



Developing a cloud-based machine learning algorithm that can detect anomalies and security threats in PDFs

(Number of Words: 10,000)

CIS7017_YR_22 Technology dissertation

Made by: Riad Anas

[Student ID: 20245669]

Module supervisor: Priyesh Dhole

MSc. Advanced Computer Science

Cardiff School of Technology

Cardiff Metropolitan University

28th July 2023

Acknowledgment

Words cannot express my gratitude and appreciation to my supervisor and mentor for his constant support, invaluable patience and the positive support.

I also could not have undertaken this journey without the support of my institution Cardiff School of Technology.

Additionally, this endeavour would not have been possible without all the generous resources made available for me by Cardiff Metropolitan University who supported me throughout my dissertation.

Of course, I could not miss mentioning my family, especially my parents but also my partner who supported me all along. Their belief in me has kept my spirit and focus high during this process. I am forever grateful for this.

Abstract

In today's world, data has sparked a significant revolution, turning data scientists into leaders in the digital field. However, this shift has raised doubts. The debate about the power of data has given rise to important areas, and cybersecurity is particularly critical in the digital era. The rapid growth of smart devices and the Internet of Things (IoT) has highlighted the need for strong cybersecurity measures to prevent harmful actions. At the same time, cloud computing has revived digital industries, providing various services like software management, storage, agile analytics, artificial intelligence, and machine learning. This master's dissertation explores the connections between data science, cybersecurity, and cloud computing, uncovering potential consequences and offering insights for sustainable practices in this complex landscape.

This study dives into the heart of data science, where massive datasets are understood, often involving mathematics, statistics, and visualisation. Big Data's rise has transformed data science, allowing complex queries and performance improvements in various sectors. Meanwhile, the increase in internet use and social media has pushed cybersecurity to the front lines, demanding strict measures against growing cyber risks. Merging these fields with dynamic cloud computing has opened new opportunities, with Amazon Web Services (AWS) SageMaker standing out as a powerful tool for developing, training, and deploying machine learning models at scale.

At the centre of this research is the need to enhance security for PDF documents through a cloud-based machine learning program. The project aims to uncover the connections between cybersecurity, cloud computing, and data science, focusing on securing universally used PDF documents. With over 70% of companies shifting to the cloud, safeguarding data from outside threats has become crucial. The project uses AWS SageMaker to create an algorithm that identifies anomalies and security risks in PDFs.

In essence, this dissertation shines a light on the complex relationship between data science, cybersecurity, and cloud computing. By using AWS SageMaker and classification machine learning algorithms, the research aims to strengthen the security of PDFs in the cloud, contributing to the broader conversation on data protection and digital integrity.

Keywords: Machine learning, Cybersecurity, Data, Data science, PDF, Cloud computing

Acknowledgment	2
Abstract	3
Chapter 1: Introduction	7
1.1 Introduction	8
1.2 Background	8
1.3 Motivations	10
1.4 Flow of the report	11
1.5 Aims, Objectives and Research Questions	12
Chapter 2: Literature Review	13
2.1 Introduction	14
2.2 Cybersecurity	14
2.3 PDFs & Security threats	16
2.4 Artificial intelligence	18
2.4.1 Broad introduction to AI	18
2.4.2 Machine learning & deep learning	19
2.4.3 Machine learning models	21
2.4.4 Machine learning biases	25
2.5 Cloud computing	26
2.5.1 Introduction to AWS	27
2.5.2 AWS Sagemaker	28
Chapter 3: Methodology	30
3.1 Introduction	31
3.2 Research Philosophy	31
3.3 Research Approach	31
3.4 Research Strategy	32
3.5 Research Design	32
3.6 Time Horizons	32
3.7 Data Collection	33
3.8 Data Analysis	33
3.9 Ethical Considerations	33
3.10 Limitations	33
3.11 Validity and Reliability	33
3.12 Dissemination of Findings	34
3.13 Conclusion	34
Chapter 4: Dataset Experiments and Results	35
4.1 Introduction	36
4.2 Experiments and results	36
4.2.1 Dataset description	36
4.2.2 Correlation and important features	37

4.2.3 Encryption effect on malicious behaviour	38
4.2.4 New dataset proposed	39
4.2.5 Machine learning models	40
4.2.6 Machine learning model chosen	41
4.2.7 Comparison with previous work	42
4.2.8 Conclusion	42
Chapter 5: Discussion	44
Chapter 6: Conclusion and future work	46
References	48
Appendix	53

List Of Figures

1. Awareness of cybersecurity terms for various age groups
2. The sections of a typical PDF file
3. Analysis of publications on the introduction of Artificial intelligence in education
4. AI / DL / ML Hierarchy
5. Supervised machine learning (Classification)
6. Unsupervised machine learning (Clustering)
7. Reinforcement learning
8. Cloud computing framework
9. General representation of the cloud
10. Privacy preserving cloud computing architecture
11. Research onion applied to this project
12. Research methods for business students (Pearson education)
13. Balanced sample of malicious and benign PDFs in the dataset
14. Correlation between independent features
15. Visualisation of the highest correlations
16. Encryption plot showing the effect on PDFs
17. New dataset proposed features
18. Model computational report with various ML models
19. Full process of the project from new dataset to ML models to Final predictions
20. Comparison of this research with previous work

List Of Abbreviations

1. ML: Machine learning
2. AI: Artificial Intelligence
3. DL: Deep learning
4. KNN: K-nearest neighbour
5. SVM: Support Vector Machine
6. PDF: Portable Document Formats
7. AWS: Amazon Web Services

Chapter 1: Introduction

1.1 Introduction

The world is currently facing one of the biggest revolutions of our time: data. Where data scientists became explorers of the digital realm. As with any human revolution, there are divided opinions on the matter. Some accept the idea that the world is ruled by data and others won't accept it. Other critical sectors have emerged during this period. Cybersecurity stands out as one of the most vulnerable and essential systems affected by digitization. Cybersecurity has been under the radar recently across the internet (Tirumala et al., 2019). The improvement of smart devices and IoT made it necessary to strengthen cybersecurity and avoid unwanted traffic and malicious behaviours (Tirumala et al., 2019). Moreover, Cloud computing has injected some fresh air into all digital industries by giving new opportunities to manage software, have access to storage, quick and detailed analytics, artificial intelligence and machine learning, and many more services (Ahamad et al., 2021). It also offers a wide range of resources, faster and innovative systems, and a lot of flexibility in pricing, computing, storing, etc (Ahamad et al., 2021). As a master's student embarking on my dissertation, I aim to explore the intricate relationship between data science, cybersecurity, and cloud computing, shedding light on the potential consequences and offering insights for sustainable practices in the face of this multifaceted issue.

1.2 Background

Data science is the art of extracting, modelling, and making sense of datasets with usually huge amounts of data (Singh and Saxena, 2021). Data science is directly intertwined with scientific disciplines such as mathematics, statistics, and visualisations (Singh and Saxena, 2021). Furthermore, the complexity of the digital realm made huge amounts of data available at hand, easily transferable, and massively reproducible. This is where Big Data emerged to answer complex queries and improve performance of companies (Austin, 2018). Researchers at EDISON suggest that data science is a combination of big data and data driven technologies and there is a need to rethink and remodel our approach to data science (Demchenko et al., 2016). NIST defines data scientists as data practitioners who manipulate data efficiently but also have overlapping expertise in business, analytics, and programming to manage the full big data lifecycle of a project (NIST, 2015). There's still a lack of concrete education concerning data science and big data programs, although many universities in Europe and America are evolving rapidly and adapting to the demand (NIST, 2015). Big data has impacted how companies view and work with data (Saltz and Grady, 2017). It has opened the door to many new skills in the realm of data science, which is now considered to be a job that deals with any activity that touches data (Saltz and Grady, 2017).

Cybersecurity has been under the radar recently across the internet (Tirumala et al., 2019). A survey on frequency of internet usage in New Zealand showed that 90% of people use the internet daily with a highest percentage (approximately 31%) of people checking their devices a couple of times a day (Tirumala et al., 2019). Thus, making the exposure to hacks higher than it was in previous generations. Nowadays, the internet has increased its users by 900% between 2000 and 2016 and the use of social media by thousands of percents (Tirumala et al., 2016). This indicates the rapid growth of technology and its importance on our day to day life, especially amongst students since the improvement of e-learning and BYOD (Bring Your Own Device) (Tirumala et al., 2016). Most cybersecurity threats come from a lack of awareness which leads to users being tricked by phishing messages either on social media, by message or by mail (Tirumala et al., 2016). Since the second half of the 20th century humankind has been investing massively in information, data, and technology, leading to high levels of improvement in some areas but also massive vulnerabilities in terms of cybersecurity, especially with the IoT technologies (Baranov and Kravchuk, 2021). It is considered that everything will be digitalized and IoT interconnected in the near future but this has to be intertwined with solid cybersecurity measures and policies (Baranov and Kravchuk, 2021). The fact that software and hardware are constantly changing and evolving creates a breach that could be exploited by attackers (Burns et al., 2013). Cybersecurity is a vast and complex scientific area that covers many disciplines related to network, data, computers, and the technologies used to keep them safe and protected (Kolar et al., 2021). Therefore, cybersecurity is a specialty that uses cybersecurity science methods (Kolar et al., 2021).

A recent study on the cloud has forecasted a 67% adoption of the cloud by IT companies (FutureScape, 2016). The cloud has increasingly offered many alternatives to the traditional workflow of information treatment going from using servers with predefined computing that could scale on demand to virtualization and containerization that could optimise cost efficiency and reduce management overload (Lynn et al., 2017). Thus, giving the option of using hardware or containers also known as FaaS or Function-as-a-Service (Lynn et al., 2017). Governments and information technology companies have dedicated a massive focus to the security in the cloud by adapting security standards into the realm of cloud computing (Di Giulio et al., 2017). Some standards involved are ISO 27001, C5, and FedRAMp (Di Giulio et al., 2017). Security standards incorporate a multitude of indicators that address IT security concerns that should be checked and verified for safe usage, especially in the cloud (Di Giulio et al., 2017). Cloud computing is not only an internet based computing system that gives access to computing and storage on demand, it has a variety of services that could benefit any industry in any domain (Gowrigolla et al., 2010). As an example, AWS (Amazon Web services) have more than 200 services ranging from computing, storage, virtual networks, streams, machine learning, security, firewalls, databases, and much

more. There are 3 deployment models available in the cloud: public, private, and hybrid (Gowrigolla et al., 2010). Public cloud is owned by the cloud provider and is available for use on demand or on reservation. On the other hand, private cloud is getting exclusive servers from the cloud provider for the companies only usage (Cavoukian and Rossos, 2002). And finally, hybrid cloud is a combination of public and private that gives the company the ability to use the cloud for a part of their workflow and to keep private servers for critical and secret data.

A powerful machine learning service developed by AWS, called SageMaker brings new opportunities to the table. Sagemaker is a fully managed machine learning service. It plays a crucial role in simplifying the intricate process of building, training, and deploying machine learning models at scale. With a comprehensive set of tools and services, SageMaker assists data scientists and developers throughout the entire machine learning workflow. Some noteworthy features and components of AWS SageMaker are its capabilities for data exploration, algorithm selection, model training, deployment, data labelling, model monitoring, and seamless integration with other AWS services. Security is a critical aspect of the cloud, making data the main source of vulnerability exploited by attackers (Nathezhtha and Yaidehi, 2018). Usually, the cloud offers a solid protection for external attacks but it's a different story with internal attackers (Nathezhtha and Yaidehi, 2018). The increasing demand in technology has created new opportunities for hackers to attack from, thus creating the need for new perspectives in security measures, and cybersecurity guidelines (Jha et al., 2022). The cloud offered an increase in economic efficiency but has known an increase in security threats (Chkirbene et al., 2021). Data scientists proposed machine learning security models, especially classification models (Chkirbene et al., 2021). Considering that cloud computing increased its revenues in the UK to more than 15 Billion dollars, machine learning combined with cybersecurity guidelines became necessary skills to master (Kumar et al., 2022).

1.3 Motivations

The motivations behind this project are numerous and they all involve an in-depth understanding of new technologies that rule the IT world. The goal is to contribute to the existing body of knowledge by exploring the relationships between data science, cybersecurity, and cloud computing at a global scale, then niche down into studying the impact of security threats on PDFs using AWS Sagemaker. By examining available data, assessing historical trends, and utilising advanced analytical techniques, this study seeks to uncover the key drivers of cybersecurity impacts on PDFs. It will be a necessary challenge that will stretch the current skill set and make it as developed as possible to get a position as a data scientist after the master's degree.

The topic chosen is focused on detecting anomalies and security threats in PDFs via the cloud. Since more than 70% of companies are using the cloud these days it became

critical to protect the data from external attacks. Moreover, the category of data chosen for this project is PDF which stands for portable document format and it is the most used document format globally. There are many factors that make Adobe PDF files popular. First, PDFs are easily readable on all devices. Second, PDFs are nicely structured and adapt to any screen. Lastly, they are not heavy on storage. Because of that, PDFs have been targeted by many hackers making them subject to cybersecurity threats and this paper will uncover all the features and attributes that are exploited by unethical attackers.

In a generation dominated by the cloud and especially by AWS, it has naturally come in mind to use its machine learning services in order to give life to this project. The main machine learning service provided by AWS is called Sagemaker and it will be the main tool for this project. It is easily integrated with Jupyter notebooks and will be used for the entire coding of this project.

1.4 Flow of the report

Section 1 will uncover the literature review and will dive in depth on notions related to data science, cybersecurity, and cloud computing. Section 2, will explain the chosen methodology of this paper. Section 3, will discuss and analyse the key concepts and results of the machine learning models. Section 4, will address the final conclusions and future work. And finally, section 5 will conclude the whole project by pointing to some key learnings.

1.5 Aims, Objectives and Research Questions

The aim of this project is to study and investigate cybersecurity threats in PDFs via the cloud and try to minimise the security threats and anomalies by developing a machine learning algorithm deployed in a cloud environment (AWS Sagemaker).

The research objectives are as follow:

- To define the current trends in cybersecurity and security threats in the cloud.
- To analyse the literature and extract features on the relationship between machine learning, cloud computing, and cybersecurity.
- To identify the machine learning algorithms mostly used to detect security threats and anomalies in the cloud.
- To create an optimised cloud environment using Sagemaker in AWS.
- To develop a machine learning algorithm that could detect anomalies and security threats in PDFs.
- To deploy and test the accuracy of the machine learning algorithm on AWS Sagemaker.
- To identify areas of improvement and future work

In this project, the algorithm developed should get high accuracy predictions to keep PDFs deployed in the cloud as safe as possible. Thus, the research questions are as follows:

- Is it safe to deploy PDFs in the cloud?
- What are the features affecting the most PDFs?
- Which machine learning algorithms detect malicious PDFs better?
- How to keep the application as safe as possible while deploying it in AWS with Sagemaker?

Chapter 2: Literature Review

2.1 Introduction

The evolution of technology led to a significant improvement in people's lives. Making it easier and giving more opportunities to innovate, create, and inspire. Furthermore, companies started investing heavily on information technology. Thus, developing super machines with highly efficient CPUs, massive memories, and all these nicely packaged into small devices. The Covid-19 pandemic has forced companies and individuals to focus their attention into the digital realm, increasing software and hardware production to satisfy the demand. This massive increase in online navigation has opened the door to many security threats, and the topic of cybersecurity became highly discussed. Hackers have always been creative in the way they access people's data, even with all the effort from tech providers, it has been a massive struggle over the years to minimise security threats and anomalies. One of the most used types of files has also been exposed to external attacks. Everyone nowadays is familiar with PDFs but not many people are aware that they represent a high risk of security threats.

2.2 Cybersecurity

Cybersecurity has dominated the internet for the past decades (Tirumala et al., 2019). This is mainly due to the presence of smart devices all around us (Tirumala et al., 2019). The introduction to IoT in recent years has opened the door to many security breaches. Software companies are constantly updating softwares to fix anomalies that could potentially be exploited by hackers. The internet became a necessity nowadays, with a massive growth of 900% of new users during the past decade (Tirumala et al., 2019). The introduction to BYOD (Bring Your Own Device) in schools has become normal and is applied in most institutions and is subject to high external cybersecurity attacks in developed countries (Tirumala et al., 2019). By adopting e-learning methods for schooling, the need for protection has drastically increased, thus raising concerns over data privacy and investing heavily in cybersecurity experts (Tirumala et al., 2019). Cybersecurity studies in New Zealand have analysed the level of concern related to data privacy and the statistics are shocking with over 93% of the population concerned by security threats and 34% concerned with security breaches (Tirumala et al., 2019). The topic of cybersecurity touches the whole world since the wide adoption of computers and the ease of access to the internet. Antiviruses are not enough anymore to protect individuals from external attacks. However, by educating ourselves about safety measures and the dos and don'ts of the internet, we could make our lives much safer. People willingly share their information on the internet and especially on social media. Facebook has grown over 28x times in 2015 (Tirumala et al., 2019). Facebook is an example of companies holding gigantic amounts of data on people around the world. This data is exploited in ways that are hidden from the public and the only goal is to leverage the data to control people's thoughts to decide their next moves, which could

be to buy a product or stay longer on the platform or millions of other reasons. Data has been widely collected from mobile devices to improve the e-learning experience of students (Pogarčić et al., 2013). (Tirumala et al., 2016) have conducted a study on cybersecurity terms familiarity for children and youngsters aged 8-21. What they noticed is that awareness increases over time and most youngsters are familiar with the main cybersecurity terms at age 18 to 21 and shown in figure 1 (Tirumala et al., 2016).

	08-12	13-17	18-21
Firewall	34	79	84
Privacy	22	65	74
Tracker	0	18	23
Private Mode	2	25	67
Antivirus	67	87	95
Phishing	1	27	43
Security Warnings	56	78	69

Figure 1: Awareness of cybersecurity terms for various age groups

Moreover, cybersecurity threats apply to all sectors, not just education. It touches heavily on public administrations, healthcare, law, economics, and more (Baranov and Kravchuk, 2021). The second half of the century has been widely dedicated to digitization and informatization (Baranov and Kravchuk, 2021). Also, the rapid and constantly changing hardware and software have created a gap between innovation and security (Burns et al., 2013). This gave the opportunity for hackers to find new ways to extort information from companies and individuals. These factors demonstrate the cruciality of cybersecurity in today's generation. Everyone is exposed to security threats, therefore making it mandatory to educate all generations on the effective use of smart devices and the internet.

A study conducted on security breaches concerning Zoom company in 2023 highlighted some key concepts of cybersecurity and the main reasons behind security breaches, here they are:

- **Risk**: Represents the probability of being hacked. Imagine having a weak password (i.e: 01234), you'll definitely increase your chances of being hacked.
- **Breach**: The hacker already has access to the data by phishing, ransomware, malware... And so many other hacking techniques.
- **Vulnerability**: This is due to a weakness in the system that makes you vulnerable to an external attack..
- **Hacker**: Hackers find ways to access the system and extract data from it. There are 3 types of hackers known to this date, white, grey, and black hackers. White

hackers usually work for companies and help them secure their system and avoid external hacks. On the other hand, black hackers are those who extort data illegally and make bad use of it. Grey hackers are in between, they are usually not harmful but work for their own benefit.

As mentioned before, Covid-19 has been a major event that made companies move all their data online and digitalize everything. Companies were obliged to work remotely thus using video calls instead of in person meetings. Furthermore, individuals became used to navigate the internet and make purchases online via e-commerce websites. Most shops shifted into cashless systems which decreases the headache of managing paper money. Smart devices have been all over the place and widely adopted by the population like watches, phones, TVs, fridges and much more.

A study conducted in 2009 found that 91% of companies were affected by security attacks (Nissim et al., 2014). Hackers usually take advantage of people to access the data, some of the widely known attacks are social engineering and phishing (Nissim et al., 2014). Social engineering is tricking a person to provide you secret data by making them think it's someone close to them (i.e: their boss or friend). Phishing is a sneaky way to hide a dangerous file sent via email, direct message, or also included in PDFs. In all these attacks, the hacker tries to incorporate a friend's name or company name or someone close to you so that you click on the link without much thinking.

2.3 PDFs & Security threats

PDFs or Portable Document Formats have been widely exploited by attackers, sending malicious PDFs via email and once opened triggering the installation of a trojan horse which gives the hacker the access to one's computer (Nissim et al., 2014). PDFs have been around the corner for over 30 years since their creation by Adobe in 1992 (Rahman et al., 2023). PDFs have a normalised structured as shown in figure 2 that consists mainly of (Yerima et al., 2022):

- The header: Contains the version of the pdf (i.e: %PDF-1.5)
- The body: contains objects, embedded files, text, images and all the code inside the PDF.
- The cross-reference table (x-ref): Contains offsets that allow access to the content of the PDF.
- The trailer: enables readers to quickly find x-ref and special objects

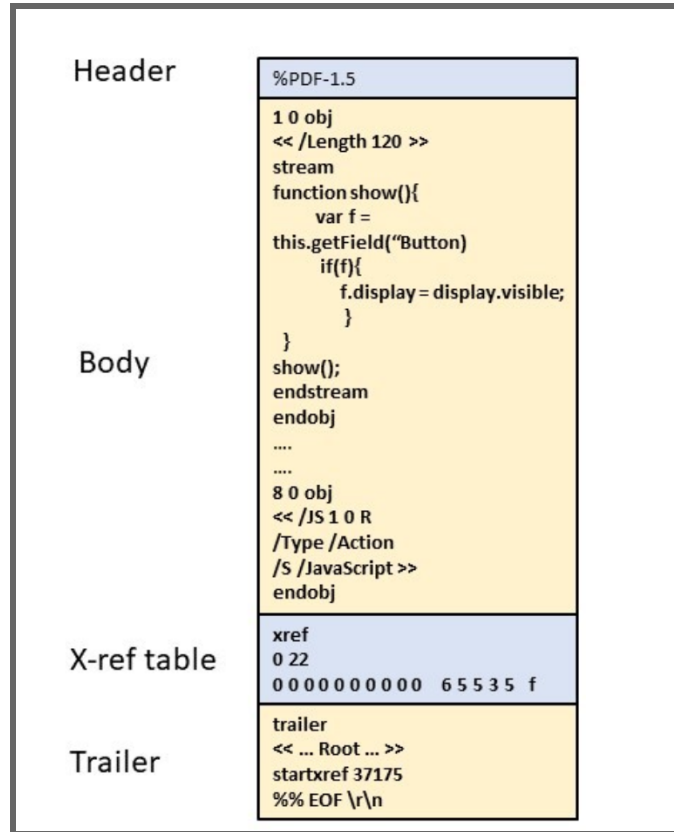


Figure 2: The sections of a typical PDF file

A study made in 2008 by F-Secure's has shown that the most targeted types of files are PDFs and word documents (Nissim et al., 2014). This pushed antivirus vendors to integrate machine learning models in order to detect anomalies and threats in PDFs (Kiem et al., 2004). However, experience has shown that these solutions were ineffective and PDFs were still being used to hide malicious programs (Nissim et al., 2014). PDFs contain a multitude of files and components inside of them, we can cite HTML, Javascript, EXE, images, objects... (Nissim et al., 2014). Malicious code is usually hidden in one of these elements (i.e: inside javascript code) making it difficult to detect without proper knowledge (Nissim et al., 2014). PDFs have been widely adopted by millions of users due to its portability and flexibility (Mohammed et al., 2021). Hackers these days exploit PDFs and spread malware across the internet. Adobe's attempt to create micro-programs from PDFs has opened the doors to install malicious programs hidden inside the code on the user's system itself (Mohammed et al., 2021). There are two main types of features present in a PDF document (Mohammed et al., 2021). The first features are called static, for example metadata, header, colour... The second features are called dynamic and they englobe everything related to the code itself such as Javascript, executables, embedded files,... (Mohammed et al., 2021) investigated a multitude of statistical methods that have been used to study the behaviour of malicious content in PDFs, we can cite:

- Binary Analysis: Analysing PDF malware files by extracting image, audio, and hash features from the binary representation.
- Keyword Analysis: Examining keywords in the PDF code to identify characteristic tags indicating malicious behaviour.
- Holistic Feature Fusion: Combining different malware detection methods to create an integrated confidence score for PDF malware detection.
- Dynamic Analysis: Running malware samples and observing their behaviour to detect and stop infections using a dynamic analysis-based approach.
- Obfuscation Removal Strategy: Disabling malicious content in PDF files by manipulating object tags using case-sensitive modifications. Also, using this technique as an augmentation method to generate additional samples for malware research.

Based on the research (Mohammed et al., 2021), the holistic feature approach gave an accuracy of 99.92% by combining features, including images, audio, and keyword-based structure.

The limitations of antivirus softwares lie in the fact that it is difficult to detect embedded code inside of JS code due to dynamic exploits from hackers (Rahman et al., 2023). Dynamic exploits refers to embedding objects, code, or cryptographic content inside of the PDF (Rahman et al., 2023). Thus, creating the necessity of developing machine learning and deep learning models to increase the odds of detecting malicious content inside PDFs (Rahman et al., 2023).

2.4 Artificial intelligence

[2.4.1 Broad introduction to AI](#)

Artificial intelligence represents technologies that aim to emulate human intelligence in various ways either by performing simple routine tasks or more complex problems that require high computation (Sumari and Syamsiana, 2021). Herbert Simon, a cognitive psychologist said that artificial intelligence (AI) is used for two main things. Firstly, for increasing the ability of humans to think and secondly, for comprehending and understanding how humans think (Sumari and Syamsiana, 2021). Most routine tasks would be classified in the first category, by using AI to assist humans in the daily tasks it gives more time to think and improve. Most complex tasks on the other hand require massive calculations and big data processing, that's why the second category is inspired by how human's think and tries to reproduce neuron functions into AI systems (Sumari and Syamsiana, 2021). There are many uses for AI, we can cite classification, regression, recognition, prediction, and many more (Sumari and Syamsiana, 2021). Artificial intelligence has had a big impact on modern life so far (Kose and Vasant, 2017). Since the evolution of computer science related subjects such as electronics, supercomputers, machine learning, deep learning, artificial intelligence have grown on

top of all these IT improvements to solve highly complex real-world problems (Kose and Vasant, 2017). Although AI has improved many aspects of our lives, it has opened the door to anxiety and many people are yet to trust these new technologies (Kose and Vasant, 2017). For this reason, AI safety was created and the main goal is to make sure that AI stays safe and is used in a proper manner to benefit people and not harm them (Kose and Vasant, 2017). Covid-19 has drastically increased people's awareness of AI (Syzdykbayeva et al., 2021). People were forced to shift their whole workflow digitally and therefore being exposed to new AI technologies. Figure 3 shows the publications on AI since 1999 and we can clearly see the spike in 2019 which is the exact time when covid-19 appeared (Syzdykbayeva et al., 2021).

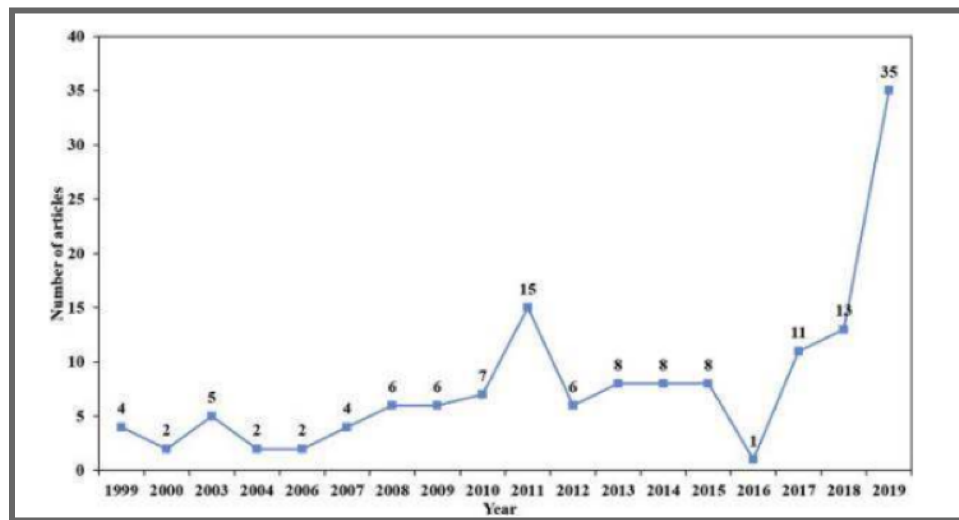


Figure 3: Analysis of publications on the introduction of Artificial intelligence in education

2.4.2 Machine learning & deep learning

There is usually a confusion about topics such as artificial intelligence, deep learning, and machine learning. They seem so interconnected and yet completely different. Figure 4 shows how intertwined and interconnected these disciplines are (Toma et al., 2022). Artificial intelligence englobes all technologies that emulate human's brain to either understand human's thinking or help with routine tasks. For that, subsections emerged such as machine learning and deep learning but they are all part of the same realm. Machine learning (ML) is basically an algorithm that learns patterns from the data and makes predictions (Nassif et al., 2021). Those predictions are possible due to a combination of statistics and computer science (Nassif et al., 2021). Deep learning on the other hand, is built on neural networks and is used for complex big data calculations involving image processing, voice treatment, face recognition, video processing, and more.

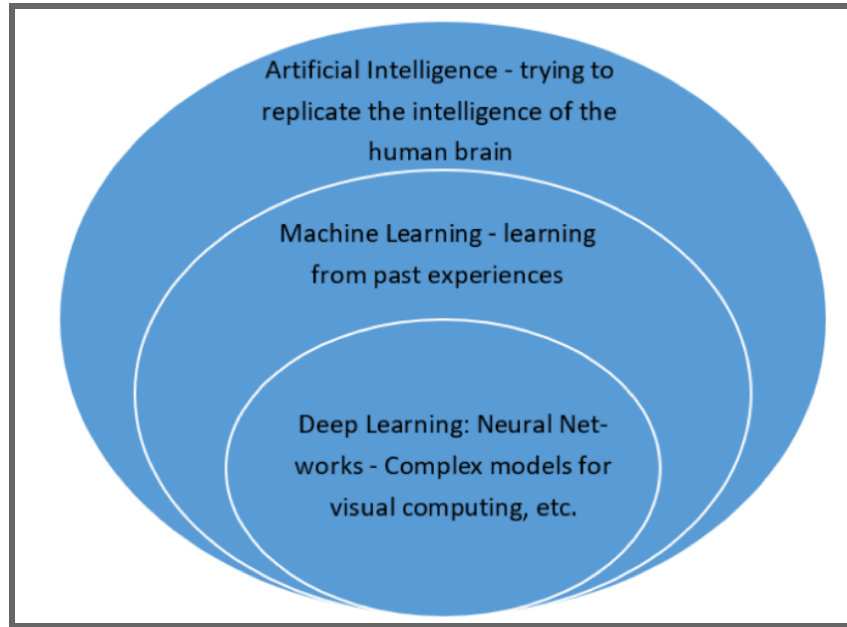


Figure 4: AI / DL / ML Hierarchy

There are three main categories of learning in machine learning: supervised, unsupervised, and semi-supervised (Arshad et al. 2021; Kwon et al., 2019). Supervised machine learning algorithms ingest data aiming to extract a specific output (classification and regression). Figure 5 illustrates a supervised machine learning model (Wang et al., 2022).

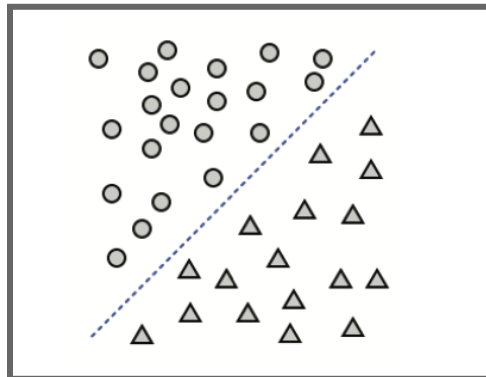


Figure 5: Supervised machine learning (Classification)

Unsupervised machine learning algorithms make sense of the data for humans to interpret (clustering and decomposition). Figure 6 illustrates an unsupervised machine learning model (Wang et al., 2022).

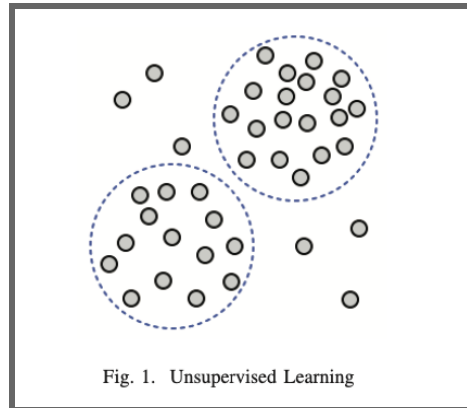


Figure 6: Unsupervised machine learning (Clustering)

Semi-supervised learning which is a combination of supervised and unsupervised learning with more common supervised learning methods.

There are also other categories such as reinforcement learning and transfer learning. Reinforcement learning is a method that continuously improves the ML model by putting an agent that measures the consequences of actions (Wang et al., 2022). Transfer learning is a reusable algorithm that solves a specific problem to different situations as shown in figure 7 (Wang et al., 2022).

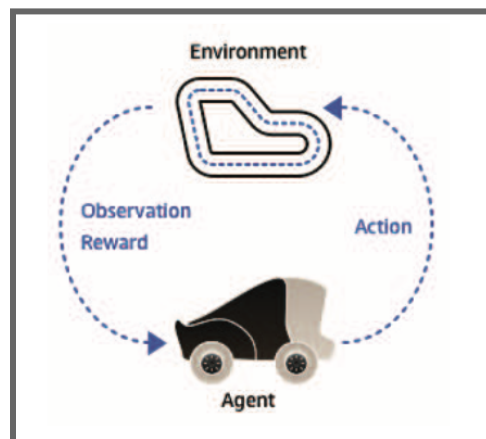


Figure 7: Reinforcement learning

2.4.3 Machine learning models

Supervised models for classification:

The goal of supervised classification models is to learn from existing data and predict new events (Banu and Kumar, 2022). There are tons of interesting classification models that have been developed throughout the years each model has unique advantages but also disadvantages, let's demystify some of them mostly used for malware classification (Gopaldinne et al., 2021):

Logistic Regression:

Logistic Regression is a statistical classifier commonly used for binary classification (Brownlee, 2016). It uses the logistic function to model the relationship between input values and a binary output. Coefficient values are used to linearly combine inputs, and the output is modelled as a binary value (Brownlee, 2016).

Advantages of Logistic Regression (LR):

- Simplicity and interpretability of the model.
- Efficient training and fast prediction time.
- Provides probabilistic outputs for classification tasks.

Disadvantages of Logistic Regression (LR):

- Limited ability to capture complex relationships.
- Assumption of linearity between features and the log-odds of the outcome.
- Susceptibility to outliers and multicollinearity.

Naive Bayes:

Naive Bayes is a supervised learning algorithm inspired by the Bayes theorem (Asiri, 2018). It assumes that the attributes are conditionally independent and uses probabilistic classification. It is scalable to large datasets and can handle high-dimensional training data (Asiri, 2018).

Advantages of Naive Bayes (NB):

- Fast training and prediction time.
- Works well with high-dimensional data.
- Performs well even with limited training data.

Disadvantages of Naive Bayes (NB):

- Assumes independence between features, which may not hold true in real-world scenarios.
- Limited expressiveness and inability to capture complex relationships.
- Sensitivity to irrelevant features.

Decision tree:

A decision tree is a hierarchical tree-like structure that represents a series of decisions or choices (Banu and Kumar, 2022). It is constructed based on training data and allows for easy interpretation and visualisation of the decision-making process. It is commonly used for classification and regression tasks (Banu and Kumar, 2022).

Advantages of Decision Tree Model:

- Interpretable and easy to understand.
- Can handle both numerical and categorical data.
- Captures complex relationships and non-linearity.

Disadvantages of Decision Tree Model:

- Prone to overfitting.

- Sensitive to small changes in data.
- Lacks global optimization.

XGBoost Classifier:

XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm that belongs to the boosting family of algorithms (Rahman et al., 2023). It is a powerful and efficient implementation of gradient boosting machines. The XGBoost classifier is specifically used for classification tasks. It involves building an ensemble of weak learners, such as decision trees, and iteratively adding them to create a strong learner. The classifier optimises a loss function using gradient descent to minimise errors and improve predictive accuracy. It allows for fine-tuning of hyperparameters, such as learning rate and maximum depth, to achieve the optimal model for classification tasks (Rahman et al., 2023).

Advantages of XGBoost Classifier:

- High performance and efficiency.
- Built-in regularisation techniques.
- Provides feature importance metric.

Disadvantages of XGBoost Classifier:

- Complexity in hyperparameter tuning.
- Memory-intensive.
- Lack of interpretability.

Random Forest:

Random Forest is a decision tree-based ensemble predictor (Tian et al., 2009). It uses a random vector sampled independently for each tree in the forest and combines their predictions. Random Forest performs well on large datasets, handles missing data, provides estimates of errors, and is known for its high accuracy (Tian et al., 2009).

Advantages of Random Forest (RF):

- High accuracy and robust performance on various types of data.
- Handles high-dimensional feature spaces and large datasets well.
- Provides estimates of feature importance for interpretability.

Disadvantages of Random Forest (RF):

- Lack of interpretability of individual trees within the ensemble.
- Requires more memory and computational resources compared to individual decision trees.
- Longer training time compared to simpler models like Naive Bayes or Logistic Regression.

SGD (Stochastic Gradient Descent):

Stochastic Gradient Descent is an optimization algorithm commonly used for training machine learning models, particularly in large-scale or online learning scenarios (Rahman et al., 2023). It is an iterative method that aims to find the minimum of a loss function by updating the model's parameters in small steps based on the gradients of the loss function (Rahman et al., 2023).

Advantages of Stochastic Gradient Descent (SGD):

- Efficiency and scalability.
- Fast convergence speed.
- Well-suited for online learning.

Disadvantages of Stochastic Gradient Descent (SGD):

- Sensitivity to learning rate.
- Potential for convergence to local minima.
- Requires data preprocessing for optimal performance.

Support Vector Machine (SVM):

SVM is a popular supervised learning model that analyses data for classification and regression (Tian et al., 2009). It finds the optimal hyperplane by maximising margins, based on Vapnik's theory of computational learning. SVM can handle large datasets, has strong regularisation properties, and aims to minimise true error (Tian et al., 2009).

Advantages of Support Vector Machine (SVM):

- Effective in high-dimensional spaces.
- Versatile with different kernel functions to handle various data types.
- Robust against overfitting with the use of regularisation.

Disadvantages of Support Vector Machine (SVM):

- Computationally expensive for large datasets.
- Difficulty in choosing appropriate kernel functions and tuning hyperparameters.
- Limited interpretability due to the complex decision boundaries.

K-Nearest Neighbor (KNN):

KNN is a classification and regression algorithm that classifies new data based on the distance to its k closest training examples (Gopaldinne et al., 2021). It stores instances corresponding to training data points and predicts the class based on the majority vote of the nearest neighbours. KNN is resistant to noisy data due to averaging (Gopaldinne et al., 2021).

Advantages of K-Nearest Neighbour (KNN):

- Simplicity in implementation and understanding.
- No training phase, as it memorised the entire dataset.
- Can handle multi-class classification and regression tasks.

Disadvantages of K-Nearest Neighbour (KNN):

- High prediction time complexity, especially for large datasets.
- Sensitivity to the choice of distance metric and number of neighbours.
- Requires careful data preprocessing, including feature scaling, to handle varying ranges and units.

ANN (Artificial Neural Network):

An Artificial Neural Network is a computational model inspired by the structure and functioning of biological neural networks in the brain (Rahman et al., 2023). It is composed of interconnected nodes or "neurons" organised in layers. ANN is used for machine learning and pattern recognition tasks, including classification and regression. It consists of multiple hidden layers with different numbers of units and activation functions, followed by an output layer. The model uses various techniques, such as weight initialization, dropout layers, and specific activation functions, to prevent overfitting and improve performance (Rahman et al., 2023).

Advantages of Artificial Neural Network (ANN):

- Ability to learn complex nonlinear relationships.
- Automatic feature extraction.
- Potential for excellent generalisation.

Disadvantages of Artificial Neural Network (ANN):

- Computationally intensive and time-consuming training.
- Prone to overfitting.
- Lack of interpretability.

2.4.4 Machine learning biases

Building trustworthy AI has been a real challenge during the past decades and it still is. This is primarily due to biases that highly impact the final outcome of AI technologies.

Biases are generally classified into these categories (Wang et al., 2022):

- Algorithm bias: When inaccurate algorithms are employed
- Data bias: Incorrect sampling when the data used doesn't reflect on the whole dataset
- Prejudicial bias: feeding the models with assumptions instead of facts, for example "drivers are male"
- Measurement bias: incorrect measurements that affect the results
- Intentional bias: intentional input of false and discriminatory data in the model

2.5 Cloud computing

Cloud computing has been around for the past 30 years. Recent studies have forecasted 67% of cloud adoption by IT companies (FutureScape, 2016). It has brought to life services such as effective storage, computing, analytics, security, private networks, machine learning and more (FutureScape, 2016). Figure 8 gives a clear picture of some use cases of the cloud (Singh, 2015).

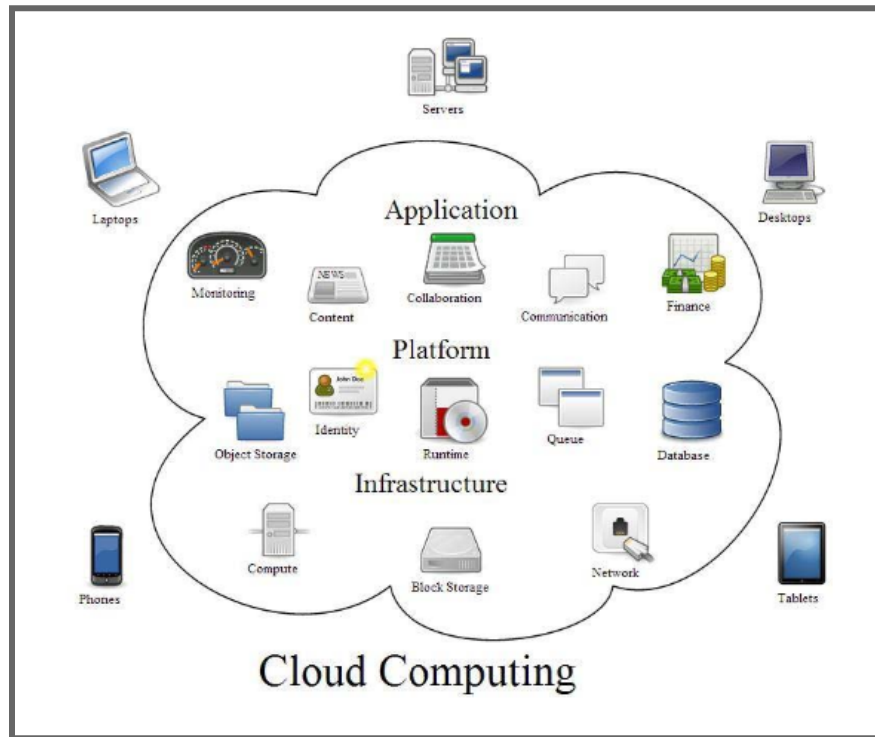


Figure 8: Cloud computing framework

Furthermore, it offers innovative and flexible resources with low maintenance cost (FutureScape, 2016). Cloud computing combined with AI have been the biggest IT breakthroughs of this generation providing a wide range of advantages for organisations as well as individuals (FutureScape, 2016). However, it has increased cybersecurity concerns with an important increase of security breaches (FutureScape, 2016). The flexibility of usage is a blessing for companies who don't have to fully rely on the cloud or choose to go all in on the cloud. Figure 9 shows a general representation of the cloud.

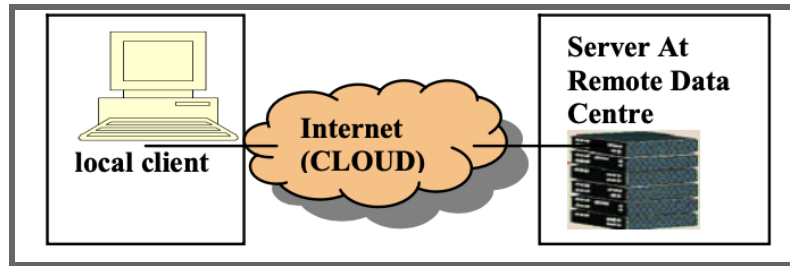


Figure 9: General representation of the cloud

Data in the cloud is a major concern and is highly regarded by cloud providers to ensure the safety of their client's data. Cloud providers such as AWS provide the platform, the infrastructure, and the software for companies or individuals to use (Gowrigolla et al., 2010). When considering data privacy we automatically think of privacy by design and default. (Gowrigolla et al., 2010) suggests that privacy by design is a proactive approach whose goal is to prevent attacks from happening in the first place. Privacy by default should be built into the system, even if an individual does nothing, their data should be kept private automatically (Gowrigolla et al., 2010). Figure 'n' shows a minimalistic architecture for cloud computing privacy especially when the data is encrypted and deployed onto the cloud (Gowrigolla et al., 2010).

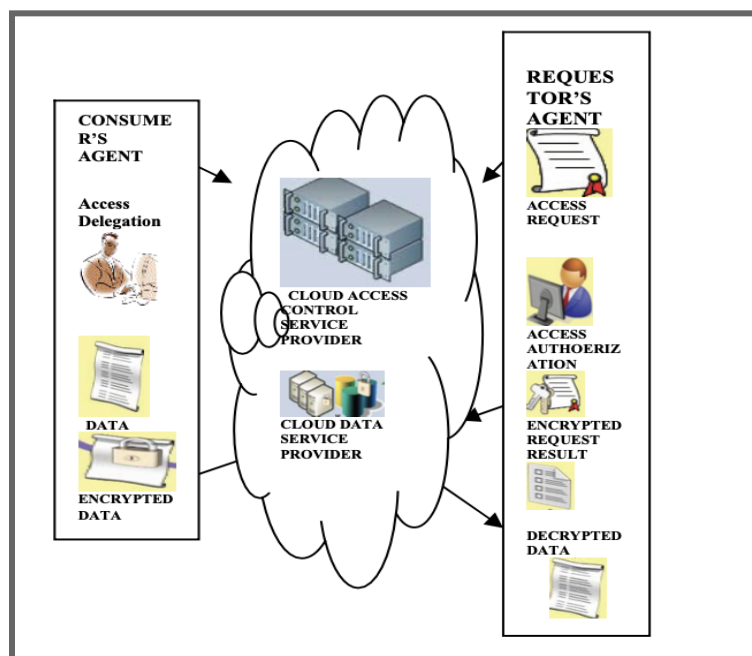


Figure 10: Privacy preserving cloud computing architecture

[2.5.1 Introduction to AWS](#)

Amazon Web Services (AWS) is the leader in cloud computing worldwide. It offers a range of services such as IaaS, HaaS, SaaS, FaaS, DaaS, and more with a pay as you

go model (Saraswat and Tripathi, 2020). This means that users pay only for the resources used (Kaur et al., 2018). Thus, making it one of the best choices for individuals and companies. AWS also offers machine learning and artificial intelligence services such as as Sagemaker (build, train, and deploy machine learning models), Polly (convert text to speech), Kendra (intelligent enterprise search), Lex (chatbots with conversational AI), Forecast (Forecast business outcomes), transcribe (convert speech to text) and much more (Mitchell, 2021).

2.5.2 AWS Sagemaker

AWS SageMaker is a fully managed machine learning service offered by Amazon Web Services (AWS) with the aim of simplifying the development, training, and deployment of machine learning models on a large scale (Mishra, 2019). It provides an extensive range of tools and services to assist data scientists and developers throughout the entire machine learning workflow.

Here are some notable features and components of AWS SageMaker (Mishra, 2019):

- Notebooks: SageMaker includes Jupyter notebook instances that facilitate tasks such as data exploration, preprocessing, and model development. These notebooks come pre-configured with popular machine learning frameworks and can seamlessly integrate with other AWS services.
- Built-in Algorithms: SageMaker offers a diverse selection of pre-built machine learning algorithms, including linear regression, classification, clustering, and image classification. These algorithms can be used as-is or customised to suit specific use cases.
- Model Training: With SageMaker, training machine learning models with large datasets becomes effortless. It supports distributed training across multiple instances to accelerate the training process. SageMaker also incorporates automatic model tuning for optimising hyperparameters.
- Model Deployment: After completing the training phase, SageMaker facilitates the deployment of trained models as web services, enabling easy integration into applications for real-time predictions. It provides automatic scaling and load balancing capabilities to handle production workloads effectively.
- Ground Truth: SageMaker Ground Truth is a service designed to aid in the labelling of data for training machine learning models. It combines human reviewers and automated labelling techniques to generate high-quality labelled datasets.
- Model Monitoring: SageMaker includes built-in functionalities for monitoring deployed machine learning models. It empowers users to detect concept drift, monitor data quality, and receive alerts for model retraining when necessary.
- Security and Integration: AWS SageMaker seamlessly integrates with other AWS services, allowing users to leverage resources like Amazon S3 for data storage,

AWS IAM for access control, and Amazon CloudWatch for monitoring purposes. It also incorporates built-in security features to ensure data privacy and compliance.

Chapter 3: Methodology

3.1 Introduction

This section outlines the methodology used in this research project, which aims to study and investigate cybersecurity threats in PDFs via the cloud and develop a machine learning algorithm deployed in a cloud environment (AWS SageMaker) to minimise security threats and anomalies.

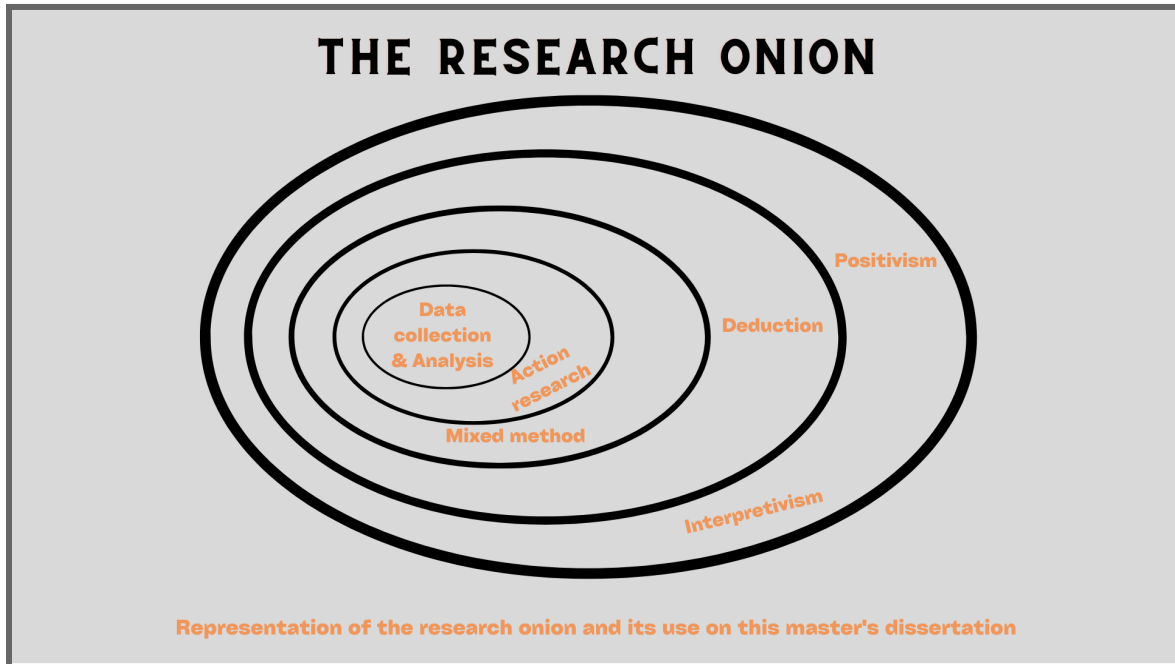


Figure 11: Research onion applied to this project

3.2 Research Philosophy

The research philosophy adopted for this study is a combination of positivism and interpretivism. Positivism will be used to analyse existing literature and identify current trends in cybersecurity, cloud computing, and machine learning algorithms. Interpretivism will be employed to gain a deeper understanding of the relationship between these concepts and to develop practical insights. Therefore, mixing between research and development.

3.3 Research Approach

The research approach for this study will be deductive, starting with a broad understanding of the research area and then narrowing down to specific research questions and hypotheses. The deductive approach will allow for the formulation and testing of hypotheses related to the effectiveness of machine learning algorithms in

detecting security threats and anomalies in PDFs deployed in the cloud as shown in figure 12 (Saunders et.al, 2009).

Deduction	
Logic	In a deductive inference, when the premises are true, the conclusion must also be true
Generalisability	Generalising from the general to the specific
Use of data	Data collection is used to evaluate propositions or hypotheses related to an existing theory
Theory	Theory falsification or verification

Figure 12: *Research methods for business students* (Pearson education)

3.4 Research Strategy

The research strategy employed in this study will be a combination of literature review and empirical analysis. The literature review will involve a comprehensive analysis of existing scholarly articles, research papers, and industry reports related to data science, cybersecurity, cloud computing, and machine learning. The empirical analysis will involve the development and implementation of a machine learning algorithm using AWS SageMaker, followed by testing and evaluation of its accuracy in detecting security threats in PDFs.

3.5 Research Design

The research design will be classified as action research, meaning it will include exploration and experimentation. The exploratory design will help in gaining insights into the complex relationship between data science, cybersecurity, and cloud computing, while the experimental design will allow for the development and testing of a machine learning algorithm.

3.6 Time Horizons

The time horizon used in this study is cross-sectional due to the pre-established and limited time available at hand for data collection and the completion of the project.

3.7 Data Collection

The data collection will consist of a newly reworked PDF dataset from The Canadian University of Cybersecurity called 'PDFMalware2022'. The Canadian university gives full access to individuals to use their data and thousands of people have already used their datasets in their research. The dataset contains 32 features and more than 10000 rows.

3.8 Data Analysis

The data analysis process will involve both qualitative and quantitative analysis techniques. Qualitative analysis will be conducted to gain insights into the features and attributes of PDFs that are exploited by attackers. This will involve a detailed examination of the content and structure of PDF files. Quantitative analysis will be performed to evaluate the performance of the machine learning algorithm in detecting security threats and anomalies in PDFs. This will include metrics such as accuracy, precision, recall, and F1 score.

3.9 Ethical Considerations

Ethical considerations will be given utmost importance throughout the research process. The data collected will be handled with confidentiality and privacy in mind. Any personal or sensitive information in the PDFs will be anonymized or removed. Additionally, proper permissions and approvals will be obtained for the use of any copyrighted materials.

3.10 Limitations

It is important to acknowledge the limitations of this research. One limitation is the availability and quality of the dataset. The sample dataset may not fully represent the diversity and complexity of real-world PDFs. Another limitation is the computational resources and time constraints for developing and testing the machine learning algorithm. These limitations may affect the generalizability and scalability of the findings.

3.11 Validity and Reliability

To ensure the validity and reliability of the research findings, appropriate measures will be taken. This includes using established machine learning techniques and algorithms, conducting rigorous data preprocessing and feature engineering, and performing statistical analyses. The results will be cross-validated using different evaluation methods and compared with existing literature and benchmarks.

3.12 Dissemination of Findings

The findings of this research will be disseminated through a comprehensive report, including a detailed analysis of the research findings, discussion of the implications and practical insights, and recommendations for future work

To summarise this section, a positivist research philosophy has been adopted, emphasising the use of quantitative methods to gather empirical evidence and test hypotheses. This philosophy aligns with the goal of uncovering objective truths about cybersecurity threats and developing a machine learning algorithm for their detection. Next, a deductive research approach has been employed, deriving specific hypotheses from existing theories and testing them with empirical data. By leveraging the existing theories and literature on cybersecurity, data science, and cloud computing, this approach provides a structured framework for hypothesis formulation and testing. The research strategy combines both quantitative and qualitative data collection and analysis techniques, utilising a mixed-methods approach. This allows for a comprehensive understanding of the subject matter, incorporating machine learning algorithms, statistical analysis, and expert insights. The time horizon chosen for this study is a cross-sectional design, collecting data at a specific point in time. This timeframe of three months provides sufficient opportunity to conduct experiments, gather data, and perform analysis, ensuring a comprehensive investigation. Data collection involves using a pre-existing dataset made available by The Canadian University of Cybersecurity. Data analysis techniques vary based on the nature of the data. Quantitative data collected from experiments will undergo statistical analysis, including descriptive statistics, correlation analysis, and machine learning algorithms, to identify patterns and relationships. Qualitative data obtained from experts will be analysed thematically, extracting key themes and patterns. Ethical considerations are carefully addressed throughout the research process, ensuring informed consent, participant privacy, and adherence to ethical guidelines and regulations.

3.13 Conclusion

In conclusion, this methodology provides a structured and comprehensive framework for investigating cybersecurity threats in PDFs via the cloud. The study incorporates a positivist and interpretivism research philosophy, a deductive research approach, a mixed-methods strategy, a cross-sectional design, and appropriate data collection and analysis techniques. This methodology lays the foundation for achieving the research objectives and addressing the research questions, contributing to the existing body of knowledge in cybersecurity, data science, and cloud computing.

Chapter 4: Dataset Experiments and Results

4.1 Introduction

The Canadian University of Cybersecurity published in 2020 their work on malware detection in PDFs by aiming to improve the shortcomings of the 'contagio' dataset which is one of the most popular PDF related dataset for malware detection. [Issakhane et al 2022] took a different approach to analysing PDFs, they categorised the data into two distinct categories: General features and structural features and they used a stacking architecture which is a combination of two machine learning models. This approach has given a very high accuracy to the research of 99.89%, however the model doesn't consider the demand in computing in a real-time world scenario. Using a stacking architecture with different machine learning models increases the computing needs to deploy the model and keep it running consistently. That's why in this paper the focus was not only put on getting the highest accuracy but also on minimising the cost of computing to make it fit into a real-life scenario by deploying the model on AWS Sagemaker.

4.2 Experiments and results

The dataset including all the features, exploratory analysis, visualisations, and experiments are described in this section. A combination of the well known jupyter notebooks and Python programming language have been used for this action research. Every graph and tables will be examined and explained in further details step by step.

[4.2.1 Dataset description](#)

The dataset has two distinct classes: Malicious and Benign. [Issakhane et al 2022] created a well balanced dataset to avoid all biased results during the machine learning process. Figure 13 shows the number of malicious and benign PDFs in the dataset that contains an overall of 10023 PDF samples.

Malicious	5555
Benign	4468
Name: Class, dtype: int64	

Figure 13: Balanced sample of malicious and benign PDFs in the dataset

The original dataset is constructed of 33 features categorised in two distinct categories: general features and structural features. Figure 14 indicates the correlation between different features and a few features stood out.

4.2.2 Correlation and important features

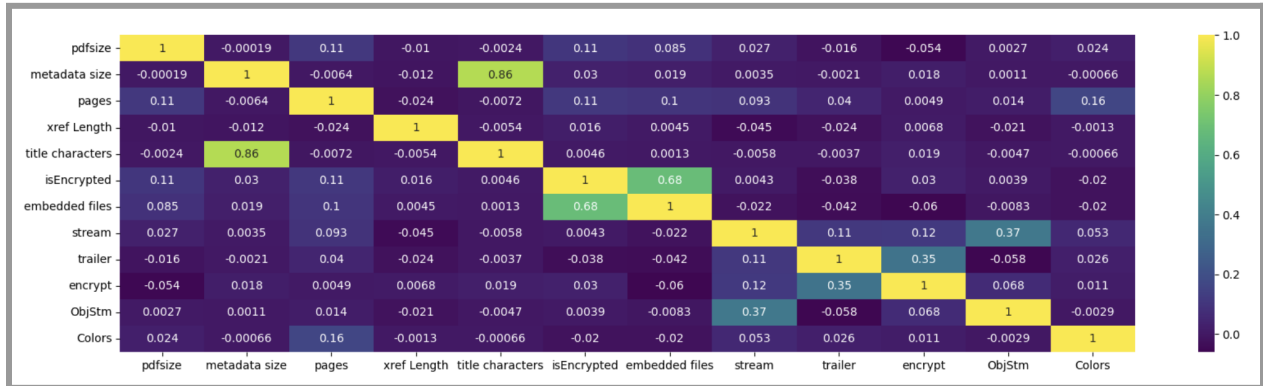


Figure 14: Correlation between independent features

There are two high correlations between 'Metadata size' and 'title characters'. Researchers found that shorter and non-impactful titles usually have a high risk of being harmful. The metadata size usually gives the details about the pdf such as the filename, date of creation, access time, access rights and so on. Second high correlation is between 'embedded files' and 'isencrypted'. There is a logical link between embedded files and encryption. An encrypted file is more secure, thus, making it difficult to embed any file on the pdf. Furthermore there are two important but lower correlations between 'stream' and 'objstm' which is basically a stream of data containing objects in the pdf and between 'trailer' and 'encrypt'. Trailer specifies how the application reading the PDF document should find the cross-reference table and other special objects. If it's not encrypted it gives easy access to external attacks. Scatter plots in figure 15 show the four correlations and their percentage.

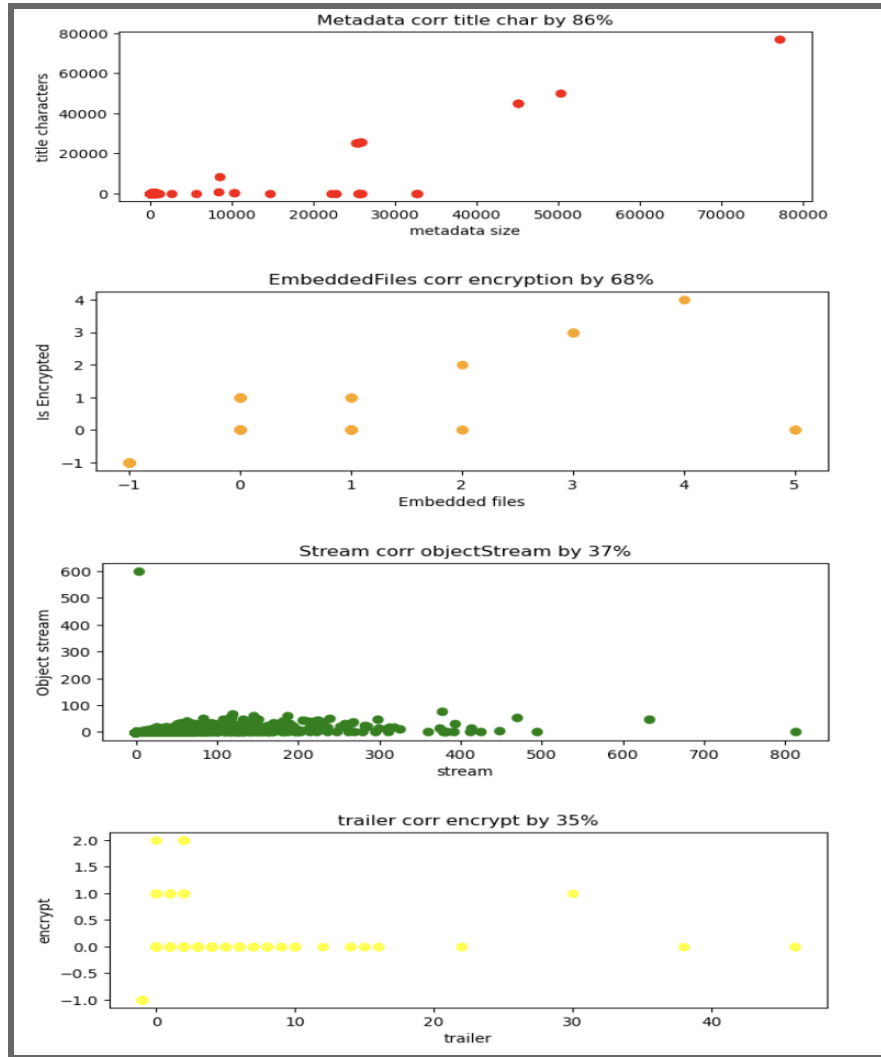


Figure 15: Visualisation of the highest correlations

4.2.3 Encryption effect on malicious behaviour

The dataset doesn't provide an indication on the effect of encryption on malicious behaviour in PDFs. As shown in figure 16, most PDFs used in this sample are non encrypted thus not giving much margin to analyse the potential of this feature. The usual role of encryption is to protect users from external attacks such as phishing, therefore adding a password encryption decreases the chances of being hacked.

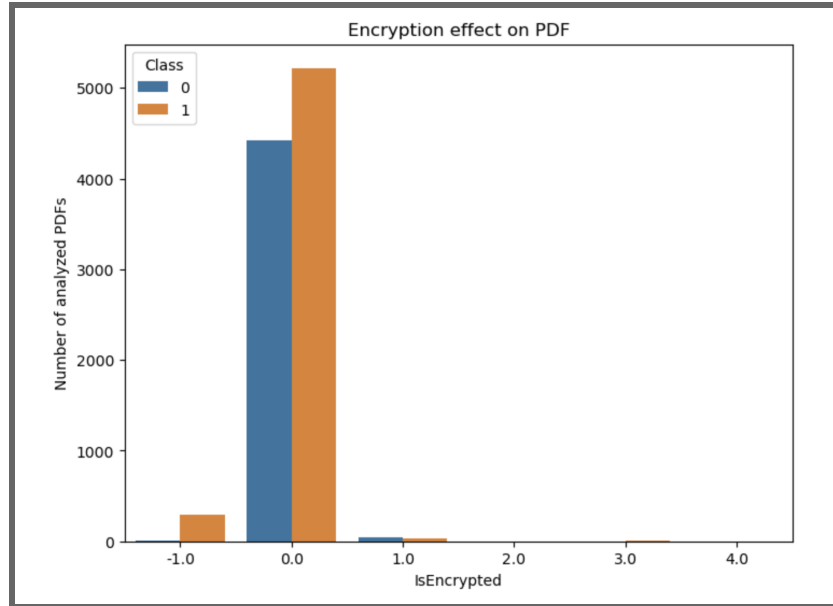


Figure 16: Encryption plot showing the effect on PDFs

4.2.4 New dataset proposed

The new dataset proposed has 17 independent features and 1 dependent output (Class). The choice was made after carefully analysing all the features and testing machine learning models on them. First, the focus was on the objects and media embedded inside PDFs thus the reason to choose the following features from the dataset:

→ embedded files | images | obj | RichMedia | EmbeddedFile

Second, this study considered including some relevant features to see how they impact the machine learning model, thus the choice of the following features:

→ AA | OpenAction: (" /AA and /OpenAction indicate an automatic action to be performed when the page/document is viewed ")

→ JBIG2Decode: The JBIG2Decode filter (PDF 1.4) decodes monochrome (1 bit per pixel) image data that has been encoded using JBIG2 encoding.

→ Javascript: This denotes the number of keywords containing javascript code

And finally, the features that had a high correlation when analysing the whole sample of data, which are:

→ metadata_size | title_characters | embedded_files | isEncrypted | stream | objstm | trailer | encrypt

To get the machine learning models working right there was a need to convert all the independent features into float as shown in figure 17 to avoid any issues during the training phase.

metadata_size	float64
title_characters	float64
isEncrypted	float64
embedded_files	float64
stream	float64
trailer	float64
encrypt	float64
ObjStm	float64
embedded_files	float64
images	float64
obj	float64
RichMedia	float64
EmbeddedFile	float64
AA	float64
OpenAction	float64
JBIG2Decode	float64
Javascript	float64
Class	object
dtype:	object

Figure 17: New dataset proposed features

[4.2.5 Machine learning models](#)

Figure 18 clearly indicated that two models give the highest accuracy:

- Decision Tree: 98.46% accuracy / training time: 14ms
- Random Forest: 99.07% accuracy / training time: 337ms

The second important thing to consider is the training and testing time. Since this model is going to be deployed in the cloud it is essential to know how much computing power the model requires. More computing means more resources allocated (bigger computing instances) and higher expenses.

	Logistic regression	Naive Bayes	Decision Tree	Random forest	Stochastic gradient descent	Support vector machine	K-nearest neighbour
Training time (ms)	88	2	14	337	15	2368	2
Testing time (ms)	89	4	15	355	16	3265	250
F1-score (%)	92.50	58.13	98.60	99.16	88.34	81.88	96.91
Accuracy score (%)	91.78	86.93	98.46	99.07	87.04	78.72	96.63

Riad Anas

Figure 18: Model computational report with various ML models

4.2.6 Machine learning model chosen

For this project, the machine learning model chosen is the 'decision tree' model since it gives a very high accuracy output and doesn't require much computing to deploy on AWS Sagemaker which will keep the project at a low cost. See figure 19 for a summary of the process followed during this project.

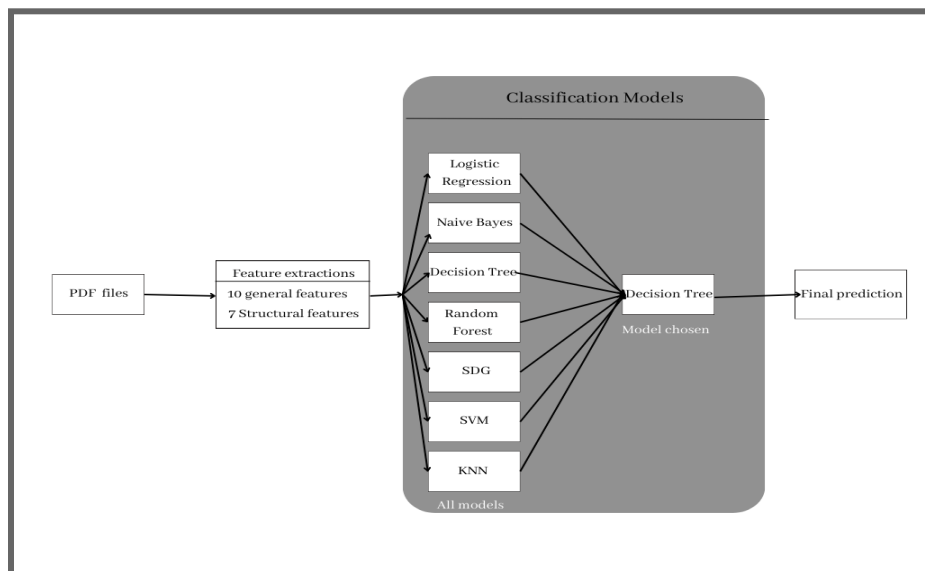
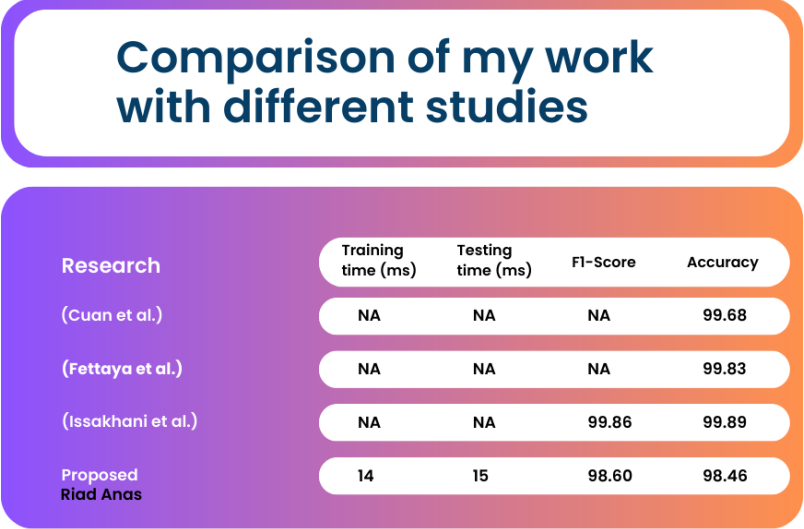


Figure 19: Full process of the project from new dataset to ML models to Final predictions

4.2.7 Comparison with previous work

Malicious content detection in PDFs is a work that has been done by few researchers during the past few years. Their goal was to find a more accurate way to detect security threats in PDFs since antiviruses weren't as effective as they used to be and hackers got more creative with how they inject malicious content. For that reason, researchers have tried a few approaches such as statistical models, machine learning models, and also deep learning models. Figure 20 shows a comparison of this study with previous studies made on the topic and shows that using machine learning models to detect security threats and anomalies in PDFs is as high as 99% which gives the confidence to normalise this approach instead of what was proposed by antiviruses during the past decades.



The figure is a table titled "Comparison of my work with different studies". It compares four research studies based on Training time (ms), Testing time (ms), F1-Score, and Accuracy. The studies are (Cuan et al.), (Fettaya et al.), (Issakhani et al.), and the Proposed method by Riad Anas. The Proposed method shows the lowest training and testing times while maintaining high F1-Score and Accuracy.

Research	Training time (ms)	Testing time (ms)	F1-Score	Accuracy
(Cuan et al.)	NA	NA	NA	99.68
(Fettaya et al.)	NA	NA	NA	99.83
(Issakhani et al.)	NA	NA	99.86	99.89
Proposed Riad Anas	14	15	98.60	98.46

Figure 20: Comparison of this research with previous work

4.2.8 Conclusion

This chapter explored the results of this study on detecting malicious content within PDFs. This topic touches primarily on cybersecurity and the importance of investing heavily in these studies to keep users and companies safe from external attacks. We've seen damaging data breaches and external attacks throughout history and these were all due to a lack of proper security and awareness. Though it is important to educate

people on how to use digital tools effectively, it is also important to build programs that could keep those same users safe in the first place. Thus the importance of building machine learning algorithms that can detect anomalies and security threats and for our study it is focused on PDFs. The results show that by analysing some important features and implementing random forest machine learning algorithms it is possible to detect more than 98% of malicious content in PDFs, which gives a high safety net for users out there. Other studies had a slightly better accuracy percentage but the resources invested were very high, thus making it costly and tricky to scale.

Chapter 5: Discussion

This action research helped me gain a better understanding of important concepts widely used in the information technology sphere. More importantly, it made me realise how critical it is to secure documents that seem safe in first regard but could keep harmful and hidden content inside. It has stretched my current skill set to gain deeper understanding of each topic and apply it into a real-life project and deploy it to AWS Sagemaker via Jupyter notebooks.

The importance of cybersecurity grows day by day since everything is digitised making it one of the top trending topics of our generation. This action research opened my eyes on the criticality of securing my workflow whether it is in the cloud or when sending a simple file via the internet. Researchers have gained a lot of confidence in this area through the massive growth and simplicity of use of machine learning and deep learning algorithms. Those relatively new tools gave us the power to make calculations that we couldn't dream of doing 20 years ago. That's why machine learning and cybersecurity nowadays go hand in hand, pushing each other to the limit and creating a safer environment for all users on the internet.

Machine learning has had many real-life applications to decrease cybersecurity threats on the internet. Some use cases involve using ML against SMS or Email scams to help people protect themselves from phishing attacks. Also, using ML to help people avoid typing errors which can lead to big issues for banks and governments. Furthermore, using ML against bots on social media to minimise the potential external attacks on social media users. And so many other use cases of ML applications out there to improve the safety of users on the internet.

Chapter 6: Conclusion and future work

In moving forward, there are several key areas where further exploration can yield valuable insights. Building on the success of this project, it is advisable to develop more sophisticated computer programs that are tailor-made to counter specific online threats. By enhancing the precision and adaptability of these programs, we can better safeguard digital assets from evolving cyber risks. Moreover, integrating blockchain technology with cloud security systems presents a promising avenue for reinforcing data integrity and authentication. Investigating the potential of quantum computing to revolutionise encryption and decryption processes offers a groundbreaking opportunity for advancing data protection strategies in the cloud. These advancements can contribute significantly to a more resilient and secure digital landscape.

However, it's important to acknowledge certain limitations in this research. The proposed solution, while effective for PDF security, may not cover all potential cybersecurity challenges. As the digital realm continues to evolve, new and unforeseen threats may emerge that require tailored solutions. Additionally, the use of emerging technologies like quantum computing may pose implementation challenges due to their complexity and evolving nature. Furthermore, ethical considerations must be an integral part of future endeavours in this field. Balancing the benefits of data science, cybersecurity, and cloud computing with the preservation of individual privacy and data rights is an ongoing challenge that demands careful attention. Striking the right balance between innovation and responsibility is crucial to ensuring the sustainable growth of these interconnected disciplines.

In conclusion, this dissertation has shed light on the symbiotic relationship between data science, cybersecurity, and cloud computing. By successfully devising a pragmatic solution to enhance PDF security in the cloud, this research has made a meaningful contribution to the ongoing dialogue surrounding data protection and digital integrity. As the digital landscape continues to evolve, these three domains will remain at the forefront of technological advancement, and their synergy will continue to drive innovation, resilience, and security in the digital age.

References

- Ahamad, S., Mohseni, M., Shekher, V., Smaisim, G.F., Tripathi, A. and Alanya-Beltran, J., 2022, April. A Detailed Analysis of the Critical Role of artificial intelligence in Enabling High-Performance Cloud Computing Systems. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 156-159). IEEE.
- Arshad, J., Townend, P. and Xu, J., 2013. A novel intrusion severity analysis approach for Clouds. *Future Generation Computer Systems*, 29(1), pp.416-428.
- Asiri, S., 2018. Machine learning classifiers. *Towards data science*.
- Austin, C.C., 2018, December. A path to big data readiness. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4844-4853). IEEE.
- Baranov, O. and Kravchuk, I., 2021, November. Internet of Things and the Problem of Cybersecurity. In 2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo) (pp. 39-42). IEEE.
- Blonce, A., Filiol, E., and Frayssignes, L. (2008). Portable document format (pdf) security analysis and malware threats. In *Presentations of Europe BlackHat 2008 Conference*.
- Brandis, R. and Steller, L. (2012). Threat modelling adobe pdf. Technical report.
- Brownlee, J., 2016. Logistic regression for machine learning. *Machine Learning Mastery*, 1.
- Burns, L., Linger, R. and Alves-Foss, J., 2013, January. Introduction to Software Cybersecurity, Assurance, and Testing Minitrack. In 2013 46th Hawaii International Conference on System Sciences (pp. 5012-5012). IEEE Computer Society.
- Carmony, C., Hu, X., Yin, H., Bhaskar, A. V., and Zhang, M. (2016). Extract me if you can: Abusing pdf parsers in malware detectors. In *NDSS*.
- Cavoukian, A. and Rossos, P.G., 2002. Information and privacy commissioner of Ontario. Personal health information: a practical tool for physicians transitioning from paper-based records to electronic health records URL: <http://www.ipc.on.ca/images/Resources/hipa-toolforphysicians.pdf> [accessed 2013-11-21][WebCite Cache ID 6LluKqbBr].
- Chkirbene, Z., Hamila, R., Erbad, A., Kiranyaz, S., Al-Emadi, N. and Hamdi, M., 2021, June. Cooperative machine learning techniques for cloud intrusion detection. In 2021 International Wireless Communications and Mobile Computing (IWCMC) (pp. 837-842). IEEE.
- Corona, I., Maiorca, D., Ariu, D., and Giacinto, G. (2014). Lux0r: Detection of malicious pdf-embedded javascript code through discriminant analysis of api references. In *workshop on artificial intelligent and security workshop*, pages 47–57.

- Cross, J. S. and Munson, M. A. (2011). Deep pdf parsing to extract features for detecting embedded malware. Sandia National Labs, Albuquerque, New Mexico, Un- limited Release SAND2011-7982.
- Cui, Y., Sun, Y., Luo, J., Huang, Y., Zhou, Y., and Li, X. (2020). Mmpd: A novel malicious pdf file detector for mobile robots. IEEE Sensors Journal.
- Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M. and Brewer, S., 2016, December. EDISON data science framework: a foundation for building data science profession for research and industry. In 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom) (pp. 620-626). IEEE.
- Di Giulio, C., Sprabery, R., Kamhoua, C., Kwiat, K., Campbell, R.H. and Bashir, M.N., 2017, June. Cloud standards in comparison: Are new security frameworks improving cloud security?. In 2017 IEEE 10th International Conference on Cloud Computing (CLOUD) (pp. 50-57). IEEE.
- FutureScape, I.D.C.I.D.C., Worldwide IT Industry 2017 Predictions; 2016. IDC.
- Gopaldinne, S.R., Kaur, H., Kaur, P. and Kaur, G., 2021, April. Overview of pdf malware classifiers. In 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 337-341). IEEE.
- Gowrigolla, B., Sivaji, S. and Masillamani, M.R., 2010, December. Design and auditing of cloud computing security. In 2010 Fifth International Conference on Information and Automation for Sustainability (pp. 292-297). IEEE.
- Jeong, Y.-S., Woo, J., and Kang, A. R. (2019). Malware detection on byte streams of pdf files using convolu- tional neural networks. Security and Communication Networks, 2019.
- Jha, S., Arora, M., Sharma, Y., Anand, A. and Sharma, D., 2022, May. Comparative Analysis of Cloud Computing Based Face Recognition Services. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (Vol. 1, pp. 567-572). IEEE.
- Kaur, A., Raj, G., Yadav, S. and Choudhury, T., 2018, December. Performance evaluation of AWS and IBM cloud platforms for security mechanism. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 516-520). IEEE.
- Kiem, H., Thuy, N.T. and Quang, T.M.N., 2004. A machine learning approach to anti-virus system. training, 500, p.1.
- Kolar, V., Delija, D. and Sirovatka, G., 2021, September. Open-Source Intelligence as the New Introduction in the Graduate Cybersecurity Curriculum. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 650-653). IEEE.

- Kose, U. and Vasant, P., 2017, September. Fading intelligence theory: A theory on keeping artificial intelligence safety for the future. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-5). IEEE.
- Kumar, S., Prethi, K.A., Singh, S., Lourens, M. and Patil, N., 2022, April. Role of machine learning in managing cloud computing security. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 2366-2369). IEEE.
- Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I. and Kim, K.J., 2019. A survey of deep learning-based network anomaly detection. *Cluster Computing*, 22, pp.949-961.
- Li, Y., Wang, Y., Wang, Y., Ke, L., and Tan, Y.-a. (2020). A feature-vector generative adversarial network for evading pdf malware classifiers. *Information Sciences*, 523:pp. 38–48.
- Liu, D., Wang, H., and Stavrou, A. (2014). Detecting malicious javascript in pdf through document instrumentation. In 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, pages 100–111. IEEE.
- Lynn, T., Rosati, P., Lejeune, A. and Emeakaroha, V., 2017, December. A preliminary review of enterprise serverless cloud computing (function-as-a-service) platforms. In 2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom) (pp. 162-169). IEEE.
- Mishra, A. (2019) Machine learning in the AWS cloud: Add intelligence to applications with Amazon Sagemaker and Amazon Rekognition, Amazon. Available at: <https://aws.amazon.com/sagemaker/> (Accessed: 05 July 2023).
- Mitchell, T.M. (2021) Machine learning, Amazon. Available at: <https://aws.amazon.com/machine-learning/> (Accessed: 07 July 2023).
- Mohammed, T.M., Nataraj, L., Chikkagoudar, S., Chandrasekaran, S. and Manjunath, B.S., 2021, November. HAPSSA: Holistic Approach to PDF malware detection using Signal and Statistical Analysis. In MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM) (pp. 709-714). IEEE.
- Nassif, A.B., Talib, M.A., Nasir, Q., Albadani, H. and Dakalbab, F.M., 2021. Machine learning for cloud security: a systematic review. *IEEE Access*, 9, pp.20717-20735.
- Nathezhtha, T. and Yaidehi, V., 2018, September. Cloud insider attack detection using machine learning. In 2018 International Conference on Recent Trends in Advance Computing (ICRTAC) (pp. 60-65). IEEE.
- Networks (CICN) (pp. 486-491). IEEE.[10]: [10]: Banu, A.N. and Kumar, K.R., 2022, November. Cloud-Based Machine Learning Techniques With Intrusion Detection System. In 2022 1st International Conference on Computational Science and Technology (ICCST) (pp. 281-286). IEEE.

- Nissim, N., Cohen, A., Moskovitch, R., Shabtai, A., Edry, M., Bar-Ad, O. and Elovici, Y., 2014, September. Alpd: Active learning framework for enhancing the detection of malicious pdf files. In 2014 IEEE Joint Intelligence and Security Informatics Conference (pp. 91-98). IEEE.
- NIST, S., 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, Sept 2015 [online] <http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf>.
- Pogarčič, I., Marković, M.G. and Davidović, V., 2013, May. BYOD: a challenge for the future digital generation. In 2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 748-752). IEEE.
- Rahman, T., Ahmed, N., Monjur, S., Haque, F.M. and Hossain, M.I., 2023, March. Interpreting Machine and Deep Learning Models for PDF Malware Detection using XAI and SHAP Framework. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-9). IEEE.
- Rahman, T., Ahmed, N., Monjur, S., Haque, F.M. and Hossain, M.I., 2023, March. Interpreting Machine and Deep Learning Models for PDF Malware Detection using XAI and SHAP Framework. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-9). IEEE.
- Saltz, J.S. and Grady, N.W., 2017, December. The ambiguity of data science team roles and the need for a data science workforce framework. In 2017 IEEE international conference on big data (Big Data) (pp. 2355-2361). IEEE.
- Samaneh MahdaviFar, Andi Fitriah Abdul Kadir, Rasool Fatemi, Dima Alhadidi, Ali A. Ghorbani; Dynamic Android Malware Category Classification using Semi-Supervised Deep Learning, The 18th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC), Aug. 17-24, 2020.
- Samaneh MahdaviFar, Dima Alhadidi, and Ali A. Ghorbani (2022). Effective and Efficient Hybrid Android Malware Classification Using Pseudo-Label Stacked Auto-Encoder, Journal of Network and Systems Management 30 (1), 1-34
- Saraswat, M. and Tripathi, R.C., 2020, December. Cloud computing: Comparison and analysis of cloud service providers-AWs, Microsoft and Google. In 2020 9th international conference system modeling and advancement in research trends (SMART) (pp. 281-285). IEEE.
- Saunders, M., Lewis, P. and Thornhill, A., 2009. Research methods for business students. Pearson education.
- Singh, A. and Saxena, N., 2021, December. Data Science: Relationship with big data, data driven predictions and machine learning. In 2021 International Conference on Computational Performance Evaluation (ComPE) (pp. 067-072). IEEE.

- Singh, M., 2015, May. Study on cloud computing and cloud database. In International Conference on Computing, Communication & Automation (pp. 708-713). IEEE.
- Stevens, D. (2011). Malicious pdf documents explained. IEEE Security & Privacy, 9(1):80–82.
- Sumari, A.D.W. and Syamsiana, I.N., 2021, November. A Simple Introduction to Cognitive Artificial Intelligence's Knowledge Growing System. In 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA) (pp. 170-175). IEEE.
- Syzdykbayeva, A., Baikulova, A. and Kerimbayeva, R., 2021, April. Introduction of Artificial Intelligence as the Basis of Modern Online Education on the Example of Higher Education. In 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-8). IEEE.
- Tian, R., Batten, L., Islam, R. and Versteeg, S., 2009, October. An automated classification system based on the strings of trojan and virus families. In 2009 4th International conference on malicious and unwanted software (MALWARE) (pp. 23-30). IEEE.
- Tirumala, S.S., Sarrafzadeh, A. and Pang, P., 2016, December. A survey on Internet usage and cybersecurity awareness in students. In 2016 14th Annual Conference on Privacy, Security and Trust (PST) (pp. 223-228). IEEE.
- Tirumala, S.S., Valluri, M.R. and Babu, G.A., 2019, January. A survey on cybersecurity awareness concerns, practices and conceptual measures. In 2019 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.
- Toma, C., Popa, M., Iancu, B., Doinea, M., Pascu, A. and Ioan-Dutescu, F., 2022. Edge machine learning for the automated decision and visual computing of the robots, IoT embedded devices or UAV-drones. Electronics, 11(21), p.3507.
- Tzermias, Z., Sykiotakis, G., Polychronakis, M., and Markatos, E. P. (2011). Combining static and dynamic analysis for the detection of malicious documents. In Proceedings of the Fourth European Workshop on System Security, pages 1–6.
- Wang, S.P., Arafin, M.T., Osuagwu, O. and Wandji, K., 2022, January. Cyber Threat Analysis and Trustworthy Artificial Intelligence. In 2022 6th International Conference on Cryptography, Security and Privacy (CSP) (pp. 86-90). IEEE.
- Yerima, S.Y., Bashar, A. and Latif, G., 2022, December. Malicious PDF detection Based on Machine Learning with Enhanced Feature Set. In 2022 14th International Conference on Computational Intelligence and Communication

Appendix