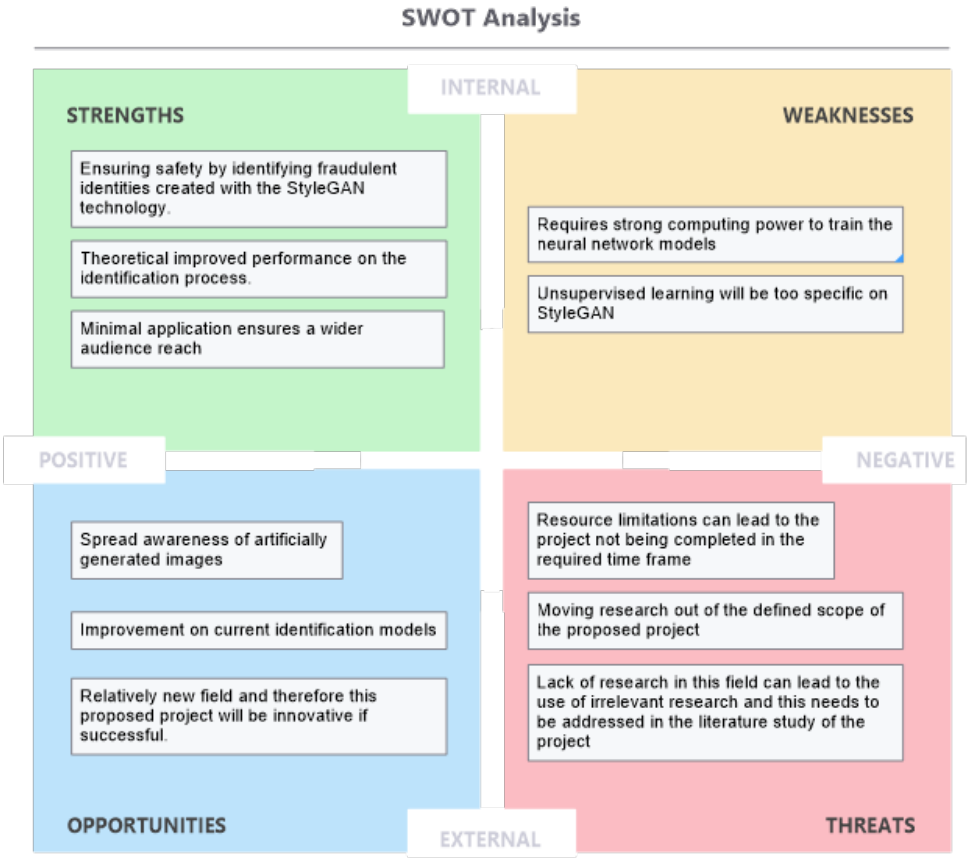


Figure 1.3: SWOT analysis in the Identification of StyleGAN images

1.6.5 Timetable

The proposed project will start on the 16th of February 2021 and will be finalized and completed on the 8th of November 2021. This project will be subdivided into 3 phases each with separate deadlines. The first phase is the research proposal that will be concluded on the 18th of April 2021. The second phase involves research that will require an in-depth analysis of the problem, possible solutions, and the discussion of StyleGAN in a literature study that must be completed on the 13th of June 2021. The last phase is the development of an artefact to practically demonstrate the solution to the identified problem.

Phase 3 of the proposed project will also require a demonstration of the developed artefact and a video demonstration of the project on the 1st of November 2021. The submission of the whole project must be completed and the final documentation will take place on the 8th of November 2021.

The Gantt chart in Figure 1.4 graphically displays the preliminary project planning with the tasks in the required sequential order for the successful completion of the proposed project.

The adherence to the project planning and the preliminary Gantt chart will enable the researcher to effectively divide the tasks into manageable time frames. The prerequisite tasks are required to begin with a dependent task and are displayed graphically in figure 1.4.

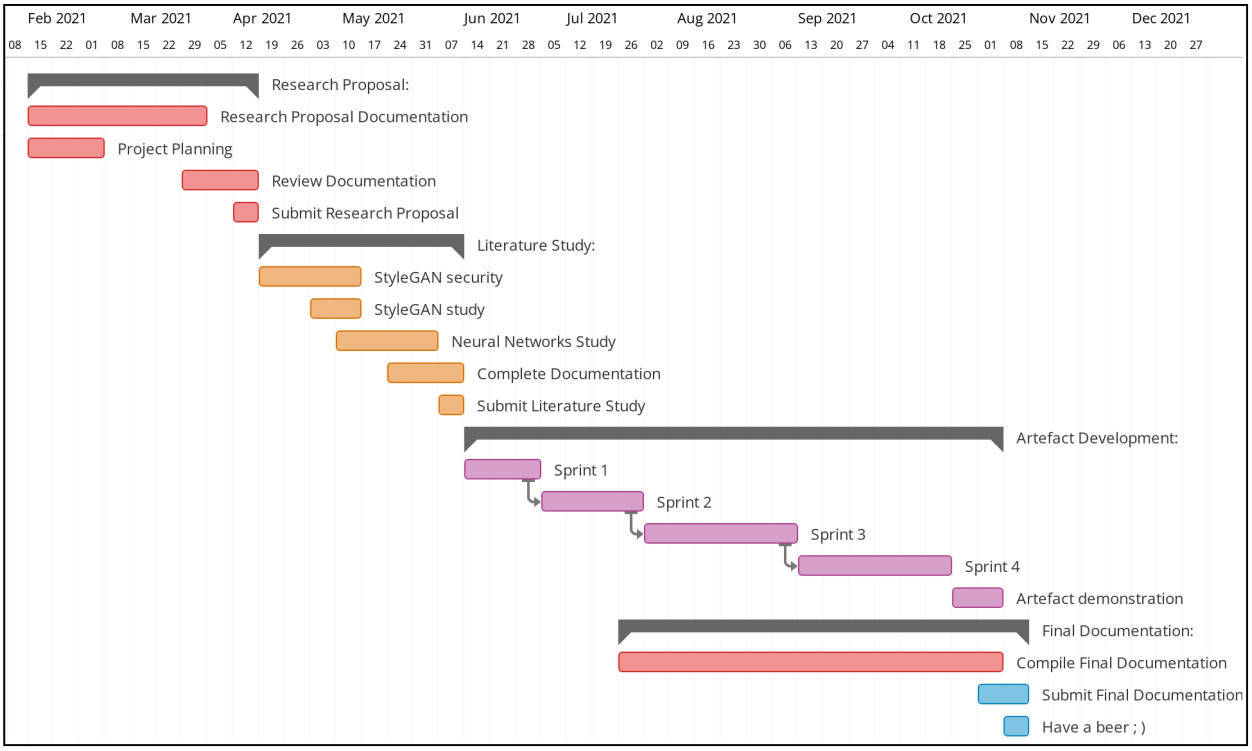


Figure 1.4: Gantt Chart graphically demonstrating the planned schedule of the proposed project

In figure 1.4 time management and planning is clearly shown to be crucial for the completion of the proposed project. The artefact development planning is subject to change based on the research outcomes and findings of the researcher in this proposed project.

1.7 Development Platform, Resources and Environment

The development of the artefact will be conducted in parallel with the research on the subject. Development of the GUI will be conducted at the same time as the literature study. The development of the final artefact and method of detection will only be conducted after the literature study based on the knowledge acquired in that phase. With the knowledge gained in the literature study will the specific method to detect StyleGAN images be implemented into the artefact. The artefact development platforms, resources and environments will be discussed in this section.

1.7.1 Web Application

The artefact will be a web application that can be scaled and hosted through a cloud services provider. The language that will be used for the front-end web application is Python-based using the scaled-down Django framework, Flask. For the implementation of the method for detection of StyleGAN generated images the language used in the back end of the artefact will be Python. The reasoning behind the selection of the above-mentioned technologies is discussed in the following section.

Developing a web-based application allows for further reach and compatibility compared to historical installation software (Murugesan et al., 2011). Developing a simple web application with a minimalistic front end will users more intuitively be able to use the application and will improve complex implementations ease of use (Murugesan et al., 2011). Web apps opened in the browser natively scale to be available on various devices and screen sizes. For the StyleGAN artefact, improved availability will provide the application with a wider reach and therefore more artificially generated images can be detected.

The web application will allow a user to load a new image where the artefact will then classify that image as a StyleGAN generated image or a real human image. The web application must keep track of basic statistics and provide them to the user. The basic statistics that will be provided in the artefact of this proposed project will be the neural network's confidence in its prediction. An uncomplicated design will be implemented where the user can intuitively navigate the web page.

1.7.2 Python

For the backend of the proposed project Python will be used. Python is an intuitive programming language that due to its active community has an abundant set of resources available to use in the artificial intelligence field. Python is widely used for artificial intelligence applications and thus is a suitable programming language to code the chosen method of detection.

1.7.3 Flask

For the front-end of the artefact's web application Flask will be used. Flask is a framework based on the larger framework Django that is a web framework using the Python language. Flask front-end features can use CSS-styling and the use of bootstrap will make the front-end uniform that enables faster development without the need for extensive visual design. Flask will enable the artefact to be used on multiple devices with a lightweight package without the explicit code and design to accommodate those devices. The framework natively scales assets to fit a wide range of devices.

1.7.4 Jupyter Notebook

Jupyter Notebooks is a Docker-like implementation of python coding. In a Jupyter notebook, cells can be run individually allowing for more control when developing neural network's. Most cloud services use Jupyter notebooks when implementing python code on their platforms. Jupyter notebooks are useful in data science and machine learning development because single python cell blocks can be executed. When training large neural networks that require long training times the cell blocks will prove useful because when training needs to be reinitiated for any reason the large code required beforehand does not have to be reinitiated.

1.7.5 Cloud Services

Because of the possible processing resource requirements that the detection of StyleGAN images require, the training of the neural network model will be conducted on the platform and infrastructure that is provided by Google Colab. A hybrid cloud services structure will be used in this proposed project because of the limited resources available.

PaaS and IaaS

In cloud services, there are multiple forms in which the cloud can be implemented for use by an individual or organisation. These implementations can range from Software to Infrastructure offered to the entity from the cloud services provider (Pfleeger and Pfleeger, 2002). For this proposed project there is a need for a platform and infrastructure as the training using large image datasets require processing power. Platform as a Service (PaaS) is a service model where the client develops software by using the languages and tools offered by the cloud services provider (Pfleeger and Pfleeger, 2002). Infrastructure as a Service (IaaS) is the second service model this proposed project will require. In IaaS the cloud services provider offers the use of processing resources and storage to name a few, to the client (Pfleeger and Pfleeger, 2002). A combination of IaaS and PaaS will be required for the completion of the artefact in this proposed project.

Google Colab with Drive

Google Colab is a free service that Google offers to data scientists to use for the development of neural network's. Google Colab workbooks are the development platform in which programming is conducted that is similar to Jupyter Notebook development on local machines. The difference in using Google Colab is the increased system resources of the web-based virtual machine. Using Google Colab for the development in this project will have the benefit of the GPU attached development environment where a Nvidea K80 GPU with 12GB of RAM can be used. Google Drive is a storage solution offered by Google that is integrated with Google Colab. The large datasets can be uploaded to an NWU google accounts drive, where storage is unlimited. The dataset can be mounted in Google Colab using Google Drive.

1.7.6 Git

For version control and deployment, the Git language will be used and the repository for the development of the artefact will be hosted on GitHub. GitHub is integrated with various cloud services and this allows the developer to easily deploy and evaluate the code.

1.7.7 Optuna

The neural network hyperparameter optimization framework Optuna will be used to increase neural network performance by optimizing the architecture of the network and the parameters within the network. Optuna is relatively new and will be studied while implementing the technology on the artefact using the DSRM.

1.8 Ethical and Legal Implications

With the development of this artefact, certain resources will be used to train the neural network. The training requires images that were generated by StyleGAN. For the comparison in the artefact images of real humans will be used. Figure 1.5 shows a sample of the real human images that will be used in this proposed project.

Figure 1.5: Gantt Chart graphically demonstrating the planned schedule of the proposed project

The StyleGAN generated images are available on the official StyleGAN GitHub repository. Included in this repository is trained StyleGAN models and multiple datasets of StyleGAN generated images. The licensing of these images is stated on the GitHub repository and is a Creative Commons license by NVIDIA Corporation (Karras et al., 2019).

For the verified human faces, the preliminary dataset that will be used is the Flickr Faces dataset that was initially used to benchmark StyleGAN. The individual images were published in Flickr by their respective authors under either Creative Commons, Public Domain. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes (Karras et al., 2019).

1.9 Provisional Chapter Division

For this whole project, the following flow of the final document can be expected. These chapters will logically flow to aid in the understanding of the topic at hand and to ultimately ensure a successful artefact.

Chapter 1: Introduction

This chapter will be concluded in the project proposal phase. It will include the research question, the project description and background. It will include the proposed project plan for the entire project.

Chapter 2: Literature Study

This chapter will be comprised of all the necessary research to understand the project and fulfil the project aims and objectives. Mostly in this project will be focusing on the specific workings of StyleGAN to effectively detect fake images.

1. StyleGAN: The history of StyleGAN and direct technologies that lead to this breakthrough in the field of GAN's will be provided as well as the specific architecture of StyleGAN.
2. Neural Network: This project will make use of Machine Learning and neural network's to detect StyleGAN generated images. A study will be completed on the field of artificial intelligence and how neural networks can aid in the detection of images.
3. Hyperparameter Optimization will prove useful in improving the created neural network model to increase the detection accuracy for StyleGAN images. A Study on hyperparameter optimization methods will aid in the creation of a hyperparameter artefact.

The study of these three sub-topics will provide the relevant background and base knowledge to develop a successful artefact and will provide the relevant foundation to improve and expand on this project in the future.

Chapter 3: Development of the artefact

Chapter 3 will apply the chosen methodologies that were identified and discussed in Chapter 1 to enable the successful development of the proposed project's artefact. This chapter will document the artefact development phase including unfamiliar problems that are identified within the development stage. The success of the artefact will be compared to the Aims and Objectives of Chapter 1 and if they are met.

Chapter 4: Results

The results of the Artefact will be introduced in this Chapter and the successful identification of StyleGAN images will be determined. The testing in this Chapter will identify the success of the artefact with the comparison in Chapter 4.

Chapter 5: Reflection

Chapter 5 will summarize the entire proposed project and conclude if the problem was solved with the successful completion of the objectives of the proposed project that allowed it to fulfil the aim of the project. The limitations that impeded the proposed project will be discussed in this section. The future expansion of the project will be discussed and explained in the context of the limitations faced.

1.10 Summary

The creation of StyleGAN images has sprouted possible security issues where falsified images can be used to create false profiles on the internet. The detection of these images will prove useful in the growing cybercrime world. StyleGAN images can be detected by implementing a machine learning approach to the problem. This project will use the DSRM in combination with Agile to develop an artefact that will detect StyleGAN images. By analysing StyleGAN and NN in a literature review can the development of the artefact commence based on the findings in the literature study.

Chapter 2

Literature Study

The aim of this proposed project is to develop a method to detect images created by a style-based generator architecture for generative adversarial networks in a set of artificial images (generated images of human faces contained in the StyleGAN dataset) and authentic images (Flickr-Faces-HQ dataset of human faces used as a benchmark for StyleGAN). Before the aim of the project can be satisfied, a literature study is required on neural networks and StyleGAN. The basic structure of neural networks and the different types of neural networks will be researched and discussed in context with the StyleGAN technology, and the detection of StyleGAN generated images. Current detection methods will be investigated and examined to evaluate their workings and relevancy towards a StyleGAN application. In conclusion, all knowledge acquired from the literature study will be used when creating the artefact to detect StyleGAN generated images.

2.1 Neural Networks

In recent times, neural networks have been regarded as a fast-growing field offering powerful tools for most types of problem-solving (Albawi et al., 2017). The increased use results from the neural network's capabilities to function even with large data sets as input effectively. Therefore, a study into neural networks is required as it is a tool to use in the detection of StyleGAN generated images because of the large datasets accompanying StyleGAN.

2.1.1 History of Neural Networks

Artificial intelligence is a modern growing field within information technology rooted in historical discoveries leading to new advances. Artificial intelligence has been in the development stages since the mid-20th century. However, early on, most advances made in artificial intelligence were developed in mathematics and computational model theory (Müller et al., 1995).

Warren McCulloch and Walter Pitts initialised the now big field of artificial intelligence when they proposed a new general theory in information processing that artificial neurons can mimic the neurons present in the human brain (Müller et al., 1995). Müller et al. (1995) also notes that the neurons Warren McCulloch and Walter Pitts proposed were more simplified than biological neurons and still promised reliable computational power. The artificial neuron that could be implemented in a network to mimic a single neuron cell in the human brain and mimic the whole brain as a network is the foundation of early artificial intelligence and machine learning theory.

The next significant advance in this field came in the 1960s by researchers Caianiello and Rosenblatt. This advance resulted from focusing on two aspects of Artificial Neural Network's (ANN): 1st being the aim to mimic their biological counterparts in their design and the 2nd is that different structures proposed different advantages (Müller et al., 1995). This difference of structure leading to a difference of perception is present in nature, where mammals all possess a biological brain, but the shape and network design allow mammals to think differently based on their specific computational needs(Müller et al., 1995). Rosenblatt coined the differing structure of ANN's as the perceptron of the network, and in modern artificial intelligence theory, the perceptron is called the perception of the neural network.

The historical, theoretical initialization of neural networks is the foundation of modern neural networks. Modern neural networks' development focuses on practical applications in modern times. Neural networks are still relatively new in artificial intelligence, with theory founded in historical mathematics

2.1.2 The Function of Neural Networks

Neural networks are present in most forms of biological computation. For example, the human brain is a neural network of neurons that allow us to compute our daily tasks. An ANN mimics the human brain in its structure (Krenker et al., 2011). Like a human brain thinking through the impulses sent and received between biological neurons, an

ANN sends and receives input and output through the artificial neurons in the network (Krenker et al., 2011). The neurons that form the network allow neural networks to imitate a biological brain's basic structure and computation.

An artificial neuron is a single node within a neural network. This neuron is an independent node residing in a neural network that receives data and applies the neuron's mathematical model to this input. The output is the result of a calculation and the operation of activation and deactivation of the node.

Krenker et al. (2011) states that the neuron structure consists of three separate stages. The first stage applies the weight to the input with multiplication. The second stage is a sum function of all the previous stage weights and biases applied to the initial input. The final stage of the artificial neuron determines if the neuron should be activated or deactivated (Krenker et al., 2011). Artificial neurons activated or deactivated artificially enable the neural network to "think" and adaptively apply its computation on inputs.

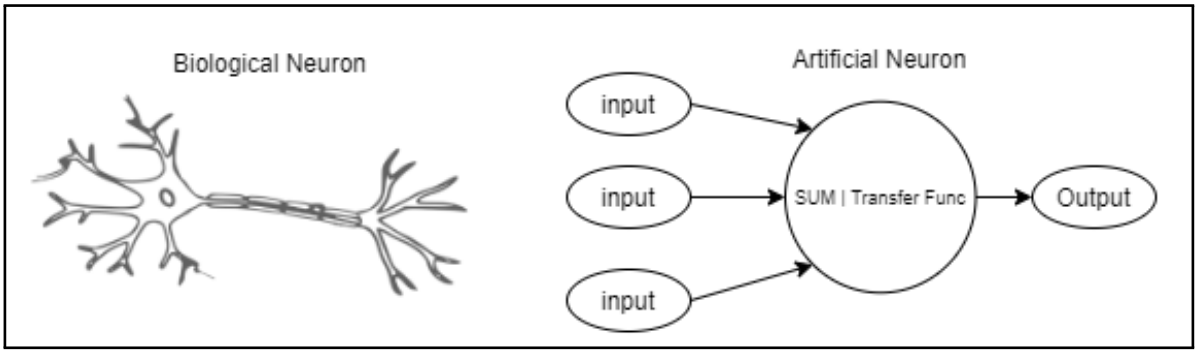


Figure 2.1: Biological Neuron and Artificial Neuron Krenker et al. (2011)

Figure 2 shows the similarities between the biological neural network and the ANN. The basis of biological neural network for the design foundation of an ANN is

The set of neurons functioning together to form a network is called a neural network. In a biological brain, the whole cellular neurons connected in the brain and artificial intelligence are the ANN with the artificial neurons working together. ANN's are capable of processing information of real-world problems that requires a more complex approach because they can distribute their neurons in a non-linear, parallel structure (Krenker et al., 2011). Multiple neurons simultaneously are the neural network structure, and neurons do not have to flow their output into the next.

Figure 3 illustrates a primary ANN with connected neurons that mimics the structure of connected biological neurons connected inside the human brain. Each neuron within the structure of the ANN in Figure 3 can be compared to a human single brain cell in

the biological brain structure. The artificial neuron present in the network only apply basic computational calculation on the data, and as output, the node only activated or deactivated itself.

2.1.3 Neural Network Learning Paradigms

Neural networks can learn in different ways, and these differences pose their respective advantages and disadvantages—the manner of learning in a neural network the network's learning paradigm. The three different learning paradigms for neural networks used in the training of neural networks are supervised learning, unsupervised learning, and reinforcement learning (Krenker et al., 2011). The learning paradigms for neural networks are also specific to the type of data the network use in training. Concerning image classification and computer vision, only two of the three learning paradigms apply, namely supervised learning and unsupervised learning (O'Shea and Nash, 2015). Different neural network topologies use different learning paradigms. The specific choice of learning paradigm is essential in the implementation of the detection method of StyleGAN images.

Supervised Learning

The supervised learning paradigm sets the parameters of the neural network based on the training data set it receives. The critical difference in supervised learning is that the input data is labelled before training the neural network (Krenker et al., 2011; O'Shea and Nash, 2015). The already labelled input is then compared to the output of the neural network to determine its error and accuracy. Training is applied to the network based on comparing the output the network created against the labelled input. If this paradigm is applied to the StyleGAN image detection problem, the training set of data consist of a set of images of human faces and a set of StyleGAN generated images. This input training set requires parameters or labels that identify a single image as either a human face or a StyleGAN generated image. The prediction made by the neural network will then be validated against these know classifiers on the images, and training will take place based on how the neural network predictions compare to the valid identifiers. In deep neural networks, supervised learning also deals explicitly with the labelled data. The advantages of supervised learning applied in deep neural networks such as CNN and GAN are that training can be conducted without initial knowledge about the differences between two different data inputs. The fallback to this advantageous method is that when there is outlying data present within the dataset, An image of a dog in the set of authentic images, the training can overstrain the boundary of decision (Alzubaidi et al., 2021).

The supervised learning paradigm is subdivided into two separate fields determined by the specialisation required while the neural network learns in the training phase. Semi-supervised learning and self-supervised learning selection are determined by how the input data is labelled to provide the neural network with information (Zhai et al., 2019). For a supervised learning approach, when aiming to identify StyleGAN generated images, the subsections of supervised learning must be evaluated in context to this project.

Semi-supervised learning is a good choice of training algorithm if the input data is both labelled and unlabelled. In most cases, the learning algorithm assumes the label of input data if the data originated from the same distribution (Zhai et al., 2019). In the case of analysing StyleGAN images, the 2 labelled data sets are authentic images and generated images. When selecting a semi-supervised learning algorithm, the initial labels in the dataset must be standardised, and the data set is then altered that only a portion of the labels is kept with the data set. The algorithm treats the rest of the dataset as unlabelled data (Zhai et al., 2019). Semi-supervised learning was based initially on different neural network architectures, and one prominent algorithm used was the GAN that is also the basis of StyleGAN. An advantage of semi-supervised learning is that it applies inductive learning through generalisation, mapping the inputs to the outputs classify the data that have the most significant impact on the output of the network.

Self-supervised learning uses only the unlabelled data to formulate mundane tasks within the network. It is commonly used to label datasets that have no classification of data (Zhai et al., 2019). Self-supervised learning algorithms can be implemented on the StyleGAN identification utilizing it labelling the dataset. The input dataset can then be a combination of StyleGAN generated images and authentic images that the self-supervised learning algorithm will then aim to solve by labelling the data as either generated by StyleGAN or is a unique actual human image. One problem with implementing this learning algorithm is that the data set containing StyleGAN images and authentic human faces are very similar. Differences are minimal between the two images, and Self-supervised learning might struggle to distinguish between the two different images, and miss labelling might occur.

Unsupervised Learning

The difference with the unsupervised learning paradigm compared with supervised learning is that with unsupervised learning, the data does not have the added information that can aid the neural network in determining its correctness of prediction. The unsupervised learning paradigm entails learning based on a set cost function and minimising the goal's cost (Krenker et al., 2011). With a StyleGAN application, this paradigm will not necessarily help the neural network to learn effectively. This

inefficiency results from an image's initial problem, either an actual image or a fake image. And a cost function to how fake an image might not necessarily lead to the neural network deciding that the image was generated using StyleGAN. This particular case is substantiated by the artefacts embedded in StyleGAN images. In specific cases, the neural network can receive a StyleGAN generated image close to an actual image with only one unique artefact created by the StyleGAN inefficiency. The cost function might still pass that image as an actual image, yet a person will quickly identify the artefact in the image with ease. More on StyleGAN artefacts will follow further in the StyleGAN analysis.

The different learning paradigms discussed each poses their own set of advantages and disadvantages. In the context of identifying StyleGAN generated images, supervised learning could be the chosen learning paradigm, more specifically semi-supervised learning, because of the two labels present in images used to train the neural network. On the other hand, self-supervised learning will not be used because of the minor differences between a StyleGAN image and an actual human face image.

2.1.4 Hot and Cold Learning

Hot and Cold learning is one of the most straightforward approaches to determining the optimal weights for machine learning problems. In a StyleGAN problem hot and cold learning will be randomly guessing the initial hyper parameters and changing them in training to increase the accuracy of the neural networks prediction.

Hot and cold learning is the process of increasing and decreasing the weights after a prediction and then training the model again. In theory, the continuous repetition of this process will lead to an error value of 0. Based on the increased or decreased error value, the changes in weights should either increase or decrease. (Trask, 2019) However, hot and cold learning is not efficient. A developer must repeat a process manually numerous times until they finally stumble on a perceived perfect combination of weights. This implementation also does not ensure that optimal values are found. A developer might start by changing a parameter to its "*optimal values*" and then changing another to the previous and ultimately closing in on a false optimal set of parameters.

Hot and cold learning is a simple form of machine learning that is not optimal and might not get the best values for training a model. Hot and Cold learning falls behind because its implementation may lead to false positives where changes in the hyper parameters leads to reduced prediction values and show that the data is optimal yet there may be an even better set of parameters. However, it is useful when implemented on minor scope problems and in the initialization in developing a neural network where hyper parameter

optimization can improve the mode further. Hot and cold learning will aid in the understanding of hyperparameters and how it connects to the machine learning model and improvements in training.

2.1.5 Neural Network Architecture

The structure in which the artificial neurons are presented within a neuron network is the neural network architecture. There are multiple different neural network architectures with different benefits and disadvantages in specific practical applications. StyleGAN is a generative adversarial neural network that applies specific styles to create a new unique image (Karras et al., 2019). A technique of identifying neural network generated images such as those generated from ProGAN, StarGAN and Deepfakes focused on the shared base convolutional neural network architecture of these technologies (Wang et al., 2020). Therefore GANs and convolutional neural networks are identified as relevant architectures that require deep analysis in this study to ensure a thorough understanding of these architectures that will be interacted with in detecting StyleGAN images.

Convolutional Neural Network

Neural networks perform exceptional at identifying patterns hidden in different large datasets, but specific neural network architectures perform better than others when implemented to detect these patterns within specific data set consisting of different data types (Liu et al., 2017). Convolutional neural networks (abbreviated as CNN) commonly used for computer vision are great neural network architecture options for identifying patterns in image datasets (Albawi et al., 2017; Yosinski et al., 2015). When trying to detect StyleGAN generated images, a large dataset containing images generated through StyleGAN and a large dataset containing images of humans will be used to train the neural network. A CNN could be an appropriate neural network architecture to use in the detection method because of the specific large image dataset required to solve this problem.

CNN's are similar to the more basic ANN's, with the only difference being the advances a CNN has towards image classifications and pattern recognition within computer vision (O'Shea and Nash, 2015). In a CNN, the primary neuron within the network improves throughout learning, and the network still takes a single weight and apply it throughout. CNN's consists of layers that interact with the raw image data and declassify it into raster data where subsequent layers preside over the fragmented raster's consecutively. CNN's are chosen for image application because ANN's struggle with the computation power needed when attempting image classification (O'Shea and Nash, 2015).

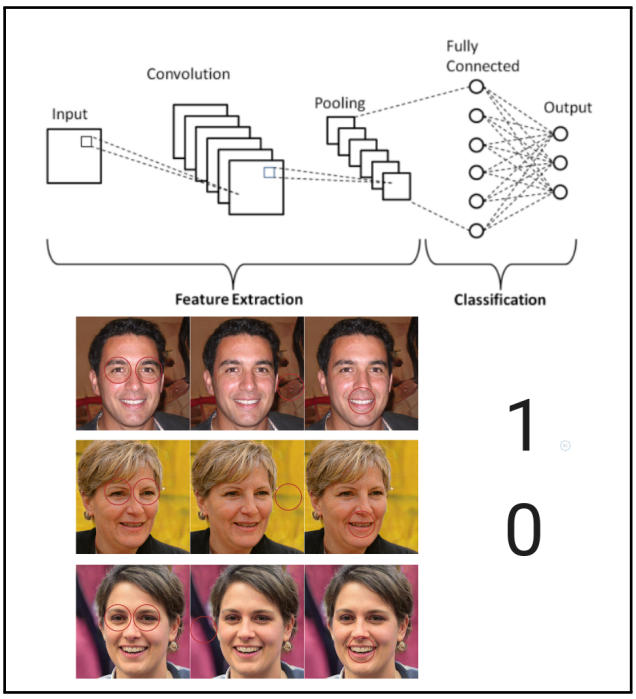


Figure 2.2: Possible Architecture of CNN applied to the StyleGAN problem (Karras et al., 2019; O'Shea and Nash, 2015)

The network architecture of a CNN is divided into three dimensions categorised as the convolutional layer, pooling layer and fully-connected layer (O'Shea and Nash, 2015). The dimensionality created by the stacked layers sets CNN's apart from ANN's in image classification. The first layer in a CNN distinguishes it from the standard ANN and produces improved performance when it is implemented for image classification.

When the input is passed through this layer, the convolution applies various filters on the data to activate two-dimensional maps (O'Shea and Nash, 2015). These 2D maps determine if features are present in the image based on pixels activated when filtered on the activation maps. The convolutions within this network can get large in dimensionality, and the pooling layer is responsible for reducing the complexity of the calculated data (O'Shea and Nash, 2015). The function of the pooling layer to reduce the dimensionality of the network can be detrimental to the data because any dimensional reduction in the architecture of the NN further reduces the data dimension simultaneously. Finally, neurons gather the data directly from nodes in the previous layer and, without connecting the preceding layers, conclude the neural network (O'Shea and Nash, 2015).

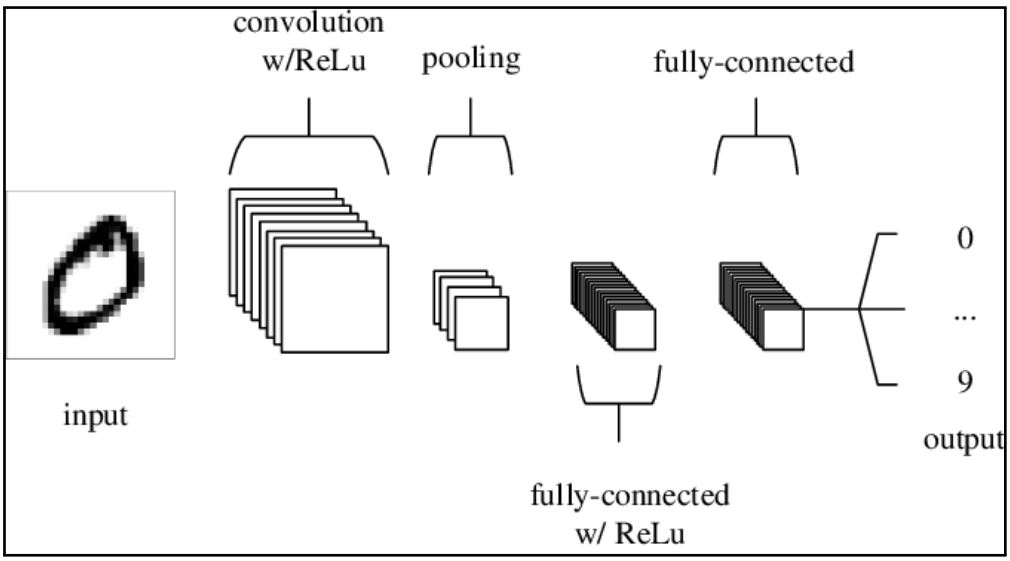


Figure 2.3: CNN used in the classification of digits O'Shea and Nash (2015)

Figure 2.3 visualizes the hidden convolution layer, pooling layer, and the fully connected layers within a CNN. Various iterations of the layers pose different advantages and disadvantages in a practical application.

The benefit when using a CNN on an image processing problem is the improved accuracy in detecting objects within the image (Albawi et al., 2017; Alzubaidi et al., 2021). The activation maps present in a CNN enables the neural network to robustly identify objects within an image through the localisation of the raster data. CNN share weight between its parameter, and this reduces the number of required nodes in training. The reduction in training nodes improves CNN generalisation and reduces overfitting created by the training process. The localisation and object detection of CNN requires extensive calculations, and the process is computationally expensive. This drawback means that without the necessary graphical processing hardware, the training of the CNN will take a long time compared to most other neural networks (Alzubaidi et al., 2021)

Generative Adverserial Networks

Competition increases the performance of sports athletes, students, and businesses (Burguillo, 2010; Hays et al., 2009; Medvedev and Zemplinerová, 2005). Neural networks can train themselves with the appropriate datasets as identified earlier. With the analogy that a neural network aims to mimic the human brain in its structure, the assumption can be made that competition might increase a neural network's performance. This assumption that competitive neural networks competing against one another led to the formulation of GANs (Creswell et al., 2018).

GAN's use a discriminator and generator to produce two neural networks competing against one another (Creswell et al., 2018). The result of pitting two networks to compete against each other leads to the network being able to improve itself more effectively and further increase application possibilities with these types of networks (Goodfellow et al., 2014). The generator constantly tries to fool the discriminator and the discriminator, in turn, try to identify when its input originated from the generator. The generator can be characterised as a criminal trying to falsify identification documents and the discriminator as a customs officer checking a passport. The criminal constantly tries to fool the customs officer. When he succeeds, the officer remembers the fake passport he missed in his identification of fake passports, and the process starts again. The officer learned from his mistake and can in future detect fake passports better. When the criminal fails, the process restarts and similar to the officer the criminal also learns from his mistakes and adapts how he generates fake documents (Goodfellow et al., 2014). The generator constantly creates, and the discriminator constantly describes. This cat and mouse game of fooling and detection in training is how a GAN evolves into the robust networks present in the field of fake images created by neural networks.

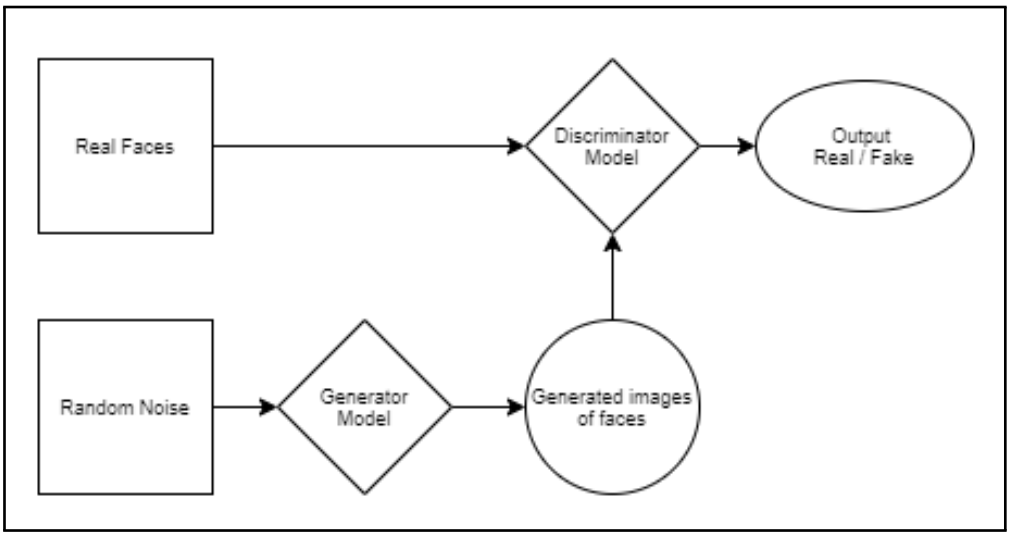


Figure 2.4: GAN architecture adapted from Creswell et al. (2018)

A demonstration of how the structure of a GAN allow the discriminator to be critical of the generator output is evident in Figure 2.4. The advantage brought on by this neural network architecture is the improved data processing performance. The GAN can process images with sharper resolution than the standard ANN that require images with reduced sharpness for model integrations (Goodfellow et al., 2014). The main disadvantage of implementing GAN is increasing complexity to keep the generator and discriminator synchronised while training the models.

StyleGAN and StyleGAN2 are GANs with an improved generator that applies specific "styles" on images, and the combination leads to the generation of new images (Karras et al., 2019, 2020). A convolutional GAN is a GAN implemented where the generator creates images with convolutional neural network architecture. This combination of the GAN checking against itself while creating images with the architecture of CNN enables improvisation in image generation (Karras et al., 2019; Wang et al., 2020).

2.1.6 Activation Functions in Neural Networks

Activation functions is a crucial part of neural networks and their processing capabilities. Without the presence of activation functions in the nodes of a neural network, the learning of the network will be reduced and the maps between the input and the output will not reach the required complexity. In this literature study an analysis on activation functions, why they are necessary and the different types of activation functions available must be conducted.

Activation functions are crucial for neural network learning

Activation functions are the processes on the nodes within the layers of neural networks that allow information to be derived from the data in specific ways. An activation function determines in what way the node will set itself on/off and allow its weight to influence the output of the network. Activation functions are used in an ANN to transform the input signal on a node to the output signal and sequentially add the output of the previous layer to the input of the next layer Sharma et al. (2017).

The inputs and weights are calculated first and then before the output is sent to the next layer the activation function is applied on the node (Sharma et al., 2017). Accuracy within neural networks can fluctuate greatly and is influenced by the number of layers within the network and the types of activation functions used. The types of activation functions within the neural network however has a more significant influence on the accuracy of the prediction of the neural network Sharma et al. (2017). There is no clear way to determine the best number of layers a neural network architecture must consist of, but there is a clear consensus between data scientists that a minimum of two layers must be used Sharma et al. (2017).

In neural networks, there are different types of activation functions but the most common set of these functions is the non-linear activation functions (Sharma et al., 2017). In neural network activation functions, there are boundaries present and these boundaries describe the type of activation function a specific approach consist of, in a linear

activation function this boundary is linear (Sharma et al., 2017). Because of these linear boundaries in linear activation functions, the neural network will only be able to change its perception of the data in linear increments. The problem however faced with these activation functions is that real-life scenarios and problems the errors present consist of non-linear characteristics (Sharma et al., 2017). Therefore data scientists opt for non-linear activation functions over linear activation functions in their functional neural network implementations.

In the development of the artefact the use of non-linear neural networks will ensure that complex information will be processed from the initial dataset. By adding non-linear activation functions to the neural network the output and steps toward the output will be non-linear in the result.

The most important aspect of using activation functions is that the functions are differentiable so that backpropagation optimization can be implemented. When backpropagation can be implemented with the use of gradient descent will the neural network have the capability of calculating the errors and losses based on the calculated weights its uses within its layers (Sharma et al., 2017).

Different types of activation functions

Different types of activation functions can be used and implemented on the layers within the neural network. The type of activation function that can be used depends on the data set and properties present in the data, and the output required from the data. As an example for binary image classification the layers in the network will have to consist of ReLU activation layers, yet the final output layer must be a Sigmoid activation layer.

Table 2.1: Different Activation Functions in Neural Networks (Sharma et al., 2017)

Linear	Sigmoid	Tanh	ReLU	SoftMax	ExpoLU

In the review on activation functions it is apparent that activation functions is a crucial aspect of neural networks and the successful output that neural network can present. Activation functions have a bigger impact on the prediction accuracy of a neural network than the number of layers present within the neural network. There is still some guessing involved into which functions will result in the best accuracy predictions of the neural network but to some degree, there is a clear consensus as to what function will work best with specific types of data. Image classification is improved with the use of ReLU functions throughout the neural network layers and a Sigmoid activation functions as it finals output layer.

With the evaluation of CNN and GAN, an understanding of neural networks in the context of the aim of this project was conducted. CNN provides enhanced capabilities in the generation and classification of image data. While GAN provides specific alterations to images and specialized focus on own output validation and improvement. For the neural network training on a dataset containing authentic images and StyleGAN generated images, supervised learning is the appropriate choice. The architecture and learning paradigms choices for neural networks should be influenced by the type of application and datasets used within training, and because of this, a CNN neural network with semi-supervised learning will be used in the development of the project's artefact.

2.2 StyleGAN

StyleGAN brought new developments in applying styles on images and changing these styles to morph images into new unique style based changes (Karras et al., 2019). Fraudsters are empowered to better create fraudulent identities by applying these styles to images of faces. A fraudster can use StyleGAN to apply styles on their faces to change how they are perceived at checkpoints and more. StyleGAN still introduced advantages to the field; however, the detection of these images is necessary in today's media-centric world. The technology of StyleGAN generated images must first be understood to create a successful implementation technique to detect the generated images.

The StyleGAN addition to the vast sets of different neural networks implemented on different problems originated sequentially with the advances made in the fields of convolutional neural networks and GANs. Because of the computational advances of CNN discussed earlier StyleGAN improved in its generation of new images. StyleGAN2 further improved on this by reducing artefacts brought on in combining styles using the GAN from StyleGAN (Karras et al., 2020).

2.2.1 StyleGAN Architecture

The structure of StyleGAN differentiates it from normal GANs and allows for the combination of "styles" to create new images of non-existing humans (Karras et al., 2019). StyleGAN initializes with an insertion of the base image and input parameter of what styles need to be applied to that image, such as age, sex, and race. Applying all these parameters in the neural network leads to an image being created that changes the styles in the base image. StyleGAN architecture altered the architecture of the generation model to allow for more control over generating with different styles.

Figure 2.5 illustrates the original StyleGAN network architecture and compares the StyleGAN architecture to the architecture of a traditional GAN. The mapping of the styles added to the architecture of traditional GANs present and a clear indication of how StyleGAN can apply the styles within images can be seen.

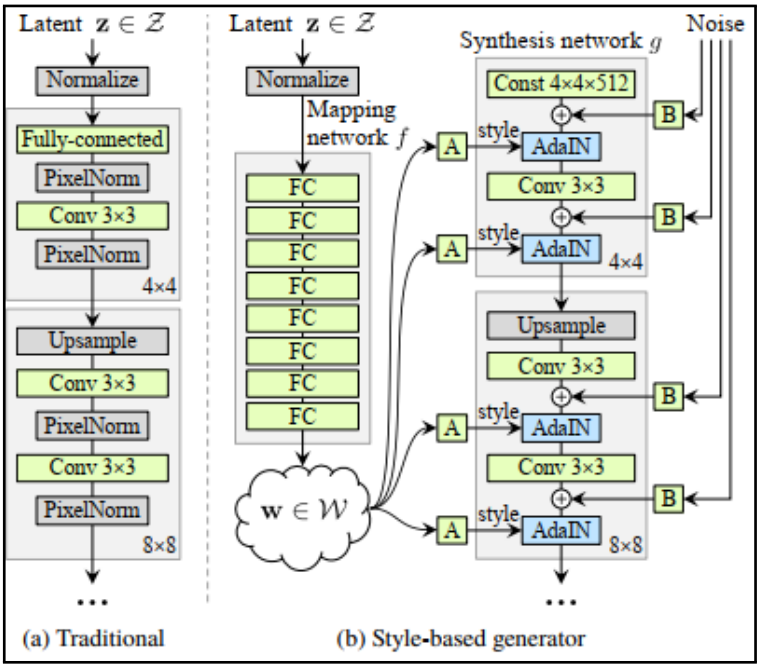


Figure 2.5: Network architecture of StyleGAN vs Traditional GAN's (Karras et al., 2019)

Differences in authentic images and StyleGAN images are minimal. The most noticeable differences are the artefacts that StyleGAN generates due to deficiencies within the StyleGAN technology. These artefacts vary from rain-drop blotches in images to hair strands not showing realistic definition (Karras et al., 2020). In the second iteration of StyleGAN, the artefacts were addressed with minor improvisations but remain present in StyleGAN2 images (Karras et al., 2020).

Figure 2.6 illustrates artefacts present in StyleGAN generated images. This water drop effect present in these images could be more accessible for humans to identify than for a neural network due to the generalization of raster data (Karras et al., 2020).

Figure 2.6: Artefacts present in StyleGAN generated images adapted from Karras et al. (2020)

2.2.2 Detection of CNN Generated Images

To detect StyleGAN generated images, the discriminator of the initial neural network can, in theory, be isolated and used to detect images created by the generator. The generator recurrently aimed to fool the discriminator, and the discriminator constantly adapted to the generator (Karras et al., 2020). The isolation of the discriminator is an invalid approach to detection because of how learning takes place within the creation of the StyleGAN network and when stoppage occurred in the training phase of the initial technology. The discriminator stops training before the generator, and thus an isolated discriminator will not be able to detect the output of the improved generator. In StyleGAN, the discriminator is not available to the public and to isolate it would require retraining of the GAN with the initial datasets of Flickr Faces images. A basic technique of supervised training to detect CNN generated images was implored by the researchers, and their results proved successful in detecting generated CNN images (Wang et al., 2020).

Table 2.2: Results of Wang et al. (2020) detecting various CNN's generating images

CNN-image generator	Detection accuracy of (Wang et al., 2020)
ProGAN	98.8%
StyleGAN	99.6%
BigGAN	66.4%
CycleGAN	88.7%
StarGAN	87.3%
Deepfake	58.1%

Wang et al. (2020) proved that by employing a neural network to train labelled images as either real or faked, CNN-generated images could be detected. Table 1 results from their application of different image generation or changing neural networks. The success that Wang et al. (2020) achieved demonstrated that a semi-supervised neural network can detect CNN-generated images. Therefore a simple neural network approach is validated, and in the context of detecting StyleGAN images, a similar approach will be taken.

The basic architecture and structure of the generator of StyleGAN discussed previously gives more understanding of how StyleGAN can create such precise replication images of human beings. The success of a previous form of detection was evaluated, and StyleGAN images can be detected. By evaluating the work conducted by Wang et al. (2020), a primary neural network with semi-supervised training is identified as a solid foundation for further detecting StyleGAN generated images.

2.3 Summary

Through this literature study, research and developments surrounding StyleGAN, neural networks, and different network architectures and learning paradigms were evaluated and discussed. CNN can create or detect images utilizing the strong computational power provided. GAN can change or initially create derived images, and StyleGAN can implement further control over the changes in the generated images. The types of neural networks involving StyleGAN and other surrounding technologies were evaluated, and an approach was discovered. A simple neural network utilizing semi-supervised learning will train on StyleGAN generated images and authentic images of human faces contained in the Flickr faces dataset will it be possible to detect these generated images as initially discovered by (Wang et al., 2020). The knowledge gained throughout this literature review will be used to develop the artefact to detect StyleGAN images. The development of the final artefact that detects these images will form the third chapter in this project

Chapter 3

Development of the Artefact

The problem of StyleGAN generated faces aiding in the malicious creation of false identities confirmed in the crackdown of Facebook profiles as shown by Chandler et al. (2016) and in the new addition of StyleGAN3 where the creator Nvidea also emphasized the problem and research aiming to detect these images (Karras et al., 2021).

In this projects literature study it was identified that a machine learning deep neural network approach will be the simplest solution to the problem. A neural network that can identify images as either real human images or images generated by StyleGAN with relative accuracy as a method for detection must be implemented in an easy to use front end for the aims of this project to be satisfied. The proposed method of detection should be the easiest solution to the problem as substantiated by Rasmussen and Ghahramani (2001) with their findings concluding that a simple neural network is usually the best neural network.

3.1 Artefact Description

The artefact for this proposed project can be divided into two main aspects namely the neural network model that will train on 2 separate datasets to classify new images between the characteristics of the data classes it learned in trained. The second aspect is the minimal front-end that will allow users to easily interact with the neural network model and receive simple output on the identification of their uploaded images. The neural network was created using the python programming language in a Jupyter notebook that was compiled and executed on the Google Colab virtual cloud-based environment.

To create the neural network popular machine learning packages in the python language was used namely Keras and TensorFlow. The neural network was evaluated and determined to be sufficient, but improvisations could be made using the new technology Optuna for hyperparameter optimization. The neural network was then implemented in a minimalistic front end web app using the python web framework Flask.

3.2 Artefact Life Cycle

The development of the artefact was conducted with the use of DSRM and an Agile combination as stated in Chapter 1. The DSRM methodology enabled new technologies to be used and implemented in the artefact whilst simultaneously learning about the technologies. An example of how DSRM aided in the development of the artefact is the implementation of Optuna. The new technology Optuna helped increase the performance of the CNN exceedingly and because of DSRM, Optuna could be implemented and studied at the same time. The combination of Agile with DSRM aided in the development of the artefact, specifically in regards to the machine learning implementation. Machine learning and Neural networks require data and time, and with the use of Agile sprints. The dataset could be prepared and finalized in a small section of the total artefact development life cycle. The next step was implementing the neural network and optimizing it, and because of the use of sprints in this project these topics could be subdivided into two separate sprints. The fragmentation of the total workload allowed for the successful completion of the artefact.

3.3 Description of the Development of the Artefact

As stated in the methodologies used in Chapter 1 and the Artefact Life Cycle the development of the artefact in this project was subdivided into Sprints. Each sprint handled a subset of tasks to allow for a manageable section of development. This partitioning of the overall workload aided in the development of the artefact and the subsequent testing of the implementations to ultimately ensure a successful project compared to the initial project aims and objectives. Therefore the overall artefact development will be discussed in terms of these sprints. The software used and implemented as well as the final implementation of the method of detection will be structured out in these sprints.

3.4 Sprints

In the development of the artefact, the overall workload was subdivided into 4 separate sprints. In the 1St Sprint the datasets were retrieved and compiled and an initial neural network was created. The 2Nd Sprint entailed training the neural network and evaluating the accuracy of the model in detecting StyleGAN generated images. The 3Rd Sprint used the results of the evaluation in the 2Nd Sprint and it was determined that improvements could be made. Optuna was used to improve the neural network model. The 4Th and final sprint completed the method of detection by implementing the neural network into a minimalistic front-end web application using the Flask framework.

3.4.1 Sprint 1

In the 1St Sprint of the development of the artefact the StyleGAN dataset was downloaded from the official StyleGAN GitHub repository where various datasets are included of StyleGAN generated images. For the images of real human faces, the FlickrFaces dataset was retrieved from the FlickrFaces GitHub repository. The original datasets contained 100 000 images each and for the detection of StyleGAN images, a subset was created of these datasets that reduced them in size by the number of images and included them together in a dataset that can be used in the training of the neural network model (Karras et al., 2019).

Downloading and the Dataset

The datasets that are provided by StyleGAN and the FlickrFaces dataset each contain 100 000 high-quality images. The problem however that was faced is that these high-quality image datasets were also very large. Conventionally datasets used and sometimes contained in software such as Google Colab and Kaggle is small in total size. The StyleGAN high quality generated faces image dataset size was a total of 175GB and the FlickrFaces dataset total size was 140GB.

The two datasets were made available on their respective GitHub repositories, redirecting to the file buckets hosted in Google Drive. To access the data locally the drives were shared to an NWU institution Google Drive account which does not have a storage limit. In Google Drive shortcuts to the original datasets could be added to the home directory of the account. the datasets could then be downloaded from the home directory.

The problem however faced when using this method is the internet limitations posed on campus at the North-West University's Potchefstroom campus. When downloading through a browser using on-campus internet infrastructure the connection will be throttled if a single file larger than 1GB is downloaded. Google Drive automatically compresses downloads into zip files when downloading and individual files cannot be downloaded sequentially using the internet browsers. Speeds up until 1GB downloaded would stay constant and at high-capacity then after reaching 1GB would be throttled down to low inconsistent speeds. A better way to retrieve the datasets locally had to be used.

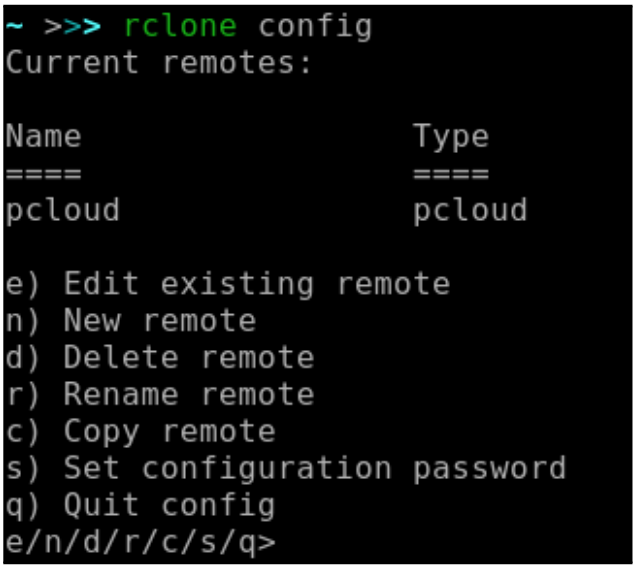


Figure 3.1: rclone setup config example (Craig-Wood, 2021)

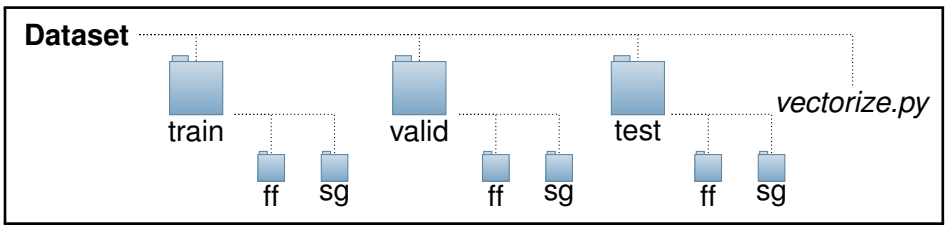
To address the problem of internet speeds being throttled when downloading the datasets the command-line software rclone was used. rclone, a Linux native package, was installed on the windows local system using the dedicated windows installer. As demonstrated in Figure 3.1 rclone could be opened up in the command line and after mounting a connection to the Google Drive account could the whole dataset be downloaded sequentially. Each file is downloaded individually from the Google Drive dataset to the local machine, and the total download times of both datasets are completed overnight.

Dataset Preparation

The datasets had to be compiled together into one dataset that was used within training the neural network. When training a neural network for a binary image classification problem the dataset directory structure is an important aspect of the data. When the neural network trains and validates its training it compares its prediction with the original

image file directory in both training and validation. If the network predicts an image to be StyleGAN generated it checks its prediction against the folder in which the file is located. The subdirectories for the datasets created in this project followed a similar structure to that of the Cat-vs-Dog problem that is used when learning about binary image classifications in neural networks. Table 3.1 shows the structure of the dataset, this directory structure remained the same for all subsequent datasets as they were created to address the resources problems that were identified in Sprint 2 and for the hyperparameter optimization trails in Sprint 3.

Table 3.1: Folder Structure of the dataset used in identifying StyleGAN images



Initially, the subset created from the StyleGAN and FlickrFaces datasets retrieved consisted of 20 000 images. The number of images used for the identification of StyleGAN images was decided based on the findings of Nasr-Esfahani et al. (2016), which that concluded the more images a neural network can use in image classification the better its prediction will be. To an extent, this is true for the problem of StyleGAN generated images versus images of real human faces due to the fine differences between the two types of images. The neural network won't classify the images based on the shape as in the Cat-vs-Dog problem but will rather focus on the small artefacts or features that are present in StyleGAN images for its classification.

Using too large a dataset can lead to the neural network overfitting the data (Trask, 2019). Overfitting occurs when the created neural network model becomes exceptionally good at being classifying the dataset it trained on, but performs much worse when classifying data that it was not trained on. Because of the scope of the proposed project the neural network that is created does not have to be the "*perfect*" model, and will in a sense overfit facial recognition features as that is the premise for the creation of the model. What this means is that if the model created gets an input image of a car and classifies it as a real human or a StyleGAN generated image, this form of overfitting is acceptable in the context of the scope of this project. Figure 3.2 illustrates the acceptable amount of overfitting that can be expected from the neural network. The network however must not overfit the images included in the training set of real human faces retrieved from the FlickrFaces dataset, as then it will classify all images as StyleGAN if it was not in the FlickrFaces training set. This issue is addressed in the second sprint when the initial neural network is created.

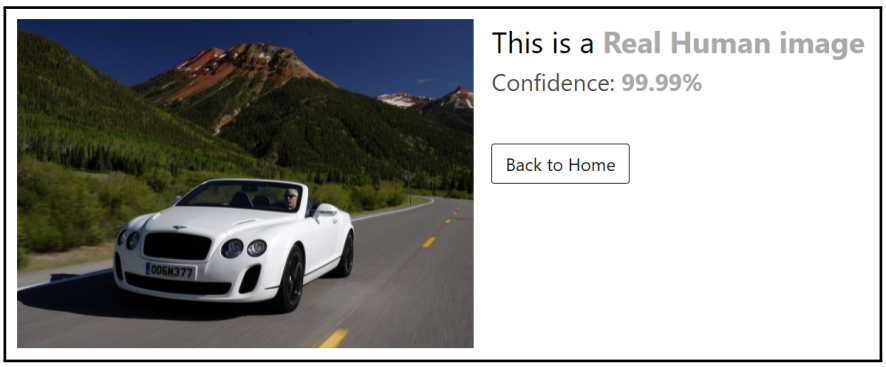


Figure 3.2: Acceptable level of Overfitting for the StyleGAN identification model

A subset of these 2 datasets will improve the development of the artefact as working with a large dataset while testing and implementing hot and cold learning will unnecessarily increase the development time. Training times are influenced by the number of epoch in training and the number of samples used. When developing a neural network works it is standard to use a small subset of the training set to keep the times in testing lower, and after the development of the neural network is completed move over to the full dataset. The Full dataset that will be used in the final training and optimization of the neural network will be larger than the small dataset used in development to ensure that the final model can identify the features present in StyleGAN generated images. Table 3.2 states the sizes of the datasets and the subdirectories that will be used in the development and the final optimization of the neural network model.

Table 3.2: Total amount of images used in the Artefact creation

Directory	Development		Training		Optimization	
	sg	ff	sg	ff	sg	ff
Train	1000	1000	2000	2000	4000	4000
Validation	500	500	1000	1000	2000	2000
Test	500	500	1000	1000	2000	2000
Total Dataset Size		4000		8000		16000

Table 3.2 and Table 3.1 shows the layout of the directories that must be consistent throughout the training of the neural network model in the artefact creation. In binary image classification, the process of training a neural network to classify two sets of images into two separate classes, the neural network references the directory in which an image resides for its check on its prediction. This means that in the training stages the neural network will make its prediction on an image, and check if that image is in the real human directory (ff) or the StyleGAN generated images directory (sg) and then

change its weights according to the correlation between its prediction and the directory path of the image receives. In short, the neural network will use the directories to "*know*" what type of image the received training image is.

The validation folder in the directory is the set of images the neural network uses to check its accuracy after each training step while training its weights. The neural network changes its weights after each image in the training set and just uses the validation set as a benchmark for each checkpoint in training. The Test set is important as it will be used to evaluate the final model. The Test set must be kept aside and never accessed in the training phase of the model. Thus a Test set was created early with unique images the will not be accessed again in the training phases.

Creating the First Neural Network

For the creation of the Artefact, it was identified that the problem of identifying StyleGAN generated images required binary image classification. To understand how binary image classification could be applied to a StyleGAN problem the Cat-vs-Dog example was used to understand binary image classification. The Cat-vs-Dog problem can be seen as a Hello World exercise that will teach the base fundamentals of binary image classification and the knowledge gained in this problem will be applied to the StyleGAN problem (Albaradei et al., 2014). The Cat-vs-Dog problem is a basic exercise to create a neural network to classify images as either a cat or a dog, and the dataset used in this exercise contains 4000 images of both cats and dogs with image sizes of 200px by 200px.

The StyleGAN problem was applied to this exercise by using the StyleGAN dataset instead of Cat-and-Dog images. A problem was identified with this implementation as StyleGAN images are very large compared to other datasets used in CNN neural network implementations, namely the MINST handwriting dataset and the Cat-vs-Dog dataset. StyleGAN images and FlickrFaces images native resolution is 1024px by 1024px, which could not be passed to a neural network catered for the classification of Cat-vs-Dog images as the input layer of the neural network must match the image resolution in CNN's (Albaradei et al., 2014; Wang et al., 2020). Figure 3.3 shows the input layers and image sizes in identifying StyleGAN images and the Cat-vs-Dog exercise.

```
input_layer = layers.Input(shape=(200, 200, 3))
input_layer = layers.Input(shape=(512, 512, 3))
input_layer = layers.Input(shape=(1024, 1024, 3))
```

Figure 3.3: Input layer of the neural network according to the image sizes

In Figure 3.4 a comparison between a dog image retrieved from the Cat-vs-Dog dataset and the dataset used for the identification of StyleGAN images in this project with the sizes of the original images (1024x1024px) the scaled-down (200x200px) images to fit into the input layer of the Cat-vs-Dog problem.

Figure 3.4: Image sizes of Cat-vs-Dog dataset, rescaled artefact dataset and StyleGAN original sizes

When the StyleGAN images were passed into the Cat-vs-Dog neural network with a change in the input layer of the network to accommodate full resolution StyleGAN images a usage limit was reached on Google Colab. This identified that some image processing will be required on the dataset. When the images were scaled down to 200x200px and the Cat-vs-Dog neural network trained on StyleGAN images an accuracy of 61% was achieved. The accuracy achieved with this basic implementation was used as a proof of concept and showed that it was possible to identify StyleGAN images with a neural network. The accuracy of the network however was not ideal and thus a network specific to the StyleGAN problem had to be created.

Image Processing

As mentioned in the previous section, the dataset of images gathered for the identification of StyleGAN images had to be processed to allow the neural network to train on these images. Image processing is an entire field on its own with theory relating to it. For the image processing needs to be required in this stage of the artefact development the python library OpenCV was identified to be the simplest solution to the problem being faced.

OpenCV is a programming library geared mostly at real-time computer vision. It was created by Intel and then supported by Willow Garage and Itseez. Under the open-source Apache 2 License, the library is cross-platform and free to use (Culjak et al., 2012). A python script was created to change the entire dataset sequentially and the need to manually scale the image was avoided. The script was used to create different datasets for later use in this proposed project. The image resolutions of the different datasets created included a 200x200 pixels, 512x512px and 1024x1024px.

Summary

The first sprint in the development of the artefact required the dataset to be retrieved, structured and the images processed. The dataset was downloaded using the rclone program to avoid the throttling limitation experienced. The layout of the directories in the dataset that was compiled for the identification of StyleGAN images was structured to enable the dataset to be used in training. The Cat-vs-Dog exercise proved that a StyleGAN identification neural network model could be created and showed that images had to be processed for a neural network to be trained on the limited resources available. For the processing of the images, a python script was created using the open-source image processing library OpenCV.

3.4.2 Sprint 2

In the second sprint of the artefact development, the first neural network was created based on the findings in the first sprint and using that dataset gathered and processed in the first sprint. The initial neural network that was created in this phase of the artefact development just required hot and cold learning and was not optimized in any form. Hot and Cold learning as analysed in Chapter 1 is an uninformed guessing game in hyperparameter optimization.

Creating the First Neural Network

To create the neural network the TensorFlow package was used. TensorFlow is a machine learning and artificial intelligence software library that is free and open-source. It may be used for a variety of applications, but it focuses on deep neural network training and inference (Abadi et al., 2016). TensorFlow was used within the Jupyter Notebooks in Google Colab and locally, and was run in the python environment.

Keras is an open-source software library for artificial neural networks that include a Python interface (Ang et al., 2017). Keras serves as a user interface for TensorFlow. Keras supports the TensorFlow package and is used for creating the neural network in the development of the artefact (Ang et al., 2017).

The initial neural network was then created using TensorFlow and Keras in Google Colab. The Code for the notebooks used can be found in Appendix C and D. As seen in Figure 3.5 the neural network consisted of 3 convolutional layers, a single dense layer and had a total amount of trainable parameters exceeding 17 million. The large amounts of trainable parameters would result in the resources available in Google Colab being exhausted in training. A hot and cold learning approach was taken and with the smaller test dataset, the model was trained and changed until some improvements could be seen. The trained accuracy of this neural network was 51% at this stage but could be improved more.

Layer (type)	Output Shape	Param #
input_3 (InputLayer)	[(None, 200, 200, 3)]	0
conv2d_13 (Conv2D)	(None, 198, 198, 16)	448
max_pooling2d_12 (MaxPooling	(None, 99, 99, 16)	0
conv2d_14 (Conv2D)	(None, 97, 97, 32)	4640
max_pooling2d_13 (MaxPooling	(None, 48, 48, 32)	0
conv2d_15 (Conv2D)	(None, 46, 46, 64)	18496
max_pooling2d_14 (MaxPooling	(None, 23, 23, 64)	0
flatten_1 (Flatten)	(None, 33856)	0
dense_2 (Dense)	(None, 512)	17334784
dense_3 (Dense)	(None, 1)	513

Total params: 17,358,881
Trainable params: 17,358,881
Non-trainable params: 0

Figure 3.5: Summary of Cat-vs-Dog network applied to the StyleGAN problem

More layers and dropout layers were added to the different convolutions using hot and cold learning. Figure 3.6 is a summary of the neural network architecture of a model created for the identification of StyleGAN images using the principles of hot and cold learning applied on the hyperparameters of the model.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 200, 200, 3)]	0
conv2d_7 (Conv2D)	(None, 198, 198, 16)	448
max_pooling2d_6 (MaxPooling2	(None, 99, 99, 16)	0
dropout_6 (Dropout)	(None, 99, 99, 16)	0
conv2d_8 (Conv2D)	(None, 97, 97, 32)	4640
max_pooling2d_7 (MaxPooling2	(None, 48, 48, 32)	0
dropout_7 (Dropout)	(None, 48, 48, 32)	0
conv2d_9 (Conv2D)	(None, 46, 46, 64)	18496
max_pooling2d_8 (MaxPooling2	(None, 23, 23, 64)	0
dropout_8 (Dropout)	(None, 23, 23, 64)	0
conv2d_10 (Conv2D)	(None, 21, 21, 64)	36928
max_pooling2d_9 (MaxPooling2	(None, 10, 10, 64)	0
dropout_9 (Dropout)	(None, 10, 10, 64)	0
conv2d_11 (Conv2D)	(None, 8, 8, 128)	73856
max_pooling2d_10 (MaxPooling	(None, 4, 4, 128)	0
dropout_10 (Dropout)	(None, 4, 4, 128)	0
conv2d_12 (Conv2D)	(None, 2, 2, 128)	147584
max_pooling2d_11 (MaxPooling	(None, 1, 1, 128)	0
dropout_11 (Dropout)	(None, 1, 1, 128)	0
flatten (Flatten)	(None, 128)	0
dense (Dense)	(None, 200)	25800
dropout_12 (Dropout)	(None, 200)	0
dense_1 (Dense)	(None, 1)	201

Total params: 307,953
Trainable params: 307,953
Non-trainable params: 0

Figure 3.6: Summary of the created Neural Network

Generalization

The dataset used in the development of the artefact was very diverse from the start. The diversity in the data allowed the created CNN to generalize the features present in the StyleGAN images that allowed for their detection. But to improve the model further without overfitting the data some image augmentation principles were used to improve the generalization of the data and improve the final artefact's usability.

Image augmentation improves the model's accuracy while providing more generalization features to the final model (Wang et al., 2020). The image augmentation process in this artefact was implemented in the training function of the CNN. When training the images passed to the neural network was randomly flipped (vertically & horizontally), scaled (large & smaller) and rotated ($180\,\mathrm{deg}$, $90\,\mathrm{deg}$, $-180\,\mathrm{deg}$ & $-90\,\mathrm{deg}$). This process increased the overall sample size of the neural network greatly without an increase in the total amount of images used in training. It is important to note that this process should not be implemented on the validation set as the validation accuracy metric describes the model learning rate (Wang et al., 2020). Figure 3.7 illustrated a visual representation of how the images were passed to the neural network in training to improve generalization in the final model.

Figure 3.7: How image augmentation passes the images to the NN

Figure 3.8 illustrates how the CNN that was created extracts features from the datasets, in its convolution layers. It can be noted that CNN's are proficient in extracting shapes and generalized features from the image.

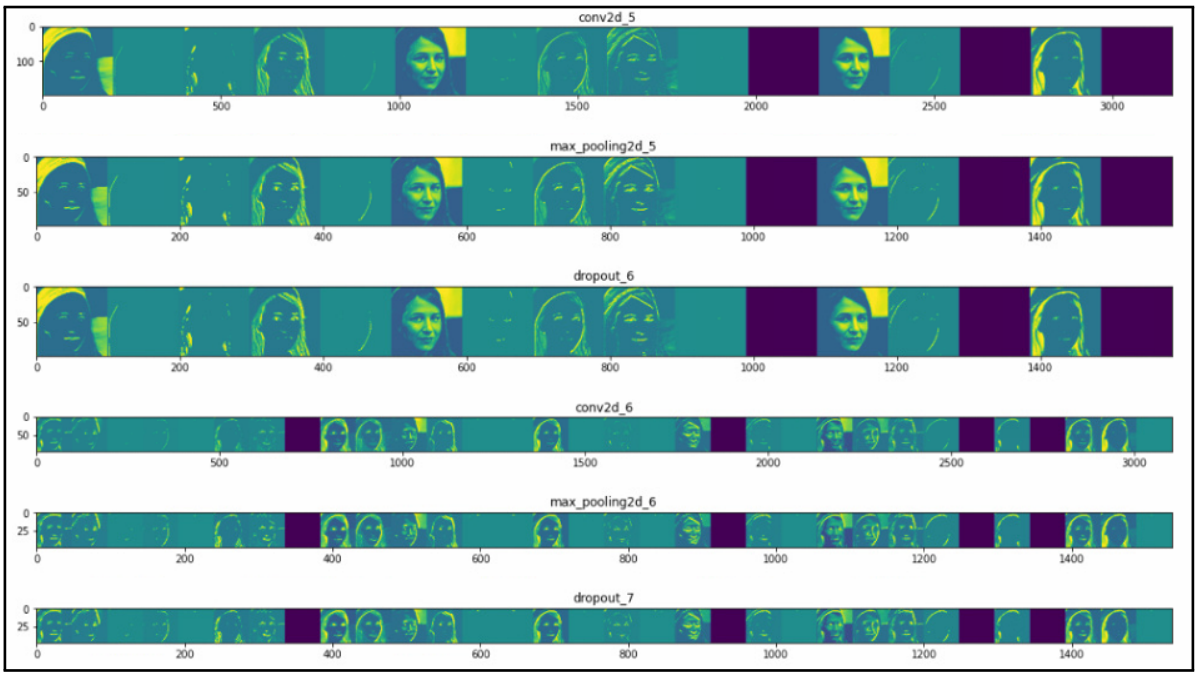


Figure 3.8: Feature extraction by the CNN on a StyleGAN image

The final self-created neural network trained on the training dataset could identify the StyleGAN generated images with 81% accuracy using the test data. The relatively high true accuracy showed promising results but because of the randomness in hot and cold learning and what was identified in the literature review of Chapter 2, the hyperparameters had to be optimized.

3.4.3 Sprint 3

Hot and Cold learning is a viable option for setting neural network parameters when developing neural networks and learning how they can be used to solve problems (Trask, 2019). But for the identification of StyleGAN images hot and cold learning proved to be an inefficient manner in which to create the neural network architecture. Optuna was defined as a possible technology to omit the hot and cold learning process and create a highly optimized neural network.

Optuna: A hyperparameter optimization framework

Optuna is a software framework for automated hyperparameter optimization that is specifically developed for machine learning. It has an imperative, define-by-run user interface. The code built with Optuna has a high level of flexibility thanks to its define-by-run Interface, and the user of Optuna may dynamically design the search spaces for the hyperparameters. The phrases "study" and "trial" are used as follows: A Trial is a single execution of the objective function, whereas a Study is an optimization based on an objective function (Akiba et al., 2019).

When the Optuna hyperparameter process was started in this sprint the trail function and study functions were created based on the documentation of the Optuna framework. The trail is a single iteration in the larger study and the study is a collection of trails where different combinations of hyperparameters are used to reach the goal of the trail. The study goal for identifying StyleGAN images was to increase the accuracy and Akiba et al. (2019) terms it was to move the accuracy metric in a maximum position. The trail function was set to suggest a neural network layer count ranging from a single layer network to a 7 layer network. The optimizer was suggested from a pool of optimizers that perform well with image classification problems as was identified by Bera and Shrivastava (2020) and Kandel et al. (2020). The Study could be changed to also train full neural networks on each trail but it was decided to train the networks the minimal amount just to evaluate the improvements of the models. When the study was completed the resulting model was then trained on the full dataset.

Optuna can provide valuable visualizations of how the hyperparameters interact with the model. Some of the visualizations can even improve the understanding of how the hyperparameters interact with the data of the neural network and similarly display what hyperparameters will have the biggest influence on the accuracy of the model. Figure 3.9 indicates what the most important hyperparameters for the model are. This is useful

when implementing a combination of hyperparameter tuning and hot and cold learning. The less important aspects can be removed from the Study and more important parameters can be researched for relevance specific to the problem at hand.

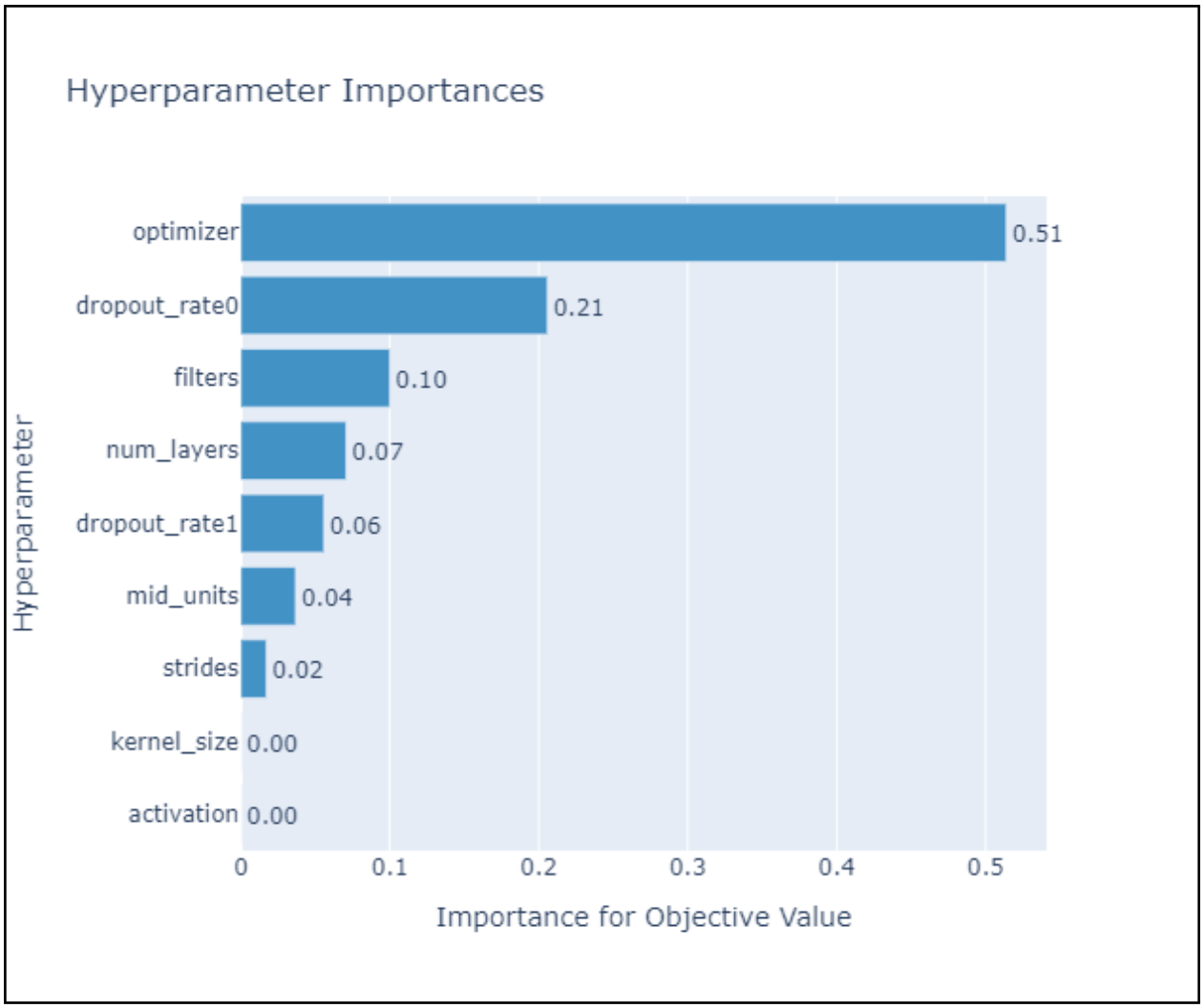


Figure 3.9: Hyperparameter importance from Optuna Study

When the Optuna study was executed on Google Colab a problem was faced regarding available resources. Google Colab provides a free service but can throttle users resources based on the number of resources used. The policies that Google use to dictate what will amount to throttling is unclear and Google can throttle any account based on their discretion. When the Optuna study started the account used on Google Colab for the development of the artefact was throttled down and exceeded the acceptable usage dictated by Google. To work around the problem the jupyter notebooks was downloaded and run on a physical computer provided by Prof. Tiny du Toit. The computer used for the study included a GPU, namely a GTX 1080 with 8GB dedicated graphics memory.

GPU's in machine learning

It was discovered while developing the artefact that GPU's are required for the training of the neural network. In the development, the Google Colab account was restricted by Google because of the resource limit reached on the account. An investigation into the processing capabilities of GPU's revealed that most GPU's on the market can be used for training neural networks. CNN applied to the problem of detecting StyleGAN images could require more processing power because of the large dataset of images. The GPU used on a local machine was a Nvidea GTX 1080 and the GPU used in Google Colab was a Nvidea K80. With the analysis of GPU's, it was found that a GTX 1080 has a better machine learning coefficient than a K80 (Nvidea, 2021). The Nvidea K80 is evaluated by Nvidea to have a deep learning coefficient of 3.7 and the Nvidea GTX 1080 has a deep learning coefficient of 6.1 (Nvidea, 2021). This is a processing increase of more than double the amount offered by Google Colab. The training that was conducted on the GTX 1080 was completed faster than the training conducted on the K80.

The model created using the Optuna framework for hyperparameter optimization was trained on the full dataset on the local machine as previously stated and training times were increased due to the improved machine learning factor of the improved GPU. The trained hyperparameter optimized model could identify StyleGAN generated images with an accuracy of 97.30% and true accuracy of 97.60% as tested with the python script. The results of the Optuna network compared to the network created in Sprint 1 will be further discussed in the Results chapter in this document.

3.4.4 Sprint 4

With the neural network model completed and trained to exceptional accuracy, a front-end application was created to allow users to easily and intuitively interact with the model as stipulated as one of the objectives of this project. By allowing users to interact with the model easily the aim of identifying StyleGAN images can be satisfied as users will be able to pass images to the neural network that in turn will be able to identify if these images were generated by StyleGAN or if the images are that of real human beings.

For the front end of the artefact a web application, that will allow for easier deployment and wider reach had to be created to satisfy the objectives of this proposed project. In the final sprint of the artefact development, various frameworks were considered and while developing the artefact it was realised that a python-based web framework will enable the implementation of the neural network.

Adding the model to the Web App

Because the neural network model was created using the Python libraries for machine learning it was realised that a python-based framework will allow the neural network to be implemented in the web app. If another front end framework like React or Angular was used the implementation of the neural network would be unnecessarily complex. React and Angular are JavaScript based frameworks and for the Python model to communicate with the front end of a JavaScript framework an API had to be implemented. The neural network requires the python machine learning libraries TensorFlow and Keras to pass an image through the model and provide the identification as feedback. Thus the libraries must reside in a python environment where TensorFlow and Keras are installed. Figure 3.10 shows how different frameworks require different aspects in implementation.

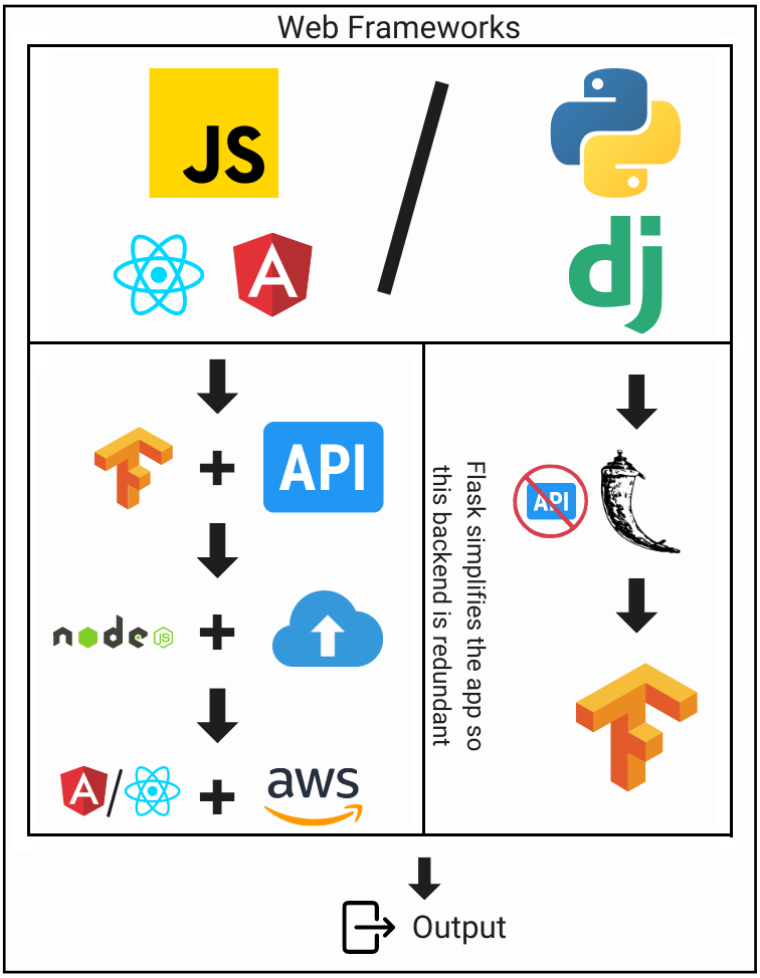


Figure 3.10: Front-end Frameworks implementation comparison

To avoid this added complexity an analysis into python web-based frameworks where conducted and the Django framework stood out. The Django framework however is a very large and clunky framework when compared to the implementation required for the artefact. The child framework Flask, which was derived from Django but with a lightweight

footprint was used for the creation of the front-end of the artefact. The Flask app was developed as a minimalistic interface that would guide users intuitively on how to use the app with clear instructions and a simplified design.

The Front End

The Artefact's final form will be the hosted model on the front end of the application. The following section will illustrate the different front end features of the artefact. The artefacts design was focused to follow a simplistic clear page layout. Minimal controls prevent users from interacting with the web app incorrectly which might lead to errors, and this minimal error handling had to be implemented.

Figure 3.11 shows the home page of the artefact front end. The simple design and clear instructions will make it easy for users to interact with the website and use it to identify StyleGAN images. Users can click on the upload control to upload images to the neural network model that will identify if the images are StyleGAN generated images.

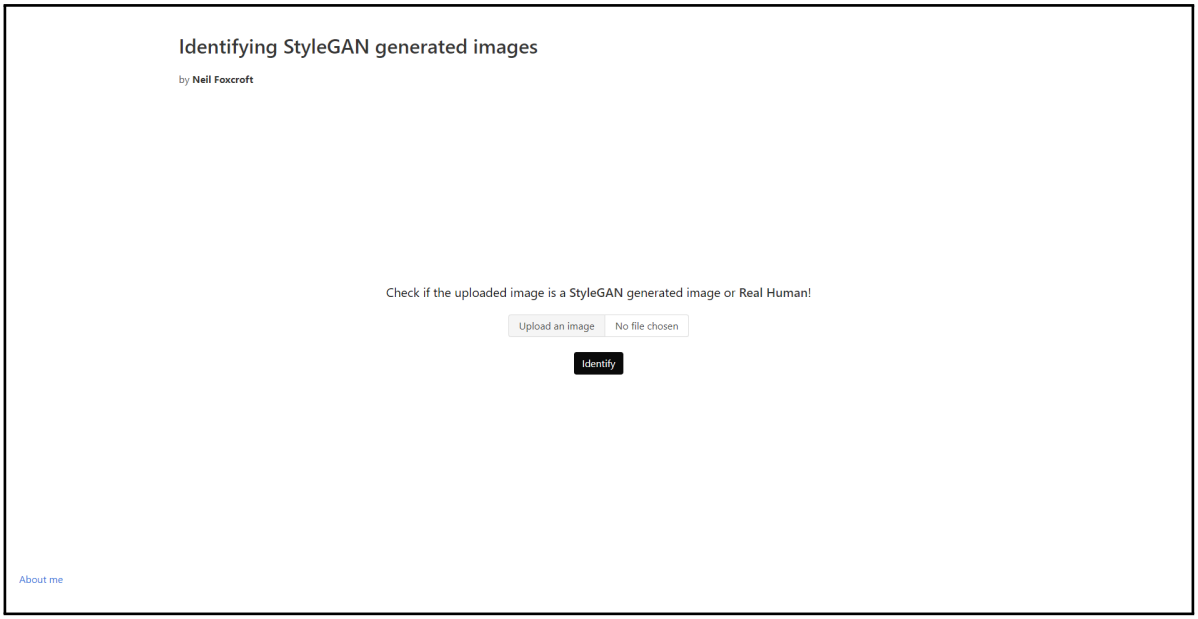


Figure 3.11: Home page of the Artefact

To improve the user-friendliness of the web app some error handling had to create. A page not found page was implemented to navigate the user back to the homepage of the application in the event of error browsing from the user. Figure 3.12 shows the page that will be displayed if a user navigates to a non-existing URL. a single button is presented that will navigate the user back to the home page of the web app.

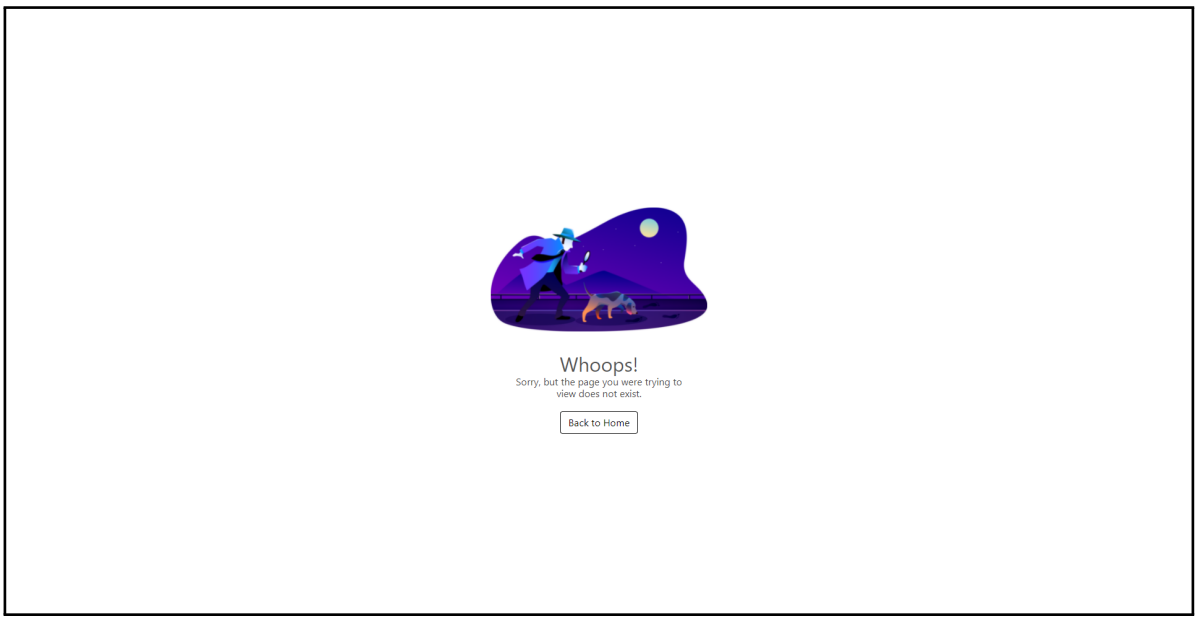


Figure 3.12: Page-not-found in the Artefact

To further improve the user experience a page was created to inform a user in the event of requesting identification without uploading an image. Figure 3.13 shows the page that will be presented if a user request identification of an image, but did not upload an image. The user can navigate back to the home page of the web app and try to upload an image to restart the process.

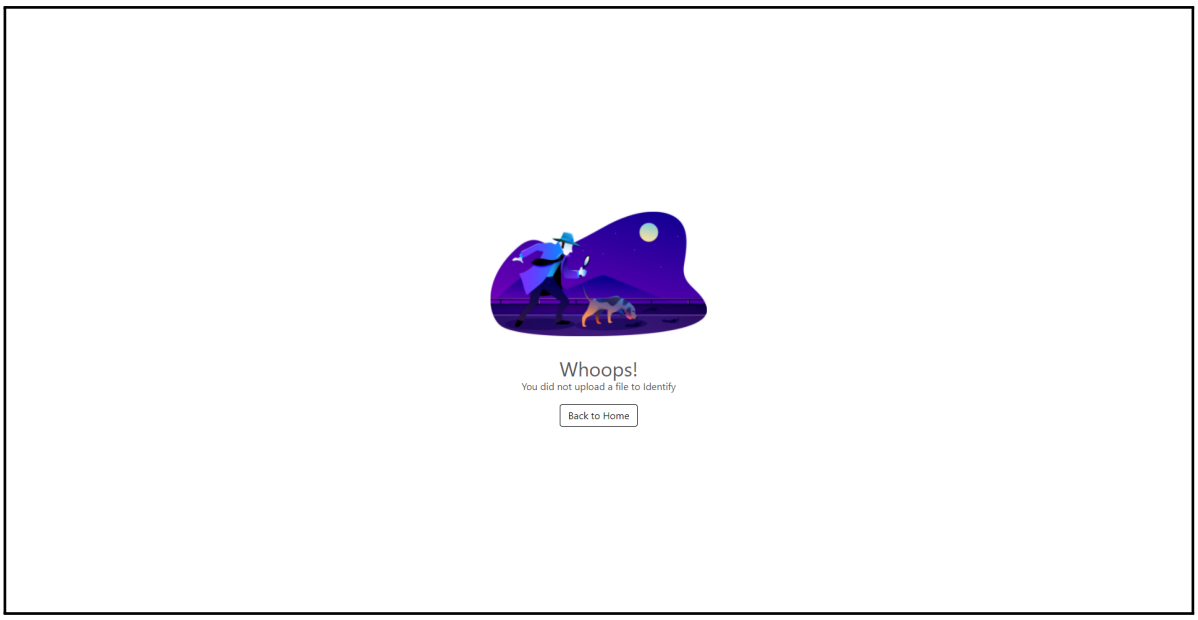


Figure 3.13: No image uploaded page in the Artefact

Figures 3.12 and 3.13 shows how a little user-friendliness consideration in the artefact design can lead to a simplified application and increased usability.

Figure 3.14 shows the output page when a user uploaded an image and the model output the users will see after the image is passed through the model. The output in Figure 3.14 shows the label the model gives the imaged based on the model's training and the confidence in its prediction. In Figure 3.14 the uploaded image was a StyleGAN image and the model labels it as such with the confidence of ≅100%.

Figure 3.14: Example of StyleGAN identification

Figure 3.15 is an example of a real human image being passed through the model and the label the model returns to the front end of the application. The confidence is also displayed to the users and in this case, the model identifies the chosen image as a real human image with a 93.99% confidence in its identification.

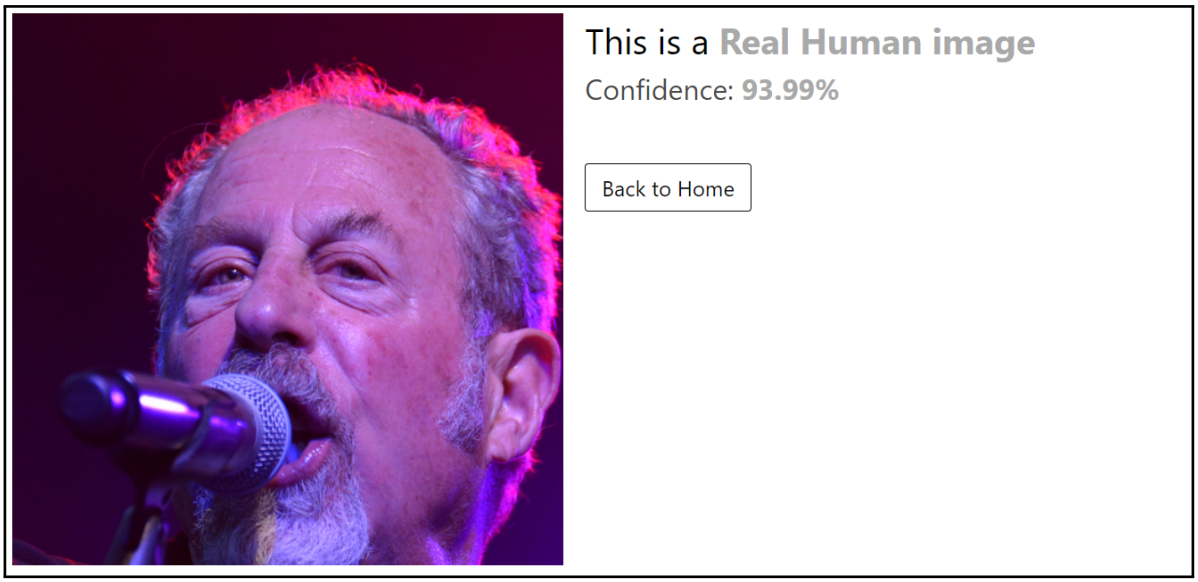


Figure 3.15: Example of real human identification

3.5 Summary

The development of the artefact was described in the manner in which the development took place. The 1st Sprint saw the gathering and processing of the datasets. rlcone was used to download the datasets to circumvent the internet limitations that were faced. The folder structure of the dataset was set out because of its importance when training the network. The 2nd Sprint entailed the creation of the initial artefact and the hot and cold learning techniques implemented in the initial model was seen as a proof of concept for the detection of StyleGAN images. In the 3rd Sprint the CNN was optimized using Optuna, a hyperparameter optimization framework. The results of the model were extremely high and the aim of the project was satisfied. The 4th and final Sprint implemented the final model into a useable lightweight front end web application that satisfied the objectives of the project.

Chapter 4

Results

The Artefact that was created in Chapter 3 could identify StyleGAN images with very high accuracy. The various results achieved in the development of the Artefact will be discussed and evaluated. The initial neural network accuracy will be discussed and example output from the artefact making the predictions will be demonstrated. The hyperparameter optimized neural network that acts as this project final model will be compared to the initial model and the model created by Wang et al. (2020) to demonstrate the large gains made to the model performance using optimization in constrained resources environment.

4.1 The First Neural Network

The neural network created in Sprint 1 produced a detection accuracy of 61% by just replacing StyleGAN generated and real human images with the Cat-vs-Dog images in the introduction to CNN's exercise. The accuracy achieved acted as a proof of concept. Using Hot and Cold learning techniques improved the model to a training accuracy of 81%. When testing the improvements made in hot and cold learning the real accuracy that the model identifies StyleGAn images was 69%, which means the model still cannot extract features from the images in the dataset but can generalize some features. The relatively high accuracy of the simplest implementation and the increase in accuracy when applying hot and cold learning shows how powerful CNN's can be in a world with increasing artificially generated images.

Figure 4.1 shows the self-created model implemented in the front-end artefact predicting on StyleGAN images and real human images from the FlickrFaces dataset. Although the model's prediction confidence is low, the prediction is advantageous compared to a random guess. A random guess and a neural network accuracy of 50% can be seen

as providing the same value. If a model predicts with only 50% accuracy then code that mimics a coin flip prediction by random selection will provide the same results to the user. Therefore with the first neural network model starting with an accuracy higher than 60% and the then reworked model providing an accuracy of 69% was the value added by this project successful from the first iteration of the artefact development. Figure 4.1 below shows the initial models deficiencies when identifying StyleGAN images.

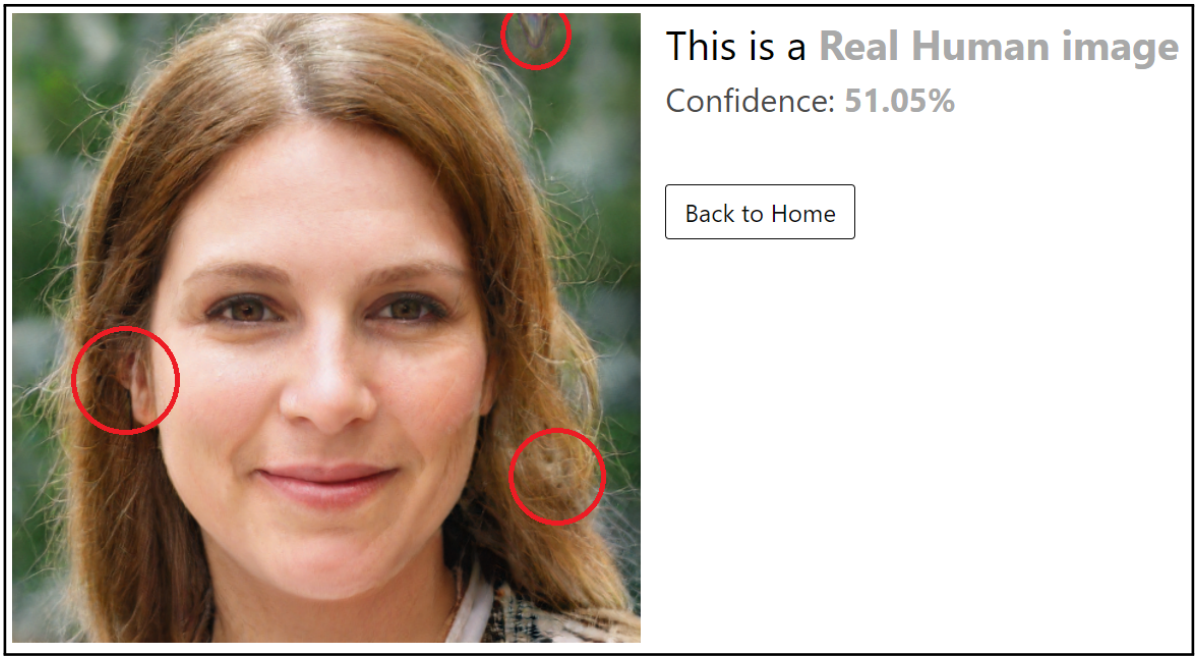


Figure 4.1: Incorrect prediction from the initial model

Figure 4.1 and Table 4.1 shows the neural network that was created based on the activities in the 1st Sprint and concluded in the second sprint using hot and cold learning. The model correctly predicts most human images it receives but struggles with identifying StyleGAN images. What was obtained from this real image tending accuracy is that in the event of a StyleGAN prediction the user can be sure that the image has counter fitting properties, in this problem that would be StyleGAN artefact. That means that this model can be used in practice but only to a certain extent and cannot provide strong assurance. The results in Table 4.1 further illustrates the model's accuracy with real human images and its struggle with StyleGAN generated images.

Table 4.1: Real Images detected vs StyleGAN images detected in the initial model

Type of Image	Identification Accuracy
StyleGAN	68.00%
Real Human	77.20%

The misclassification of the initial model can partly be due to it not being optimized and therefore missing fine features in StyleGAN images. Some generalization occurred in this model and thus the results of the accuracy in SyleGAN misclassification echoed into the results of Real Human detection with high accuracy.

4.2 The Optimized model

The final model that was created in Sprint 3 improved in accuracy drastically. The model compared to the previous model is shown in Table below and a clear gain inaccuracy can be seen. The optimized model improved to such a high level of feature recognition in the StyleGAN data that the model's prediction on StyleGAN2 generated images is similar to the 1st model created in the artefact developments prediction on StyleGAN1 images. Optuna greatly improved to model to such a high accuracy level in StyleGAN1 models that the model can contend with the model created by Wang et al. (2020) for the identification of CNN images. If using only StyleGAN1 images the model that was created for the artefact of this project is as effective as the model created by Wang et al. (2020). The improvement that was made on the process for the model created in this project is the less restraining training and requirements for resources.

Table 4.2: Model comparison in terms of Accuracy and Confidence

Created Models	Tested Accuracy	Average Confidence
Cat-vs-Dog replacement	61.00%	57.00%
Hot and Cold learning	69.00%	65,37%
Hyperparameter Optimized	97,60%	96,98%

As shown in Table 4.2 implementing hyperparameter optimization on the StyleGAN identification problem improved the model to such a high accuracy that the model can be used in the industry, such as image verification services on dating websites. The higher confidence average that the optimized model produces on a single prediction is an important metric of the model's overall performance. When a model receives an image and makes its prediction the confidence metric that is also returned to the user in the front-end of the application will aid in a better identification "accuracy" that the model can provide. In the event of a miss classification, the overall confidence level of the model's prediction will be lower than the average confidence of the network. Therefore even in the event of a miss identification of a StyleGAN image, the user can reference the low confidence level and know that something regarding that image is not standard and a second evaluation might be necessary.

The improvements in neural network accuracy and feature extraction of the created CNN with hyperparameter optimization is clear when looking at the comparison in Table 4.3. A select few images contained in the datasets are used to compare the difference in the performance of the Hot and Cold learning neural network and the Hyperparameter optimized network. The increase in confidence and change in identification substantiates the use of Optuna in Sprint 3 further.

Table 4.3: Specific images comparison between the two CNN created

			Hot-and-Cold learning model		**Hyperparameter optimized model**	
	Image number	Type of image	Prediction	Confidence	Prediction	Confidence
1	09020.png	StyleGAN	Real Human	54,82%	StyleGAN	99,77%
2	09022.png	StyleGAN	StyleGAN	50,02%	StyleGAN	100,00%
3	09063.png	StyleGAN	Real Human	60,66%	Real Human	66,51%
4	09016.png	Real Human	Real Human	57,06%	Real Human	99,99%
5	09018.png	Real Human	Real Human	100,00%	Real Human	99,99%
6	09046.png	Real Human	Real Human	59,77%	StyleGAN	53,22%

In Table 4.3 the increased made by hyperparameter optimization is clear. However, in image number 6 the optimized model incorrectly classifies a human image as a StyleGAN image where the first model created could correctly identify the image. This can be partly to the randomness in prediction when a CNN does not have a full grasp of the data and cannot extract the correct features from the dataset. It is also interesting to note that the optimized model never predicts a Real Human image with 100% confidence. The model is great at generalizing the features in the images but with images of real human beings being very unique and diverse in nature this phenomenon is justified. StyleGAN images on the other hand predicted by the model with a 100% accuracy and a reason for this can be the predictability that accompanies images being created by the same neural network.

The overall confidence of prediction is higher in the optimized model than in the hot and cold learning model. The effect of the increase in confidence is that when implemented the prediction of the final model can be used by users aiming to detect StyleGAN images as an extra metric to prevent false identification. If the final model predicts an image incorrectly the confidence in its prediction is very low in comparison to the confidence average of the model seen in Table 4.2. When implemented users can recheck identifications with confidence in prediction between a certain range to further improve on the identification capabilities that this model presents.

In Figure 4.2, a random StyleGAN image and a real human image not contained in the FlickrFaces dataset is passed to the neural network model in the working complete artefact. The Optuna optimized model identifies the images correctly and illustrates its high confidence in its correct prediction.

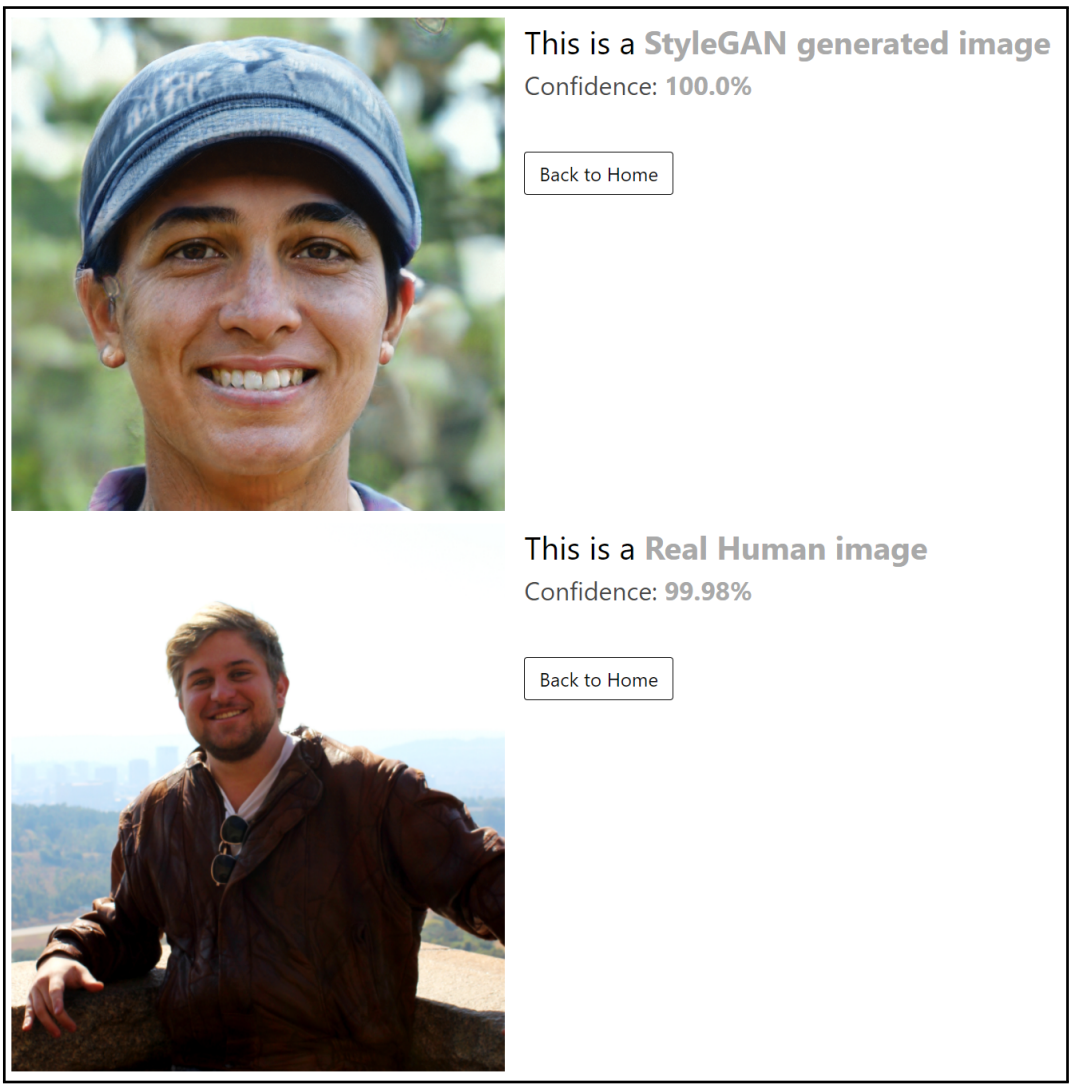


Figure 4.2: Optimized model high confidence and correct prediction

The optimized model that was created as the final neural network model for this proposed project can identify StyleGAN images and Real human images with high accuracy. Because of the image augmentation previously mentioned the model is also great at generalizing the data. The generalization enables the model to also correctly predict human images not contained in the FlickrFaces dataset. Generalization present in the model helped in the StyleGAN specific images identification but was not as prominent in aiding the function of detection as with generalization with real images. StyleGAN images all originate from the same datasets and GAN, the differences being the version of StyleGAN generated images.

The small gains in generalizing features from StyleGAN aided in the detection of StyleGAN1 images with fast gains, but when applying the network model to StyleGAN2 images only a small part of generalization aids in the detection of these images. Figure 4.3 shows the detection of StyleGAN2 images using the model of the artefact.

The model can predict on the images but with the limited testing conducted the prediction of StyleGAN2 images is lower than that of StyleGAN1. This means that this model when implemented will be able to detect some StyleGAN2 images but proves that implementing the same scope of this proposed project on StyleGAN2 and respectively will create a model that can identify these images. A method that can be followed to allow the model to identify all versions of StyleGAN images can be the creation of a dataset that contains StyleGAN, StyleGAN2 and StyleGAN3 images and train a single model on that dataset.

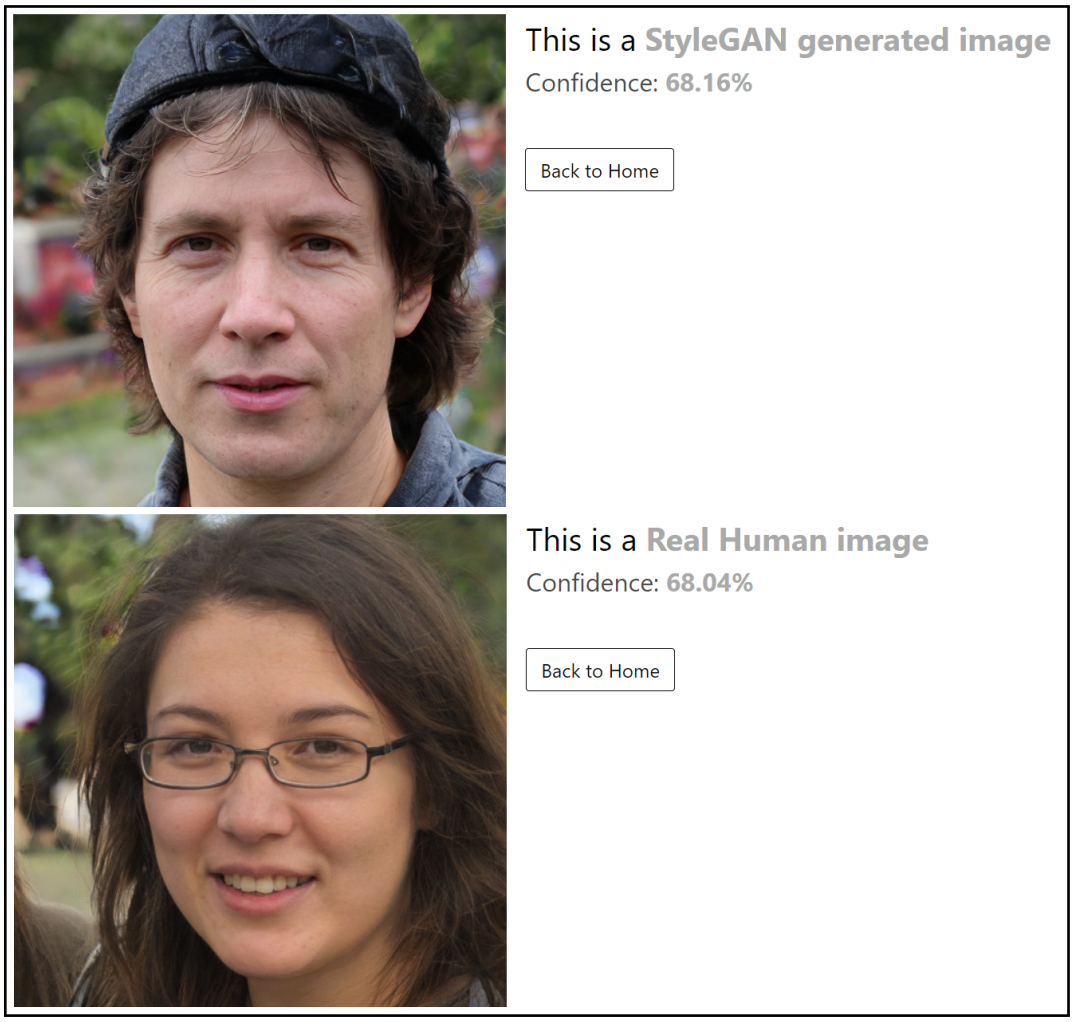


Figure 4.3: StyleGAN2 tested on the Artefact

The relatively low confidence in Figure 4.3 is a result of the increase in difficulty in detecting StyleGAN2 images. The low confidence also shows that the model is optimized towards the detection of StyleGAN1 images without overfitting. Thus the possibility of a neural network that can detect all three versions of StyleGAN images is proven and can be created similar to the network created by Wang et al. (2020) but by implementing hyperparameter optimization and reducing the resource needs when creating the model.

4.3 Summary

The result presented by the CNN of the artefact proves the capabilities of Hyperparameter optimization, specifically to Optuna and Binary image classification. The model created in the development of the artefact using hot and cold learning, the simplest approach, showed the performance of CNN's when training models on image datasets. The initial accuracy of the 1st model was relatively high and could be used but the final model's accuracy improved to such an excellent accuracy that it can be used in practice. The scope aim of the project was successfully achieved when analysing the performance of the final model in the detection of StyleGAN images.

Chapter 5

Reflection

While completing the proposed project I learned a lot about Deep Learning and CNN. The problem of identifying StyleGAN images with the use of Deep Learning and CNN directed me towards learning about how the problem can be solved using technologies such as TensorFlow, Keras and Optuna to train a final model that can identify these generated images with high accuracy. The new technology Optuna increased my understanding of how neural networks function by forcing me to dive in deep and apply my foundational previous knowledge in the field of AI.

The strong point of the artefact created in this proposed project is the high accuracy in which StyleGAN images can be identified. The generalization of the final model is also sufficient that allows it to identify images that differ from the original dataset that was used. The font-end allows users to easily interact with such a complex method of detection without the need for any previous knowledge.

Weak points in the created Artefact is that the model can only detect StyleGAN1 images with high accuracy and falls short in detecting StyleGAN2 and StyleGAN3 generated images. The model sometimes misidentifying the images it receives can be seen as a weakness as when it misclassifies a user might use the misclassification for their decision-making process and in that sense, the artefact will aid in the passing of StyleGAN generated images as real images. The frequency of the above-mentioned weakness is so low that this possibility is negligible.

The aim of the proposed project was fully satisfied and the final method of detection can identify StyleGAN images consistently with high accuracy. The objectives of the project were satisfied in the study in StyleGAN and a full understanding of the technology that supported the development of the method of detection and the final implementation of the method. These objectives were all satisfied based on the high accuracy model being implemented in the frontend web application.

A hurdle in the completion of the proposed project was the managing of the timeframe. The planning in the initial stages of the project allowed for all target dates to be met. The takeaway for me will be to plan for the unexpected as the increase in load-shedding times and other events cause some deviations of the original timeframe of the project.

The process followed in completing the project could be improved upon. The DSRM methodology was useful in the development of the artefact but a possible better methodology could be used to specifically cater for the machine learning project. Resource planning would have allowed for better training using larger datasets and more training steps that ultimately would have led to a model that would have been one of the best performing models for StyleGAN identification in current times.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Albaradei, S., Wang, Y., Cao, L., and Li, L.-J. (2014). Learning mid-level features from object hierarchy for image classification. In *IEEE Winter Conference on Applications of Computer Vision*, pages 235–240. IEEE.

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. *IEEE*, pages 1–6.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74.

Ang, L., LI, Y.-x., and LI, X.-h. (2017). Tensorflow and keras-based convolutional neural network in cat image recognition. *DEStech Transactions on Computer Science and Engineering*, cmsam(cmsam).

Bera, S. and Shrivastava, V. K. (2020). Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification. *International Journal of Remote Sensing*, 41(7):2664–2683.

Burguillo, J. C. (2010). Using game theory and competition-based learning to stimulate student motivation and performance. *Computers & education*, 55(2):566–575.

Chandler, D., Munday, R., and Oxford University, P. (2016). *A dictionary of social media*. Oxford University Press.

Collins, H. (2018). *Creative research: the theory and practice of research for the creative industries*. Bloomsbury Publishing.

Craig-Wood, N. (2021). rclone documentation.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Culjak, I., Abram, D., Pribanic, T., Dzapo, H., and Cifrek, M. (2012). A brief introduction to opencv. In *2012 proceedings of the 35th international convention MIPRO*, pages 1725–1730. IEEE.

Fleishman, G. (2019). How to spot the realistic fake people creeping into your timelines. *Fast Company*.

Fysh, M. C. and Bindemann, M. (2018). Human–computer interaction in face matching. *Cognitive Science*, 42(5):1714–1732.

Gallagher, F. and Calabrese, E. (2019). *Facebook's latest takedown has a twist - AI-generated profile pictures*. ABC News.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Hays, K., Thomas, O., Maynard, I., and Bawden, M. (2009). The role of confidence in world-class sport performance. *Journal of sports sciences*, 27(11):1185–1199.

Kandel, I., Castelli, M., and Popovič, A. (2020). Comparative study of first order optimizers for image classification using convolutional neural networks on histopathology images. *Journal of imaging*, 6(9):92.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. In *Proc. NeurIPS*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, PP.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116.

Krenker, A., Bešter, J., and Kos, A. (2011). Introduction to the artificial neural networks. *Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech*, pages 1–18.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26.

Medvedev, A. and Zemplinerová, A. (2005). Does competition improve performance? evidence from the czech manufacturing industries. *Prague Economic Papers*, 14(4):317–330.

Mitra, A., Mohanty, S. P., Corcoran, P., and Kougianos, E. (2021). A machine learning based approach for deepfake detection in social media through key video frame extraction. *SN Computer Science*, 2(2).

Müller, B., Reinhardt, J., and Strickland, M. T. (1995). *Neural networks: an introduction*. Springer Science & Business Media.

Murugesan, S., Rossi, G., Wilbanks, L., and Djavanshir, R. (2011). The future of web apps. *IT Professional*, 13(5):12–14.

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S., Jafari, M., Ward, K., and Najarian, K. (2016). Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1373–1376.

Nvidea (2021). Cuda gpus.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Peffers, K. E. N., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77.

Pfleeger, C. and Pfleeger, S. L. (2002). *Security in Computing*. Prentice Hall.

Rasmussen, C. E. and Ghahramani, Z. (2001). Occam's razor. *Advances in neural information processing systems*, pages 294–300.

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *towards data science*, 6(12):310–316.

Skilton, M. and Hovsepian, F. (2017). *The 4th industrial revolution : responding to the impact of artificial intelligence on business*. Springer International Publishing AG.

Trask, A. (2019). *Grokking Deep Learning*. Manning Publications Co.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020). Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704.

Weiyin, H., Thong, J. Y. L., Chasalow, L. C., and Dhillon, G. (2011). User acceptance of agile information systems: A model and empirical test. *Journal of Management Information Systems*, 28(1):235–272.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv e-prints*, page arXiv: 1506.06579.

Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485.

Appendix A

Ethics Form

Title of project:

Identifying StyleGAN images

Name: Neil Foxcroft
Supervisor: Dr. R. Serfontein

Starting date: 16/02/2021
End date: 08/11/2021

1. Have you read the information available related to research ethics (Chapter 5 of Researching Information Systems and Computing; BJ Oates and Chapter 13 of Writing for computer science, J Zobel; Manual for postgraduate studies, available on eFundi)?	Yes ✔	No

2. Do you make use of people as the source of data in your project (for example the completion of questionnaires or evaluation of products)?	Yes	No ✔

3. Are there any aspects of your research that you need permission from another party to use (for example use of property or tools)? If yes, provide more detail.	Yes	No ✔

4. Describe your research question and give a brief description of your plans for the collection of data.

With the development of this artefact, certain resources will be used to train the neural network. The training requires images that were generated by StyleGAN. For the comparison in the artefact images of real humans will be used.

The StyleGAN generated images are available on the official StyleGAN GitHub repository. Included in this repository is trained StyleGAN models and multiple datasets of StyleGAN generated images. The licencing of these images is stated on the GitHub repository and is a Creative Commons license by NVIDIA Corporation (Karras *et al*., 2019).

For the verified human faces, the preliminary dataset that will be used is the Flickr Faces dataset that was initially used to benchmark StyleGAN. The individual images were published in Flickr by their respective authors under either Creative Commons, Public Domain. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes (Karras *et al*., 2019).

With the identified need for detection of StyleGAN images and the discussed security implications that the invention of StyleGAN and similar methods introduced the proposed project aims to detect these images with the use of a trained neural network. The main research question that this proposed project aims to answer is: How can StyleGAN generated images be detected?

Data that will be used in this proposed project is free to use data sets provided in the StyleGAN repository and the Flickr Faces dataset. These datasets stated in their respective repositories that they are included with a creative commons license that will allow me to use them for my research purposes.

5. Describe how you plan to provide information about yourself and the goals of your research to participants.

This proposed project will not require any participants

6. Describe what methods you will use to get permission from participants in your study.

This proposed project will not require any participants

7. Will you be able to ensure that participants' information will be used in an anonymous, private, and confidential way? How? This proposed project will not require any participants	Yes ✔	No

8. Are there any foreseeable risks of damage (physical, social, or psychological) to participants or the environment? If you answer yes, give detail of the preventative measures you will follow.	Yes	No ✔

9. Are there any foreseeable risks to the NWU, for example, lawful actions that may follow the research or damage the image of the university? If yes, give detail.	Yes	No ✔

10. Are there any other ethical issues that may occur during the execution of the research (for example conflicting interests)? If yes, provide detail and explain how you plan to manage them.	Yes	No ✔

I hereby declare that the information contained in this form is accurate. I have attempted to identify the risks that may arise in conducting this research and acknowledge my obligations and the rights of the participants. I confirm that the research will be conducted in line with all University, legal and ethical standards.

Name of student: Neil Foxcroft

Signature:_____
Date: 18/04/2021

Name of study leader: Dr R. Serfontein

Signature: _____
Date:

Name of an additional moderator:

Signature: _____
Date:

Appendix B

Research Proposal

SUBJECT GROUP COMPUTER SCIENCE AND INFORMATION SYSTEMS

Research Proposal for an Honours project

The student and the supervisor must consult the *Manual for Postgraduate Studies* before writing the research proposal. The *Manual for Postgraduate Studies* explains in detail what is expected at each of the subheadings below. The proposal should not be longer than 5 pages.

The Subject Group requires that the research proposal will be submitted through the use of this form and in the format below. Please complete using a computer.

1 Student initials, surname, and student number

Initials	N	Surname	Foxcroft	Student number	28418077

2 Degree for which student is registered

BSc Honours in Computer Science and Information Technology

3 Name of supervisor

Initials and surname	Dr R. Serfontein

4 Proposed title

Title (preferably not more than 12 words)	Identifying StyleGAN images

5 Problem statement and substantiation

Provide the theme and link with gaps in the literature and recent research in the area. Indicate the research question, its actuality and how the research will endeavour to answer the question.

StyleGAN is an open-source Generative Adversarial Network (GAN) that can be used to generate faces of people that do not exist (such as those shown on thispersondoesnotexist.com). This means that fraudsters can use StyleGAN generated faces that normally would pass a visual inspection conducted by a human inspector as part of false identities. The detection of such images with the use of artificial intelligence will be useful because of the factors that currently lead to misidentification.

Possible misidentification of StyleGAN images is a reality that needs to be addressed. Humans in the role of identifying artificially generated human faces may be susceptible to external factors hindering their capabilities and increasing the rate of error in which they identify fraudulent images (Fysh & Bindemann, 2018). Fysh and Bindemann also noted that in the specific use case of passport officers that were tested on passport images captured on the same day against the "traveller" presenting that image, that the officers made substantial errors in a controlled environment when comparing the picture identity to that of the traveller. These results enforced their original statement that humans struggle with unfamiliar face identification.

By looking at how the technology has been used since its release, the possible use-cases for GAN generated images and the always growing cybercrime industry the possible detection of these images is identified as a crucial function in the 4th industrial revolution. With these identified factors will the proposed project aim to solve the problem of detecting StyleGAN images by using an artificial intelligence approach to solve the problem.

6 Research aims and objectives

Provide the different general as well as the specific aspects which will form part of the research.

Aims:

The main purpose of the proposed project is to develop a method that can detect fraudulent human faces created by StyleGAN with relative accuracy. Various techniques and approaches to the detection of GAN generated images will be researched and the simplest implemented approach that can still detect these types of images with relative surety will be selected for the artefact.

Objectives:

The success of the project will be weighed against the completion of the secondary objectives that have been identified as listed below.

- Perform a literature study on GANs and specifically analyse the architecture and function of StyleGAN to understand the technology.
- Develop an approach to the successful identification of generated images.
- Develop an artefact that will use the selected method to detect a fake identity.

The successful completion of the above-identified objectives will aid the researcher in satisfying the aim of the proposed research project.

7 Basic hypotheses (where applicable)

The use of Neural Networks will aid in the successful detection of StyleGAN generated images.

8 Method of investigation

8.1 Literature study

Provide an indication only of which literature will be used in the study with key references. A summary of the literature is not required here.

In this modern world with humanity currently in its 4th industrial revolution, the additions of innovative technologies require original approaches to implement and maintain these technologies. The change from the digital age to the automation age is accelerated by the breakthroughs in the fields of artificial intelligence (AI) and security. (Skilton & Hovsepian, 2017)

One of the big advances in artificial intelligence is the creation of StyleGAN, this new approach to a generative adversarial network allowed for more control in an image than its predecessors. (Karras et al., 2019) StyleGAN uses the principles of a GAN to create new images derived from input that specifies what “styles” need to be included in the image. According to Karras et al. (2019) a style is defined in this contexts as a set of parameters that modifies the input of the image to result in different outputs. If the input received is that the image needs to be in the style of a person with glasses, red hair and must be female, StyleGAN will then generate that image based on images used of that similar styles in the initial training of the model. The resulting image will thus be that of the “styles” required in the input. While this functionality can be utilized for various positive use cases – this breakthrough also creates various challenges and setbacks in the field of security, more specifically the aspects of facial recognition and identity verification as malicious use of this GAN might aid in the creation of fraudulent identities. (Mitra *et al.*, 2021)

The big advancement in StyleGAN that led to the conception of this proposed project was the release of StyleGAN2 in February 2020. (Karras, Laine, Aittala*, et al.*, 2020) Following the release of version 2, there were improvements in the generation process of these images with this updated version of StyleGAN. StyleGAN2 saw that images were no longer subverted with artefacts or traces left on the image because of the processing method used in the previous iterations. (Karras, Laine, Aittala*, et al.*, 2020) These traces were easy to identify and clearly showed out of place in the context of the image. With the removal of the traces, usually, in the form of drop-like spots, the difficulty in detecting the generated images increased and similarly the need to detect StyleGAN generated images used maliciously in security verification processes.

8.2 Methods of investigation

The proposed design, data acquisition, procedures, data processing, funding sources (but not a budget), mathematical methods, computer methods.

The positivistic research paradigm is suitable for the proposed project because with the creation of the artefact, the data collected will be examined and an unobjective interpretation of the data is necessary. The data collected will be the results of the artefact's successful identification of StyleGAN generated images.

DSRM is an information systems specific methodology that focuses on research and iterative design. (Peffers *et al.*, 2007) Because the researcher is studying the field and technologies in which they want to solve the specific problem the background knowledge of the problem will be explored parallel to the design of the problem solution. DSRM will enable iterative design and development throughout the completion of the proposed project

The artefact that will be developed will aim to detect StyleGAN generated images between a data set of real images of human faces and StyleGAN generated images. The artefact will be hosted online as this will simplify the development of user interaction. Because of these factors, Agile will be the most suitable methodology for artefact development.

Neural networks and computer vision will be used once an extensive literature study identifies these methods suitable for StyleGAN image detection.

9 Provisional chapter division

Chapter 1: Introduction - This chapter will be concluded in the project proposal phase. It will include the research question, the project description and background. It will include the proposed project plan for the entire project.

Chapter 2: Literature Study - This chapter will be comprised of all the necessary research to understand the project and fulfil the project aims and objectives. Mostly in this project will be focusing on the specific workings of StyleGAN to effectively detect fake images.

Chapter 3: Development of the artefact - Chapter 3 will apply the chosen methodologies that were identified and discussed in Chapter 1 to enable the successful development of the proposed project's artefact. This chapter will document the artefact development phase including unfamiliar problems that are identified within the development stage. The success of the artefact will be compared to the Aims and Objectives of Chapter 1 and if they are met.

Chapter 4: Testing and Results - The results of the Artefact will be introduced in this Chapter and the successful identification of StyleGAN images will be determined. The testing in this Chapter will identify the success of the artefact with the comparison in Chapter 3.

Chapter 5: Conclusion - Chapter 5 will summarize the entire proposed project and conclude if the problem was solved with the successful completion of the objectives of the proposed project that allowed it to fulfil the aim of the project. The limitations that impeded the proposed project will be discussed in this section. Future expansion of the project will be discussed and explained in the context of the limitations faced

10 Literature references

Provide complete references to the literature referenced in this proposal only.

Gallagher, F. and Calabrese, E. (2019).Facebook's latest takedown has a twist - AI-generated profile pictures. ABC News. 31 December 2019. https://abcnews.go.com/US/facebooks-latest-takedown-twist-ai-generated-profile-pictures/story?id=67925292 Date of access: 24 March 2021.

Fysh, M. C. and Bindemann, M. (2018). Human–computer interaction in face matching.Cognitive Science, 42(5):1714–1732.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture forgenerative adversarial networks.IEEE transactions on pattern analysis and machineintelligence, PP.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzingand improving the image quality of stylegan.2020 IEEE/CVF Conference on ComputerVision and Pattern Recognition (CVPR), pages 8107–8116.

Mitra, A., Mohanty, S. P., Corcoran, P., and Kougianos, E. (2021). A machine learningbased approach for deepfake detection in social media through key video frameextraction.SN Computer Science, 2(2)

Peffers, K. E. N., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). Adesign science research methodology for information systems research.Journal ofManagement Information Systems, 24(3):45–77.

Skilton, M. and Hovsepian, F. (2017).The 4th industrial revolution : responding to theimpact of artificial intelligence on business. Springer International Publishing AG.

Boxcroft

.....

Student

18 / 04 / 2021

Date

.....

Supervisor

.....

Date

Appendix C

Jupyter Notebook: 1

[ ]:
```
import zipfile
from google.colab import drive

drive.mount('/content/drive/')

#poc_DATASET
#zip_ref = zipfile.ZipFile("/content/drive/My Drive/sg_ff_filtered_red.zip", 'r')

#ISGI_20000_200gray_DATASET
# zip_ref = zipfile.ZipFile("/content/drive/My Drive/ISGI_dataset_200g.zip", 'r')

#ISGI_20000_200rgb_DATASET
zip_ref = zipfile.ZipFile("/content/drive/My Drive/ISGI_dataset_200rgb.zip", 'r')

zip_ref.extractall("/tmp/")
zip_ref.close()
```

[ ]:
```
import os

base_dir = '/tmp/ISGI_dataset_200rgb'
train_dir = os.path.join(base_dir, 'train')
validation_dir = os.path.join(base_dir, 'valid') #valid or validations

# Directory with our training FlickerFaces pictures
```

```
train_ff_dir = os.path.join(train_dir, 'ff')

# Directory with our training StyleGAN pictures
train_sg_dir = os.path.join(train_dir, 'sg')

# Directory with our validation FlickerFaces pictures
validation_ff_dir = os.path.join(validation_dir, 'ff')

# Directory with our validation StyleGAN pictures
validation_sg_dir = os.path.join(validation_dir, 'sg')
```

[ ]:
```
train_ff_fnames = os.listdir(train_ff_dir)
train_ff_fnames.sort()
print(train_ff_fnames[:10])

train_sg_fnames = os.listdir(train_sg_dir)
train_sg_fnames.sort()
print(train_sg_fnames[:10])
```

[ ]:
```
print('Training_FlickerFaces images total: \t', len(os.listdir(train_ff_dir)))
print('Training_StyleGAN images total: \t', len(os.listdir(train_sg_dir)))
print('Validation_FlickerFaces images total: \t', len(os.
 ↪listdir(validation_ff_dir)))
print('Validation_StyleGAN images total: \t', len(os.listdir(validation_sg_dir)))
```

[ ]:
```
%matplotlib inline

import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import random

#params for graph
nrows = 4
ncols = 4

#index for iteration
pic_index = random.randint(0, 990)
```

[ ]:
```
fig = plt.gcf()
fig.set_size_inches(ncols * 4, nrows * 4)

pic_index += 8
next_ff_pix =  [os.path.join(train_ff_dir, fname)
                for fname in train_ff_fnames[pic_index-8:pic_index]]
next_sg_pix =  [os.path.join(train_sg_dir, fname)
                for fname in train_sg_fnames[pic_index-8:pic_index]]
```

```
for i, img_path in enumerate(next_ff_pix+next_sg_pix):
  sp = plt.subplot(nrows, ncols, i + 1)
  sp.axis('Off')

  img = mpimg.imread(img_path)
  plt.imshow(img)

  plt.show
```

[ ]:
```
from tensorflow.keras import layers
from tensorflow.keras import Model
from tensorflow.keras.layers import BatchNormalization, Dropout
```

[ ]:
```
input_layer = layers.Input(shape=(200, 200, 3))

x = layers.Conv2D(16, 3, activation='relu')(input_layer)
x = layers.MaxPooling2D(2)(x)
x = Dropout(rate = 0.2)(x)

x = layers.Conv2D(32, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)
x = Dropout(rate = 0.3)(x)

x = layers.Conv2D(64, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)
x = Dropout(rate = 0.4)(x)

x = layers.Conv2D(64, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)
x = Dropout(rate = 0.5)(x)

x = layers.Conv2D(128, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)
x = Dropout(rate = 0.5)(x)

x = layers.Conv2D(128, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)
x = Dropout(rate = 0.5)(x)

x = layers.Flatten()(x)

x = layers.Dense(200, activation='relu')(x)
x = Dropout(rate = 0.5)(x)

output_layer = layers.Dense(1, activation='sigmoid')(x)

model = Model(input_layer, output_layer)
```

```
model.summary()
```

[ ]:
```
input_layer = layers.Input(shape=(200, 200, 3))

x = layers.Conv2D(16, 3, activation='relu')(input_layer)
x = layers.MaxPooling2D(2)(x)

x = layers.Conv2D(32, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)

x = layers.Conv2D(64, 3, activation='relu')(x)
x = layers.MaxPooling2D(2)(x)

x = layers.Flatten()(x)

x = layers.Dense(512, activation='relu')(x)

output_layer = layers.Dense(1, activation='sigmoid')(x)

model = Model(input_layer, output_layer)

model.summary()
```

[ ]:
```
from tensorflow.keras.optimizers import RMSprop

model.compile(loss='binary_crossentropy',
              optimizer=RMSprop(learning_rate=0.001),
              metrics=['acc'])
```

[ ]:
```
from tensorflow.keras.preprocessing.image import ImageDataGenerator

train_datagen = ImageDataGenerator(rescale=1./255,
                                   zoom_range = 0.2,
                                   horizontal_flip = True,
                                   vertical_flip = True)

val_datagen = ImageDataGenerator(rescale=1./255)

# Flow training images in batches of 20 using train_datagen generator
train_generator = train_datagen.flow_from_directory(
        train_dir,  # This is the source directory for training images
        target_size=(200, 200),
        batch_size=20,
        # Since we use binary_crossentropy loss, we need binary labels
        class_mode='binary')
```

```
# Flow validation images in batches of 20 using val_datagen generator
validation_generator = val_datagen.flow_from_directory(
        validation_dir,
        target_size=(200, 200),
        batch_size=20,
        class_mode='binary')
```

[ ]:
```
history = model.fit(
      train_generator,
      steps_per_epoch=800,  # 2000 images = batch_size * steps
      epochs=15,
      validation_data=validation_generator,
      validation_steps=100,  # 1000 images = batch_size * steps
      verbose=2)
```

[ ]:
```
scores = model.evaluate(validation_generator, verbose=0)
print("Accuracy: %.2f%%" % (scores[1]*100))
```

[ ]:
```
import numpy as np
import random
from tensorflow.keras.preprocessing.image import img_to_array, load_img

# define a new Model that will take an image as input, and will output
# intermediate representations for all layers in the previous model after
# the first.
successive_outputs = [layer.output for layer in model.layers[1:]]
visualization_model = Model(input_layer, successive_outputs)

# prepare a random input image of a FlickerFaces or StyleGAN from the training␣
 ↪set.
ff_img_files = [os.path.join(train_ff_dir, f) for f in train_ff_fnames]
sg_img_files = [os.path.join(train_sg_dir, f) for f in train_sg_fnames]
img_path = random.choice(ff_img_files + sg_img_files)

img = load_img(img_path, target_size=(200, 200))  # this is a PIL image
x = img_to_array(img)  # Numpy array with shape (150, 150, 3)
x = x.reshape((1,) + x.shape)  # Numpy array with shape (1, 150, 150, 3)

# Rescale by 1/255
x /= 255

#  run our image through our network, thus obtaining all
# intermediate representations for this image.
successive_feature_maps = visualization_model.predict(x)

# These are the names of the layers, so can have them as part of our plot
layer_names = [layer.name for layer in model.layers[1:]]
```

```
# Now  display our representations
for layer_name, feature_map in zip(layer_names, successive_feature_maps):
  if len(feature_map.shape) == 4:
    # Just do this for the conv / maxpool layers, not the fully-connected layers
    n_features = feature_map.shape[-1]  # number of features in feature map
    # The feature map has shape (1, size, size, n_features)
    size = feature_map.shape[1]
    # We will tile our images in this matrix
    display_grid = np.zeros((size, size * n_features))
    for i in range(n_features):
      # Postprocess the feature to make it visually palatable
      x = feature_map[0, :, :, i]
      x -= x.mean()
      x /= x.std()
      x *= 64
      x += 128
      x = np.clip(x, 0, 255).astype('uint8')
      # We'll tile each filter into this big horizontal grid
      display_grid[:, i * size : (i + 1) * size] = x
    # Display the grid
    scale = 20. / n_features
    plt.figure(figsize=(scale * n_features, scale))
    plt.title(layer_name)
    plt.grid(False)
    plt.imshow(display_grid, aspect='auto', cmap='viridis')
```

[ ]:
```
# Retrieve a list of accuracy results on training and validation data
# sets for each training epoch
acc = history.history['acc']
val_acc = history.history['val_acc']

# Retrieve a list of list results on training and validation data
# sets for each training epoch
loss = history.history['loss']
val_loss = history.history['val_loss']

# Get number of epochs
epochs = range(len(acc))

# Plot training and validation accuracy per epoch
plt.plot(epochs, acc)
plt.plot(epochs, val_acc)
plt.title('Training and validation accuracy')

plt.figure()
```

```
# Plot training and validation loss per epoch
plt.plot(epochs, loss)
plt.plot(epochs, val_loss)
plt.title('Training and validation loss')
```

[ ]:
```
model.save('placeholderm2.h5')
```

[ ]:
```
!zip -r /content/model_1 /content/model
```

Appendix D

Jupyter Notebook: 2

[ ]:
```
import zipfile
from google.colab import drive

drive.mount('/content/drive/')

#poc_DATASET
zip_ref = zipfile.ZipFile("/content/drive/My Drive/sg_ff_filtered_red.zip", 'r')

#ISGI_20000_200gray_DATASET
# zip_ref = zipfile.ZipFile("/content/drive/My Drive/ISGI_dataset_200g.zip", 'r')

#ISGI_20000_200rgb_DATASET
#zip_ref = zipfile.ZipFile("/content/drive/My Drive/ISGI_dataset_200rgb.zip",␣
 ↪'r')

zip_ref.extractall("/tmp/")
zip_ref.close()
```

[ ]:
```
import os

base_dir = '/tmp/sg_ff_filtered_red'
train_dir = os.path.join(base_dir, 'train')
validation_dir = os.path.join(base_dir, 'validation')
```

```
# Directory with our training FlickerFaces pictures
train_ff_dir = os.path.join(train_dir, 'ff')

# Directory with our training StyleGAN pictures
train_sg_dir = os.path.join(train_dir, 'sg')

# Directory with our validation FlickerFaces pictures
validation_ff_dir = os.path.join(validation_dir, 'ff')

# Directory with our validation StyleGAN pictures
validation_sg_dir = os.path.join(validation_dir, 'sg')
```

[ ]:
```
#imports
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout
from tensorflow.keras.backend import clear_session
import tensorflow as tf
tf.test.gpu_device_name()
```

[ ]:
```
print('Training_FlickerFaces images total: \t', len(os.listdir(train_ff_dir)))
print('Training_StyleGAN images total: \t', len(os.listdir(train_sg_dir)))
print('Validation_FlickerFaces images total: \t', len(os.
 ↪listdir(validation_ff_dir)))
print('Validation_StyleGAN images total: \t', len(os.listdir(validation_sg_dir)))
```

[ ]:
```
!pip install optuna
```

[ ]:
```
import optuna
```

[ ]:
```
dropout_rate = [0] * 2

def create_model(trial):

    num_layers = trial.suggest_int("num_layers", 1, 7)
    activation = trial.suggest_categorical("activation", ["relu"])
    dropout_rate[0] = trial.suggest_uniform('dropout_rate'+str(0), 0.0, 0.5)
    dropout_rate[1] = trial.suggest_uniform('dropout_rate'+str(1), 0.0, 0.5)
    mid_units = int(trial.suggest_discrete_uniform("mid_units", 100, 300, 100))
    filters=trial.suggest_categorical("filters", [16, 32, 64, 128])
    kernel_size=trial.suggest_categorical("kernel_size", [3, 3])
    strides=trial.suggest_categorical("strides", [1, 2])

    classifier = Sequential()

    #step 1  - Convolution Layers

    classifier.add(
```

```
        Conv2D(
            filters=filters,
            kernel_size=kernel_size,
            strides=1,
            activation = activation,
            input_shape=(200, 200, 3),
        )
    )

    classifier.add(MaxPooling2D(pool_size=(2, 2)))
    for i in range(1, num_layers):
        classifier.add(
            Conv2D(
                filters=filters,
                kernel_size=kernel_size,
                strides=1,
                activation = activation,
            )
        )
    classifier.add(MaxPooling2D(pool_size=(2, 2)))
    classifier.add(Dropout(dropout_rate[0]))
    classifier.add(Flatten())
    classifier.add(Dense(units = mid_units, activation = activation))
    classifier.add(Dropout(dropout_rate[1]))
    classifier.add(Dense(units = 1, activation ='sigmoid'))

    return classifier
```

[ ]:
```
#image augumentation
from keras.preprocessing.image import ImageDataGenerator

#Data Preparation

train_datagen = ImageDataGenerator(rescale = 1./255,
                                   shear_range = 0.2,
                                   zoom_range = 0.2,
                                   horizontal_flip = True)

test_datagen = ImageDataGenerator(rescale = 1./255)

training_set = train_datagen.flow_from_directory(train_dir,
                                                 target_size = (200, 200),
                                                 batch_size = 10,
                                                 class_mode = 'binary')

test_set = test_datagen.flow_from_directory(validation_dir,
                                            target_size = (200, 200),
```

```
                                            batch_size = 10,
                                            class_mode = 'binary')
```

[ ]:
```
training_set
```

[ ]:
```
def objective(trial):

    optimizer = trial.suggest_categorical("optimizer", ["sgd", "adam",␣
 ↪"rmsprop", "adadelta", "adagrad", "adamax"])

    classifier = create_model(trial)

    classifier.compile(optimizer = optimizer, loss = 'binary_crossentropy',␣
 ↪metrics = ['accuracy'])

    history = classifier.fit(training_set,
                         steps_per_epoch = 100, # num_samples // batch_size
                         epochs = 5, # entire iteration over dataset
                         validation_data = test_set,
                         validation_steps = 50) #https://keras.io/api/models/
 ↪model_training_apis/

    classifier.save('/drive/MyDrive/Models/trialmodel_' + str(history.
 ↪history['val_accuracy'][-1]) +".h5")


    return history.history["val_accuracy"][-1]
```

[ ]:
```
import pickle

study = optuna.create_study(direction="maximize", )

#studypik = pickle.load(open('study.pickle', 'rb'))
study.optimize(objective, n_trials = 10, timeout = 60 * 60 * 3,␣
 ↪show_progress_bar=True)
print(studypik.best_params)
print(studypik.best_value)
pickle.dump(studypik, open('study.pickle', 'wb'))
```

[ ]:
```
study = optuna.create_study(direction="maximize", )
study.optimize(objective, n_trials = 10, timeout = 60 * 60 * 3,␣
 ↪show_progress_bar=True)
print(study.best_params)
print(study.best_value)
```

[ ]:
```
print(study.best_params)
print(study.best_value)
```

[ ]:
```
fig = optuna.visualization.plot_optimization_history(study)
fig.show()
```

[ ]:
```
fig = optuna.visualization.plot_param_importances(study)
fig.show()
```

[ ]:
```
print(studypik.best_params)
print(studypik.best_value)
pickle.dump(studypik, open('study.pickle', 'wb'))
```

[ ]:
```
print("Number of finished trials: {}".format(len(study.trials)))

print("Best trial:")
trial = study.best_trial

print("  Value: {}".format(trial.value))

print("  Params: ")
for key, value in trial.params.items():
  print("    {}: {}".format(key, value))
```

[ ]:
```
import pickle

studypik = pickle.load(open('study.pickle', 'rb'))
print(studypik.best_params)
print(studypik.best_value)
pickle.dump(studypik, open('study.pickle', 'wb'))
```

[ ]:
```
!pip install pyyaml h5py
```

[ ]:
```
import os

import tensorflow as tf
from tensorflow import keras

print(tf.version.VERSION)
```

[ ]:
```
new_model = tf.keras.models.load_model('/content/drive/MyDrive/optunam.h5')

# Check its architecture
new_model.summary()
```

[ ]:
```
import os


model = tf.keras.models.load_model("/content/trialmodel_0.9764999747276306.h5")
```