Stay Local or Go Global – A Comparison of Preference of Explainability

Lydia Channon
745065

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Master of Science

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

September 30th 2021

Declaration

This work has not been previously accepted in substance for any degree and is not being con- currently submitted in candidature for any degree.

Signed (candidate)

Date 28/09/2021

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed (candidate)

Date 28/09/2021

Statement 2

I hereby give my consent for my thesis, if accepted, to be made available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed (candidate)

Date 28/09/2021

Abstract

Machine learning has been at the core of many technological advancements in recent years, unfortunately the role of the human is sometimes overlooked. The problem that initially arises here is that if there is no trust in the model then it will not be used. A big step in resolving this is understanding of what the machine does and understanding it is behaviour [62]. The model of acceptance believes that the acceptance of technology in households needs to be understood and accepted for the same technology ideas to be widely adopted into society. In recent years there has been a huge increase in demand for cloud based services, this in turn has caused a prompt boost in the levels of Internet traffic and topological complexity. This creates several different requirements for accurate classification of applications and Internet traffic. Machine learning continues to grow and gain notoriety in socially important decision making, interpretability remains a critical dilemma especially when it comes to predictive models. If a model lacks optimality, then it could have substantial societal implications [5]. This study employed a mixed method approach. Participants were initially sent a pre-study questionnaire to garner information regarding attitudes towards technology. Then three separate neural networks were trained on the darknet dataset to classify information. Then participants contributed further by watching a presentation video and completing a second questionnaire to establish whether global or local explanation methods are preferred when understanding what each model is contributing.

Acknowledgements

First many thanks go to my supervisors Dr Matt Roach, Dr Alma Rahat and Dr Lella Nouri.

I would also like to dedicate this work to my wife Karen, and all my colleagues who have helped me in the last year.

Lastly, to the army of those who are here or passed that believed in my ability to just get on with it.

Table of Contents

Chapter 1

Introduction

1.1 Motivations

Machine learning has been at the core of many technological advancements in recent years, unfortunately the role of the human is sometimes overlooked. The problem that initially arises here is that if there is no trust in the model then it will not be used. A big step in resolving this is understanding of what the machine does and understanding it is behaviour [62]. The adoption of machine learning models in everyday tasks is growing rapidly, so is the requirement to consider the ethical, moral, and societal implications of using these machines. Several important questions are raised here such as, how did the model conclude the predictions it made? Does the prediction mean a favourable outcome for some and not others? Has it been unfavourable to a particular group? Can the model be manipulated into changing its predictions? [66] to answer at least some of these questions there is a growing need for research in this area. This paper looks at interpretability and attempts to answer where there is a preference for global or local model explanations. What are people's expectations from the adoption of new technology? There are fundamental requirements for advancements in technology to ensure society is up to date with essential facilities, but do users understand that security applications need to evolve in line with these innovations too.

A fundamental step for any network design task is a good understanding of the traffic the network is supposed to carry [27]. Artificial intelligence is one of the most prominent areas of science and technology. It can have a tremendous socio-economic impact and a ubiquitous adoption in modern society [12]. Network traffic identification provides an effective technical tool to aid in the classification of traffic to and from applications. By classifying, identifying, and distinguishing the application of network traffic, the traffic can then be sub-divided to provide users with personalised network services improving the quality of service and user satisfaction [77] and security. Artificial intelligence systems that are based on machine learning excel in many areas and some can even out-perform humans in compound tasks [64]. Building machine learning models which are transparent, is a convergent approach to extracting novel domain knowledge and performing model validation [52]. There has been a surge in the use of machine learning methods in areas such as, healthcare, law, autonomous car and policy regulation, decisions are increasingly being made by algorithms. These models are usually black box models and therefore do not provide explanations for how and why they make decisions. This becomes a problem when decisions are biased and enforce inequality. The understanding of decisions of machine learning models and the processes behind them can help understand the rules the models use and therefore attempt to prevent and train out any potential bias [55]. Deep learning models have been instrumental in solving complex problems in the areas of vision and speech. However, a key bottle neck problem with acceptance of these models in real life applications is due to the issues of interpretability and trust. Algorithms are usually trained on limited data and therefore often different to that of the real world. Subject to human error or unwarranted correlations in data create bias which can have an adverse effect on the hypothesis learned by the model [16].

1.2 Objectives

Leading machine learning models are typically very opaque but are increasingly being used for making critical decisions. As such there is a growing urgency to understand these models and ensure they are accurate, fair, and unbiased [59]. Interpretability can be defined as something which is human simulatable [43]. By this it is assumed that humans can carry out all calculation required in a reasonable time, rule out functions which are not required and provide a systematic description of all calculations, models that can be considered human simulatable are, nearest neighbours and small decision trees, among others [59]. Like teachers and students, transparency means the objective, assessment and the outcome are clear.

The objective of this study is to ascertain participants explanation method preference between global and local perspectives of machine learning models. This will be done through various methods utilizing questionnaires, video, and interview techniques. They will be presented with three different models of classifier represented by colour, green, blue, and orange. These will be representative of random forest, neural network, decision tree classifiers, this however, will be unknown to the participants to prevent preference and reduce influence. Participants will be given the accuracy score for each model and then presented with various visualisations, firstly confusion matrices which should add depth and richness to the accuracy of each model. Following on from this will be a basic decision tree, then permutations feature importance and finally ShAP feature importance. The intention of this study is to highlight where transparency and explainability meet and can be enhanced and new technology regarding internet security can be readily accepted and adopted.

Chapter 2

Literature Review

2.1 The Darknet

The Internet has vacant address space which is not speculated to interact with other computers. It is named the darknet due to its anonymous nature, virtual marketplace, and cryptocurrency. There are illegitimate hosts in the darknet, as such any traffic is treated as probe, back scatter, or misconfiguration. They are also known as network telescopes, sinkholes, or black holes. Analysis of darknet traffic can assist in identifying and monitoring of malware before the offensive and detection of malicious intentions after an outbreak [42]. However, not all activity on the darknet is nefarious, some users prefer to use it for the extra layer of privacy it affords them.

2.2 Fear of New Technology

Fear is fuelled by people being exposed to greater risk in the job market without the means to gain security. It is important to understand what this fear is as it impacts the perception and therefore the development and adoption of new technology [20] ultimately also affecting economic growth potential, opportunity, and security. Community debates reaching back as far as the industrial revolution, highlight the public's perceived threat of technological innovation to job opportunities [20]. There are social concerns regarding technological developments, the main one being the computerisation and automation of some tasks are a perceived source for job losses, this is a phenomenon known as technological unemployment [23]. Arntz et al [6] argue that there is support for contention that jobs at risk should not be equated with technological advancements. Mokyr et al [050] noted that the adoption of new technology is a much slower process than what is commonly imagined. Computerisation can create new jobs, more jobs, and the opportunity to upskill and train. Workers can adjust to computerisation by increasing their human capital in the complementary areas with the creation of new occupations and products [1] this study also highlighted that America has in the last 30 years created new occupations which have accounted for a large percentage of the employment growth. Graetz and Michaels [31] show that the introduction of industrial robotics has led to a decline in the hours worked for both low and middle skilled works. They also highlight that it is not clear who is the most fearful or why.

Concern of job security is not the only type of concern that is occurring in parallel to this there is also the anxiety of technology and immoral pervasive material that can be observed online. Society is concerned about protecting the vulnerable, specifically children and others from falling victim to this material. Back in 2018 the UK’s then Prime Minister, Theresa May stated that "Technology companies still need to go further in stepping up to their responsibilities for dealing with harmful and illegal online activity” [58]. But how do companies deal with this on such a large scale, spanning the globe and several different legal structures. Facebook is by far the largest social media company; at the end of 2018 it had employed over 30,000 content moderators worldwide [49]. But even with a workforce this large, it will never be enough to manage the millions of social media posts published every single day, also it does not consider other content distributed on the Internet via other methods. Consideration also needs to be paid to the type of content being viewed, some of which may be violent, traumatic, and disturbing, how do we keep the moderators safe. It is therefore easy to conceive why companies have started employing Artificial Intelligence (A.I.) into their security frameworks. To improve security measures companies have started collecting as much web traffic information as possible with the aim to analyse it by correlating it with the services that are provided and compare it to a log file to optimise decision making processes. This can allow safe conclusions to be drawn about network users and optimise the network resources according to monitoring needs and control legal and security issues. The information ecosystem and the importance of its applications requires a cybersecurity environment with fully automated solutions. They also need real time incident handling, analysis, and security information to identify known and unknown threats. There are recommendations for agnostic neural network framework for darknet traffic, big data analysis and network management to real-time automate the malicious intent decision process [21].

2.3 Technology Acceptance Model

Technology acceptance model, diffusion of innovation are unified theories of acceptance and the use of technology [57]. Model of acceptance believes that the acceptance of technology in households needs to be understood and accepted for the same technology ideas to be widely adopted into society. The model was designed to explain an individual's adoption of electronic mail system in organisational settings, based on work purposes with the individual employee of the organisation in this type of setting [57] however, this type of adoption is usually mandatory. The adoption of technology in the home is different, even down to desktop and mobile applications. The first questionnaire looks at what causes people to accept or reject technology. Among the many variables that can and do influence system use there are two main determinants as defined by Davis [19]; will it help me do my job better? Which makes the system useful and is it easy to use? Perceived usefulness can be defined as 'the degree to which a person believes that using a particular system would enhance his or her job performance' [19]. Useful is defined as 'capable of being used advantageously' [19]. Any system which has perceived usefulness is one which a user believes in the existence of and forms and positive performance relationship regarding its application. Perceived ease of use refers to the degree in which the user is free of any effort. Ease can be defined as 'freedom from difficulty or great effort' Davis [19]. Mobile users perceived ease of use behaviour could be influence by the ease of access, speed, screen size, text input facilities, storage and battery life span compared to desktop counterparts [57]. It is from this research that the first questionnaire has been designed.

2.4 Perception and Identifying Internet Trends

The perception here is that through machine learning techniques the A.I. learns to identify deplorable posts and imagery and either highlights to a team of reviewers or removes it entirely itself. However, the problem here is that any method developed is only as good as its programming. There is no human like organic learning and evolving, or memory. It may also need updating with reclassified data and therefore is open to faults from regression testing. Other negative aspects of using A.I. have been seen more recently with the outbreak of COVID-19 pandemic. As with many companies, Facebook sent its employees home as part of the safety measures of lockdown, giving it is A.I. system full autonomy for the first time over content moderation. Unfortunately, this saw legitimate news feeds from *The Atlantic* and *Times of Israel* amongst others being removed for violating Facebook's spam rules. The problem was attributed to a bug but, could also be an indication of what is to come as the world moves forward in a different direction. The online environment is becoming a more and more vital way to communicate and share information [49]. But how does an algorithm identify the difference between what is potentially two identical images with very different content attached. To combat these issues some social platforms have introduced a report function which provides self-governance to users in conveying social opinions of acceptable content. However, the request for removal of specified content by a user is not guaranteed under the same rules in which the A.I. and company content arbitrators hold such content to account. With less moderators and more users moving to online services for information, there is potential for those users to come to more harm from nefarious content such as violent propaganda and sexual exploitation. With such high margins of error, it is difficult to imagine how all of this is going to be resolved swiftly and could have a cumulative chilling effect on free speech and the flow of information. There is a requisite for an in depth understanding of Internet traffic, but this is challenging for Internet service providers [72].

2.5 Classifying Internet Traffic

There are three conventional methods to uncover attack trends on the Internet, these are observing activity on the darknet, checking for attacks on honey pots and accumulating warnings by utilising intrusion detection devices. Darknets are Internet Protocol address spaces which can be reached via the internet but where no host exists [2]. Most activity here is malevolent interests such as backscatter. For researchers there is a source of valuable information which can be analysed and used to learn attack trends. The problem here though is that only the first packet of each flow can be observed [2], meaning other methods are required to obtain more detailed and accurate indication of activity. Users can only access the darknet through the application of specific software such as 'the onion router' which can be accessed through the Tor browser. The Tor network provides privacy and high level anonymity which attracts dubious service agents and therefore new critical challenges to world security is created [3]. Generating the increasing necessity for new methods of detection, monitoring and elimination techniques to maintain safety for other users of the Internet as a whole.

In recent years there has been a huge increase in demand for cloud based services, this in turn has caused a prompt boost in the levels of Internet traffic and topological complexity. This creates several different requirements for accurate classification of applications and Internet traffic. One example of this would be quality of service treatment, different networks and cloud applications enforce different quality of service constraints such as, low end-to-end delay of applications which are interactive, through to high throughput for file transfer needs [36]. Network management tasks, like bandwidth budget for crucial applications rely on classification of network traffic [4]. The limitation of classification of ports led to the progression of Deep Packet Inspection (DPI). This method inspects the packet application headers alongside the payload and matches it alongside the application pattern signatures. This has been a successful technique in identifying applications carried by traffic through dynamic ports. Unfortunately, there are shortcomings related to several issues such as scale, cost and reliability and reliance of the availability of packet payload. New approaches have been proposed to augment both DPI and port-based classification, to assist in identifying applications through analysis of statistical properties of traffic and the characteristics of applications based on flow-feature and host behaviour. This is where successful machine learning tools have been applied to classify network traffic. Patterns here fluctuate during busy hours, by building a new network classification model with new data and applying it to historical training datasets could help boost real time accuracy [36]. Analysis monitoring and categorisation of Internet network traffic is an extremely important security task. Information that is exchanged may be requests, responses and control data that is fragmented in the form of network packets. If it is difficult to make any conclusion when looking at individual network packets this can be because the transmission of information is fragmented between devices, this can be into several packets which are interconnected [21].

As stated above, one way to categorize network traffic is to look at the payload of every packet. This can be extremely accurate as long and the payload is not encrypted. However, there are many applications that employ some form of encryption making classification accuracy opaque. Another approach would be to use well-known TCP/UDP port numbers, however, again this can become difficult when applications use non-traditional ports to bypass firewalls or circumvent operating system restrictions [4]. One way would be to identify features of network traffic and then used those identified to guide further classification. Research in this area has used machine learning practices such as Naïve Bayesian Models, Adaboost or Maximum Entropy methods. Although, having an encoded payload and running different applications engaging an encrypted channel makes it a difficult problem to classify encrypted traffic from a given file log [54], [53]. Alshammari and Zincir-Heywood [4] used machine learning methods for regression in their project work, the accurate type of classifier identified was Support Vector Machines (SVM). They also recognized that future work should follow a similar route by comparing the classification approach against clustering based techniques.

A neural network is a processing system consisting of an input layer and output layer, with one or more hidden layers which are connected by neurons. The neurons are a processing element which receives the input and produces an output signal through a transfer function [74]. Neurons receive an input from the neuron activations of the previous layer and perform a computation. The neurons of the network implement a complex non-linear mapping from the input to the output, the mapping is learned from a practice called error backpropagation [52]. There are many types of neural networks that are used for different purposes an example being, Support Vector Machines (SVM), these are used for binary classifications, such as facial recognition, classification, and text categorisation [67]. There are also Convolution Neural Networks (CNNs) which contain a 3-dimensional arrangement of neurons. Inputs are multiplied by weights and then fed into the activation function. Convolution uses RELU and multilayer perceptron uses nonlinear function and softmax. This type of neural network is particularly effective for image and video recognition. They can be used for deep learning with minimal parameters but are complex to design and maintain [30].

The prerequisite of building an SVM classification model is that each training sample can determine parameter for each class. The decision weight for each class is set by the parameter vector of all training samples contributing to the decision weight of the corresponding class and less emphasis on the decision weight of other classes. Statistical property of SVM decision making requires a large training set, therefore a cost effective way of achieving this is selecting a training sample and optimising its parameters, while keeping all other parameter vectors fixed. To avoid parameter optimisation bias, the order of different samples needs to perform indefinite iterations of the parameter vector of each sample until global optimality is reached. The application of iterative tuning to the SVM also achieves better classification accuracy and speed of training in comparison to non-iterative SVM models [36]. Whilst SVM techniques seem to be a future model for classification there are still problems regarding the speed of model training even with a small data sample size the model can be lethargic in its training abilities.

Machine learning methods such as Deep Neural Networks (DNNs) are an indispensable too for a comprehensive range of applications such as image recognition, speech, and natural language processing. Montavon et al [52] states that techniques for interpretation and understanding have become a key part to a robust validation procedure. Simple models are easier to understand, and therefore they also support the fact that linear models and basic decision trees tend to be the dominate choice in this area. However, here is an apparent lack of a formal understanding of what it means to explain a classifier, many of the insightful visualisations produced are heuristic and therefore their meaning can remain unclear [26]. An ideal explanation would extract the whole process chain from input to output, unfortunately, this technology is not available yet [40]. What is clear is that as these models become increasingly predictive and adopted into industry, it also becomes progressively necessary that they work as intended and their decisions are as transparent as possible [7]. Deep learning models are theory agnostic this means that designers do not programme them into a model that reflects their understanding of the causal structure of the problem that requires solving but the structure learns a model from a large set of data [44]. The architecture of these models contains layers of connected nodes much like the neurons in the human brain [44].

Machine learning models based on trees are the most popular non-linear models currently in use, Random Forest gradient boosted tree and others are used in finance, medicine, biology, customer retention and other areas. They make predictions based on input features, it is important that these models are accurate and interpretable. There are three main ways to explain individual predictions from trees, Lundberg et al [45] identify these as; first, reporting the decision path, second an unpublished heuristic approach that assigns credit to each input feature and third, a variety of model-agnostic approaches that require executing the model many times for each explanation. Each of these approaches have constraints, simply reporting the decision path is not helpful. The behaviour of heuristic methods has not been meticulously analysed to date. Model-agnostic approaches can be slow and suffer from sampling variability. Due to the criticality of some of the decision being made there is a growing urgency to understand the decisions being made to ensure that they are correct, fair, unbiased, and ethical. Local explanations are typically derived from the model directly, or a model that approximates the predictive model in the neighbourhood around a specific point. They indicate the situation in which it is possible to understand only the reasons for a specific decision, meaning that only the single prediction/decision is interpretable. Global explanations can take the form of a series of roles. They can understand and follow the whole logic of the model and therefore the entire reasoning, leading to all different outcomes [59], [32].

The impressive performance of A.I. approaches to prediction, recommendation and decision making tends to come from the adoption of complex machine learning models that do not present the logic of the internal processes [12], these machines are known as black boxes. Machine learning black box models are created from data using an algorithm in a way that we cannot understand. Variables are created and combined to make predictions even if a list of input variables are available to be utilised. Trusting a black box model means confiding in the models' calculations and the whole dataset from which it has been built, can trust be formed without understanding what is being presented? There are other alternative technically equivalent more interpretable models available, which also provide a better understanding of the predictions being made [63] but may not offer the accuracy or swiftness required. Dempster-Shafer evidence theory [68] attempts to offer an effective tool for uncertainty reasoning without access to prior information. Results generated from this theory are fault accepting which can assist in supporting any decision making [71]. Dempster-Shafer method fulfils associatory laws and has therefore been comprehensively applied the field. Though, Xiao [76], found that none of these models have the capability to express the fluctuations of data at a given phase of time during their execution [76].

There has been a proliferation of new applications that do not have IANA registered ports. Registered ports are assigned in several different ways on a first come, first serve basis, and distinguish between different services that run over transport protocols such as TCP, UDP, DCCP, and SCTP. System and user ports should not be used without or prior to IANA registration. However, registration does not guarantee authenticity of an application and any traffic flowing through or from a registered port should not be considered genuine or respectable [71]. The incentive for users to take advantage of already registered ports is to disguise any traffic and can circumvent any filtering or firewalls [18]. This also has the capability of motivating pervasive deployment of network and port address translation, where several servers offer services through the same public IP addresses but via different ports. Dainotti et al [18] also state that payload examination techniques are a reliable source of Internet traffic classification, but it faces formidable privacy challenges. Privacy policies and laws may prevent researchers access to or archiving any packet content. There are also technological and economical challenges to this method such as it is easily bypassed by encryption methods, protocol obfuscation or encapsulation. It is also prohibitively computationally expensive for general use. Nevertheless, these interests have driven new discriminating properties of internet traffic classes and other classification that no longer requires payload examination. The creation of algorithms from pattern recognition using machine learning techniques seem to show promising results especially in the use of classification of encrypted traffic and supervised machine learning tactics have proven to achieve comparable results to payload examination. Unsupervised machine learning is showing promising results in dealing with evolving network traffic. These are the areas that require renewed attention as applications using encryption technology is proving to be increasingly chosen areas for nefarious activity as security and removal techniques are employed by the larger and popular media application companies.

Machine learning continues to grow and gain notoriety in socially important decision making, interpretability remains a critical dilemma especially when it comes to predictive models. If a model lacks optimality, then it could have substantial societal implications [5]. For instance, COMPAS users surreptitiously assumed that a transparent model would not be accurate enough to produce effective recidivism predictions. However, COMPAS scores are racially biased and because the decision was made to not use a transparent model, no-one can determine the root of the bias or its extent. Another fault with this system is that it does not provide any reasoning for a given prediction [41]. Another problem to note here would be the fact that interpretability is open to perception and therefore has many different meanings. Montavon et al [52] provide definitions for interpretability and expainability. Definition one; an interpretation is the mapping of an abstract concept into a domain that the human can comprehend. Definition two; an explanation is a collection of features of the interpretable domain that have contributed to a given decision. Black box models have such a high performance and accuracy history that it has helped encourage the adoption of non-interpretable machine learning models even if the denseness inherent issues from training or unfair data, as seen with COMPAS. A substantial risk is presented by relying on opaque model may lead to implementing decisions that lack understanding or violate ethical principles. Companies and individuals, which employ these methods, increasingly embedding machine learning models into A.I. products are incurring all kinds of risk, potential loss of safety and trust [12].

The technology presented in some of the machine learning systems is revolutionary in its nature but, despite this they face challenges to their deployment. The large complexity and high energy demand of deep learning models are example of this. There is also a shortage of hardiness to antagonistic attacks that can cause major complications and security risks in application. An example of this would be autonomous driving. There are issues around transparency and explainable techniques reducing any trust in the verifiability of any decisions made by the A.I. system [64]. In part the ability to verify the decision making of an A.I. system does help foster trust in that system. The capacity to explain the rationale behind any decisions is important for human interaction, explanations are an essential part of human education and learning and help reinforce trust. Straightforward classifiers such as linear models or shallow decision trees are intrinsically interpretable, complex classifiers such as deep neural networks that contain several layers of non-linear transformations complicates the understanding of how they make predictions. A method to overcome some of these challenges is to locally approximate them with a surrogate function which is interpretable [64]. One popular method of this is Shapely Additive Explanations (ShAP), ShAP is considered one of the better techniques for explainability due to its reliability and consistency ensuring that features that are most important features are always given the highest score. Whereas, for instance a tree-based model may give two equally important features separate scores based on the level of splitting that was done using the features [61]. This is the main rationale for using ShAP as an explanation method.

One cannot appeal a decision without knowing the basis upon which it has been made. The transparency and explainability of the decision are a crucial pre-requisite of democratic governance but the level of understanding is not the same for everyone involved, for instance judge, solicitor, and defendant, it is different for each. So, what is the minimum requirement? The assumption seems to be that it is fair to impose a higher standard of transparency on A.I. systems than human decision makers [81]. Problems such as racial bias and vulnerabilities to nefariously motivated attacks have led to calls for models to be interpretable. This will not only lead to identifiable bias but help build trustworthiness in the system [25]. The issues presented around explainability have become problematic with the prominent view that accuracy of an A.I system is often a trade-off for explainability [25]. Mittelstadt et al [50] ask are explanations important on an individual level? And if so, who is affected by the decision making process? Transparency addresses the internal function of a given model and can therefore be further specified to its target. Post-hoc human interpretable explanations of a model and specific decisions do not seek to answer how a model functions but how it behaved and why [50]. What is more important accuracy or understanding? They are two very different components. Should simplicity be favoured over accuracy? What would increase interpretability? Do any models stand out in particular? Or is it the explanation method? Do people have a preference to global or local explanation techniques This study seeks to answer these questions to some extent. What does seem clear is that the problem plaguing current literature is that explainability and understanding are synonymous with interpretability.

Explainability is essential for users to effectively understand, trust and manage powerful A.I applications [33].

Chapter 3

Methodology

3.1 Approach

This study employed a mixed method approach. Participants were initially sent a pre-study questionnaire to garner information regarding attitudes towards technology. Then three separate neural networks were trained on the darknet dataset to classify information. Then participants contributed further by watching a presentation video and completing a second questionnaire to establish whether global or local explanation methods are preferred when understanding what each model is contributing.
Prior to any live study the researcher will conduct a pilot study with a couple of individual participants to ensure that the final study will run smoothly. Each participant at this stage will complete the questionnaires and watch the video as intended but also provide quality feedback to the researcher. This will enable fine tuning of the final questionnaires and the video to facilitate better functionality and understanding to all participants regardless of their background and knowledge of the technology and principles being discussed. The feedback obtained at this stage will prove to be vital to the researcher as it may present novel thoughts and opinions circumvented at this initial phase of the study. Pilot studies are important as it will also provide a feasibility response and time constraints regarding the structure of the research being conducted [46] this is important as it will allow the researcher to provide accurate length of the study to participants at the time of inviting them to take part.

This study comprises of a classification of a darknet dataset, followed by a live video and questionnaire event with participants. The questionnaire will consist of two parts. The initial part will be emailed with the consent form to be completed prior to the meeting. This is to obtain what technological devices participants own and their thoughts on technology advancements in work and the home. The thinking behind this part is that there may be a pattern between opinion of technology devices and how participants perceive advancements such as network classifiers. The second questionnaire has been designed in hand with a video presentation, the intention is to show the presentation to participants which has been broken into three sections, at each section the video will be paused, and the participant will be asked to complete the relevant section of the questionnaire. The researcher will also have at hand the individual presentation slides so that if a participant wishes to recap anything then it can be done swiftly and easily. Participants will also be made aware that the video can be paused at any time if any explanation is required or if they have any questions.

3.2 Pilot Study Results

Doody and Doody [24] state that conducting a pilot study can lead to higher quality and relevant research and is intentionally planned before the actual study. This initial study was put together and presented to ensure that the final process flowed efficiently and was as concise as possible. It also presented an opportunity for the researcher to receive feedback to make certain that the video and questionnaires were appropriate for the study and aimed to answer the research question. The researcher believes that it is important that the information is presented in an understandable format [48] so that participants with no background knowledge do not feel swamped with technological language. A pilot study is like a feasibility study and is used figure out the best methods for pursuing it and estimate the time and resources will be required to complete the larger version, among other things [17], which is important when asking people to give up their time to take part. Feedback from this study suggested adding a summary slide to the end of each section of the video to enable participants the ability to review what had been covered. The questionnaire needed an overhaul to ensure it flowed better with the video presentation. Also, it became clear that presenting via Zoom made it difficult for a two way conversation between the researcher and participant. As such it was suggested that the questionnaire be expanded slightly to include some more open-ended questions to collect useful response data that would have been collected from conversation. All this feedback was valuable and implemented prior to the undertaking of the formal study.

3.3 Final Study Design

The presentation is a PowerPoint video which has been put together to show to participants. It outlines the accuracy of the three models that have been designed to classify the Darknet Dataset. The Darknet Dataset which will be utilised consists of 158,659 records. This has been broken down into 134,348 benevolent samples and 24,311 darknet examples. In terms of traffic the highest is audio-streaming with 13,284 samples [42]. The three models present to the participants are the orange model, blue model, and the green model. They have been presented this way to prevent any preference bias should the participants have any prior knowledge of classifiers. One presented with the accuracy the video moves one to show performance in a form of confusion matrix. Next, they are provided with the fact that the blue model can produce decision trees. Performance is then offered in the form of global and local explanation methods. The methods chosen by the researcher are permutation scores and ShAP (Shapely Additive Explanations). The purpose of picking these two methods is to be able to present participants with two very different techniques to ascertain which is a preferable explanation approach. Fig. 1 show the list of features and the first line of data of the dataset to highlight the type of data used:

Fig. 1 – Extract of data from the darknet dataset

Src Port	57158
Dst Port	443
Protocol	6
Flow Duration	229
Total Fwd Packet	1
Total Bwd packets	1
Total Length of Fwd Packet	0
Total Length of Bwd Packet	0
Fwd Packet Length Max	0
Fwd Packet Length Min	0
Fwd Packet Length Mean	0
Fwd Packet Length Std	0
Bwd Packet Length Max	0

Bwd Packet Length Min	0
Bwd Packet Length Mean	0
Bwd Packet Length Std	0
Flow Bytes/s	0
Flow Packets/s	8733.624
Flow IAT Mean	229
Flow IAT Std	0
Flow IAT Max	229
Flow IAT Min	229
Fwd IAT Total	0
Fwd IAT Mean	0
Fwd IAT Std	0
Fwd IAT Max	0
Fwd IAT Min	0
Bwd IAT Total	0
Bwd IAT Mean	0
Bwd IAT Std	0
Bwd IAT Max	0
Bwd IAT Min	0
Fwd PSH Flags	0
Bwd PSH Flags	0
Fwd URG Flags	0
Bwd URG Flags	0
Fwd Header Length	20
Bwd Header Length	20
Fwd Packets/s	4366.812
Bwd Packets/s	4366.812227
Packet Length Min	0
Packet Length Max	0
Packet Length Mean	0
Packet Length Std	0
Packet Length Variance	0
FIN Flag Count	2
SYN Flag Count	0
RST Flag Count	0
PSH Flag Count	0
ACK Flag Count URG Flag Count	2
CWE Flag Count	0
ECE Flag Count	0
Down/Up Ratio	1
Average Packet Size	0
Fwd Segment Size Avg	0
Bwd Segment Size Avg	0
Fwd Bytes/Bulk Avg	0
Fwd Packet/Bulk Avg	0
Fwd Bulk Rate Avg	0

Bwd Bytes/Bulk Avg	0
Bwd Packet/Bulk Avg	0
Bwd Bulk Rate Avg	0
Subflow Fwd Packets	0
Subflow Fwd Bytes	0
Subflow Bwd Packets	0
Subflow Bwd Bytes	0
FWD Init Win Bytes	1892
Bwd Init Win Bytes	1047
Fwd Act Data Pkts	0
Fwd Seg Size Min	20
Active Mean	0
Active Std	0
Active Max	0
Active Min	0
Idle Mean	0
Idle Std	0
Idle Max	0
Idle Min	0
Labels 1	Non-Tor
Labels 2	Audio-Streaming

As demonstrated from the above table most of the data is numerical with two columns that are labels. Labels 1 identifies whether traffic is darknet or not and is labelled as, Tor, VPN, Non-Tor, or Non-VPN. Labels 2 recognises the type of category of traffic and contains the following identifiers, Audio-Stream, Browsing, Chat, Email, P2P, File Transfer, Video-Stream and VOIP. Pre-processing of the data removed, table entries that had no data or entries such as N/A, it was also necessary to rename the label columns Labels 1 and Labels 2 to obtain a distinction between the two as they were run through the classifiers separately. Prior to classification the data was split into test and train sections, this split was set at an 80%/20% divide and was maintained at this level for all sections. The point of separating the data this way is to prevent overfitting [69].

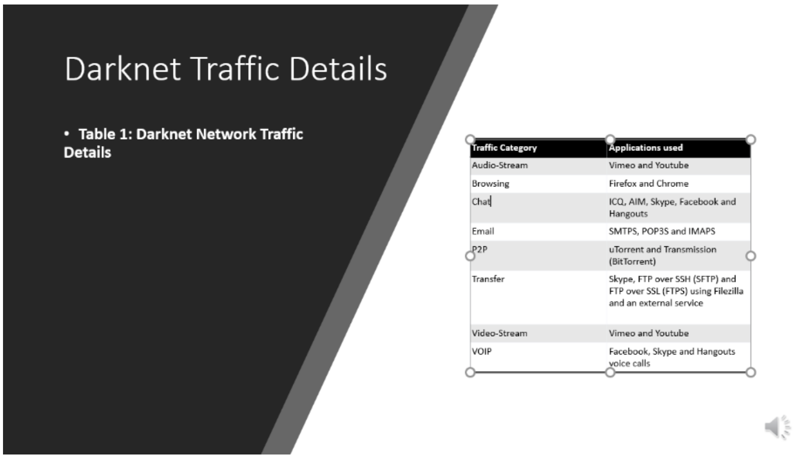
The video presentation will be split into three sections, at the end of each section, participants are asked to fill in the relevant section of the questionnaire. At the beginning of the video the participants are informed that they can pause the video at any point and ask questions or seek further explanation should it be necessary and that the full PowerPoint is available should they wish to revisit a section. The design of the study is illustrated in fig. 2

Fig. 2 – Study Design

Introduction	Orange Model	Blue Model	Green Model	End of section
Accuracy scores including confusion matrix and decision trees	Model accuracy	Model accuracy	Model accuracy	Discussion with participant including relevant section of questionnaire.
Global explanation	Visualisation	Visualisation	Visualisation	Discussion with participant including relevant section of questionnaire.
Local explanation	Visualisation	Visualisation	Visualisation	Discussion with participant including relevant section of questionnaire.

Fig. 3 below shows the slides as presented in the video; this will give an idea of what the participants observed. Each slide had contained brief but relevant explanations as to what was being proposed:

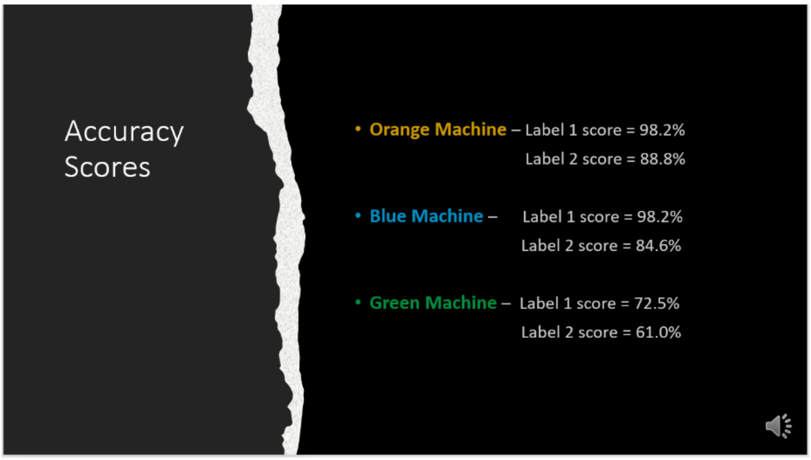
Fig.3 Presentation Slides



Darknet Traffic Details

- **Table 1: Darknet Network Traffic Details**

Traffic Category	Applications used
Audio-Stream	Vimeo and Youtube
Browsing	Firefox and Chrome
Chat	ICQ, AIM, Skype, Facebook and Hangouts
Email	SMTPS, POP3S and IMAPS
P2P	uTorrent and Transmission (BitTorrent)
Transfer	Skype, FTP over SSH (SFTP) and FTP over SSL (FTPS) using Filezilla and an external service
Video-Stream	Vimeo and Youtube
VOIP	Facebook, Skype and Hangouts voice calls

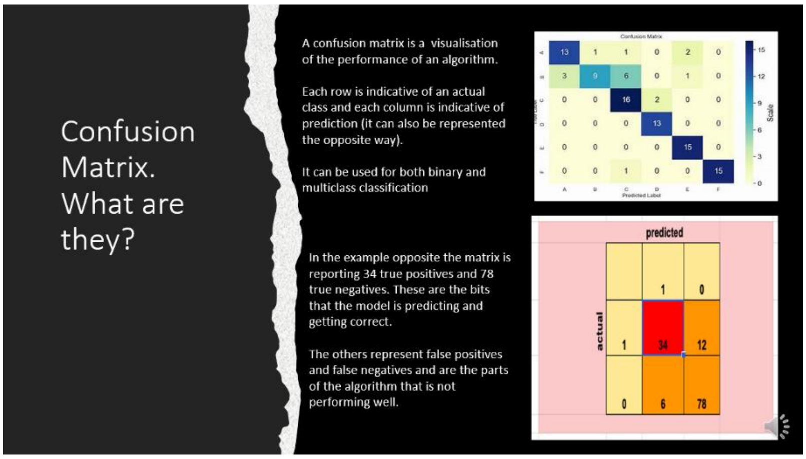


Accuracy Scores

- Orange Machine – Label 1 score = 98.2%
  Label 2 score = 88.8%

- Blue Machine – Label 1 score = 98.2%
  Label 2 score = 84.6%

- Green Machine – Label 1 score = 72.5%
  Label 2 score = 61.0%



Confusion Matrix. What are they?

A confusion matrix is a visualisation of the performance of an algorithm.

Each row is indicative of an actual class and each column is indicative of prediction (it can also be represented the opposite way).

It can be used for both binary and multiclass classification

In the example opposite the matrix is reporting 34 true positives and 78 true negatives. These are the bits that the model is predicting and getting correct.

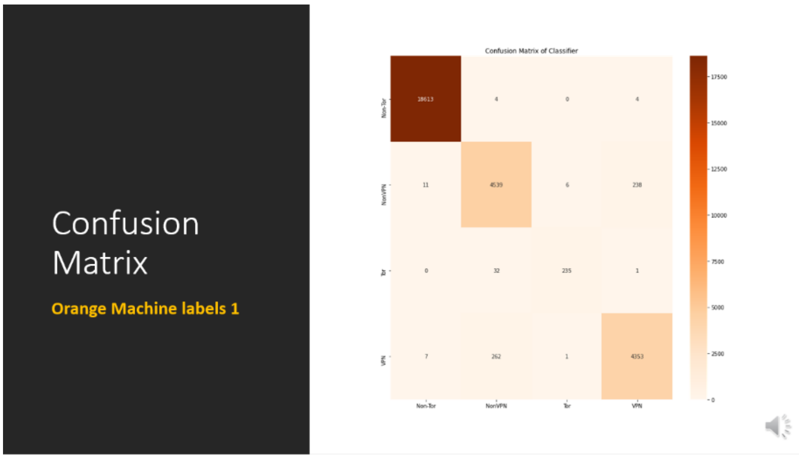
The others represent false positives and false negatives and are the parts of the algorithm that is not performing well.

Confusion Matrix

	A	B	C	D	E	F
A	13	1	1	0	2	0
B	3	9	6	0	1	0
C	0	0	16	2	0	0
D	0	0	0	13	0	0
E	0	0	0	0	15	0
F	0	0	1	0	0	15

predicted

actual		
	1	0
1	34	12
0	6	78

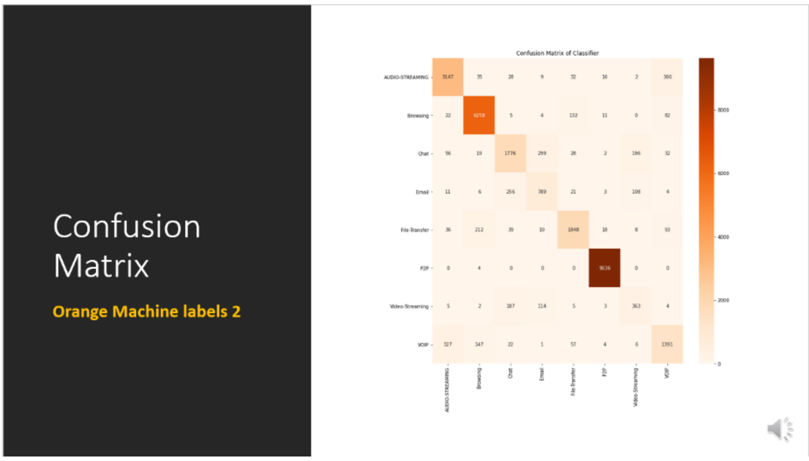


Confusion Matrix

Orange Machine labels 1

Confusion Matrix of Classifier

	Non-Tor	NonVPN	Tor	VPN
Non-Tor	18613	4	0	4
NonVPN	11	4539	6	238
Tor	0	32	235	1
VPN	7	262	1	4353

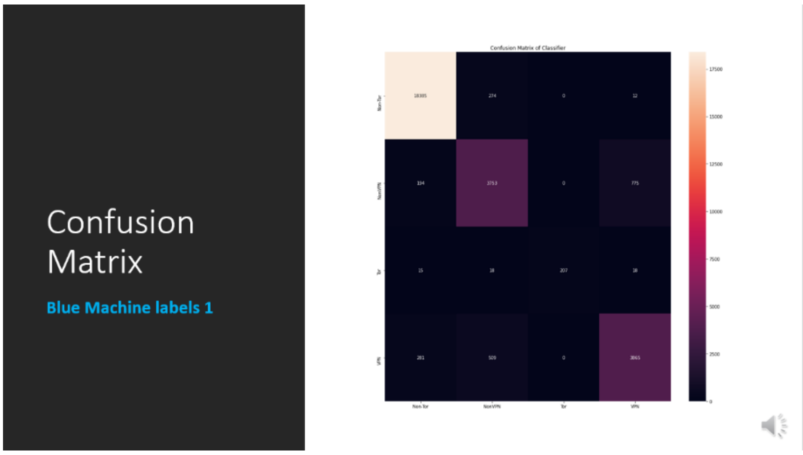


Confusion Matrix

Orange Machine labels 2

Confusion Matrix of Classifier

	AUDIO-STREAMING	Browsing	Chat	Email	File-Transfer	P2P	Video-Streaming	VOIP
AUDIO-STREAMING	3147	35	28	9	32	16	2	366
Browsing	22	6258	5	4	132	11	0	82
Chat	56	19	1776	299	28	2	196	32
Email	11	6	256	789	21	3	108	4
File-Transfer	36	212	39	19	1848	18	8	93
P2P	0	4	0	0	0	9636	0	0
Video-Streaming	5	2	187	114	5	3	363	4
VOIP	327	147	22	1	57	4	6	1391

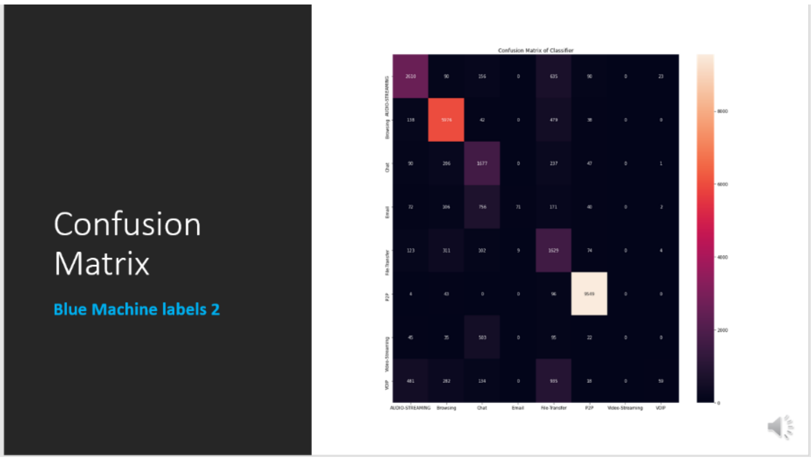


Confusion Matrix

Blue Machine labels 1

Confusion Matrix of Classifier

	Non-Tor	NonVPN	Tor	VPN
Non-Tor	18385	274	0	12
NonVPN	194	3753	0	775
Tor	15	18	207	18
VPN	281	509	0	3865

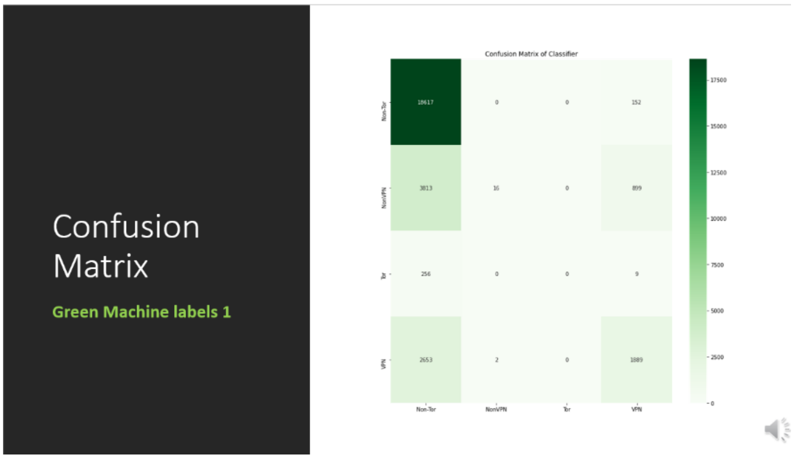


Confusion Matrix

Blue Machine labels 2

Confusion Matrix of Classifier

	AUDIO-STREAMING	Browsing	Chat	Email	File-Transfer	P2P	Video-Streaming	VOIP
AUDIO-STREAMING	2610	90	156	0	635	90	0	23
Browsing	138	5976	42	0	479	38	0	0
Chat	90	206	1677	0	237	47	0	1
Email	72	106	756	71	171	40	0	2
File-Transfer	123	311	102	9	1629	74	0	4
P2P	4	43	0	0	96	9549	0	0
Video-Streaming	45	35	503	0	95	22	0	0
VOIP	481	282	134	0	935	18	0	59

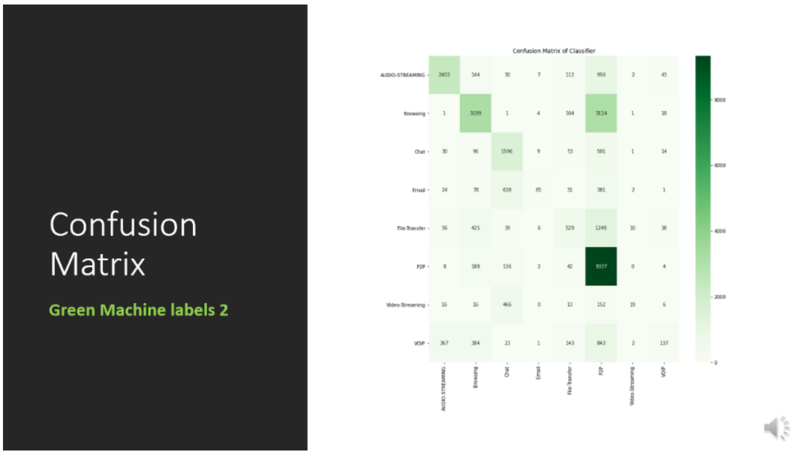


Confusion Matrix

Green Machine labels 1

Confusion Matrix of Classifier

	Non-Tor	NonVPN	Tor	VPN
Non-Tor	18617	0	0	152
NonVPN	3813	16	0	899
Tor	256	0	0	9
VPN	2653	2	0	1889

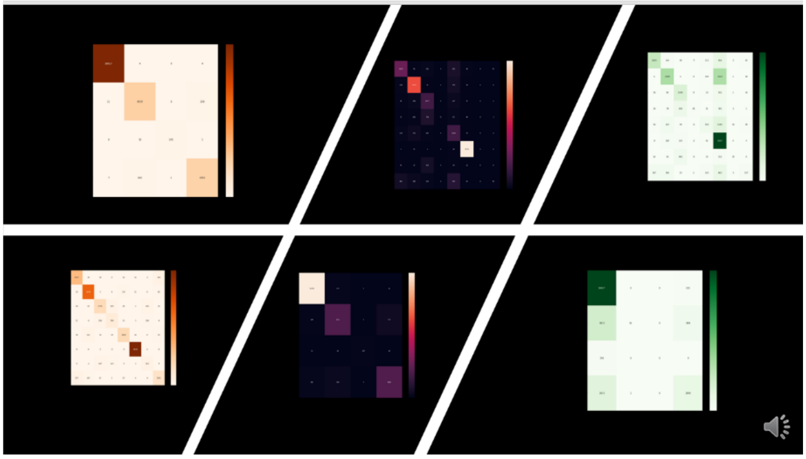


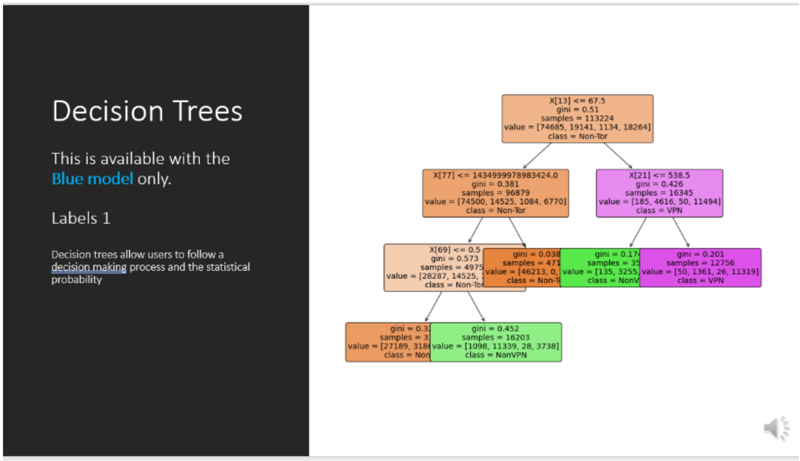
Confusion Matrix

Green Machine labels 2

Confusion Matrix of Classifier

	AUDIO-STREAMING	Browsing	Chat	Email	File-Transfer	P2P	Video-Streaming	VOIP
AUDIO-STREAMING	2403	144	30	7	112	956	2	43
Browsing	1	3189	1	4	104	3114	1	18
Chat	30	96	1596	9	53	501	1	14
Email	24	78	638	65	31	381	2	1
File-Transfer	56	425	39	6	529	1249	10	38
P2P	8	188	136	2	42	9337	0	4
Video-Streaming	16	16	466	0	13	152	19	6
VOIP	367	384	23	1	143	843	2	137





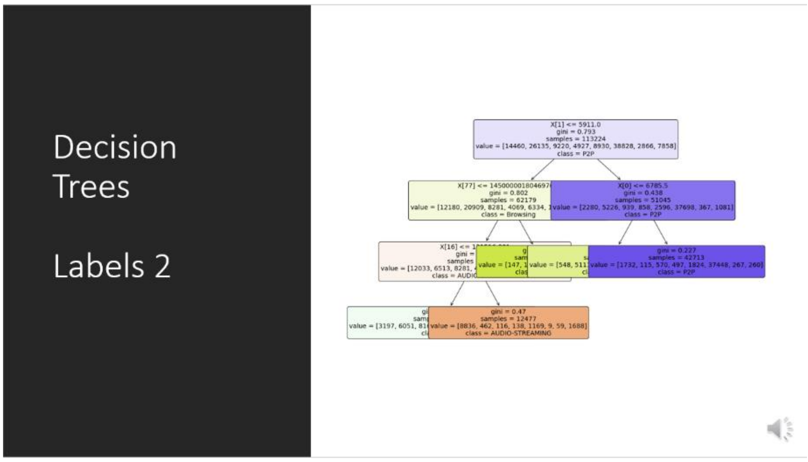
Decision Trees

This is available with the Blue model only.

Labels 1

Decision trees allow users to follow a decision making process and the statistical probability

The decision tree visualisations were illustrating only one feature, as producing against each feature would have been too big and complex for a PowerPoint presentation and the researcher wanted it to be as clear as possible. This was explained to each participant.



Decision Trees

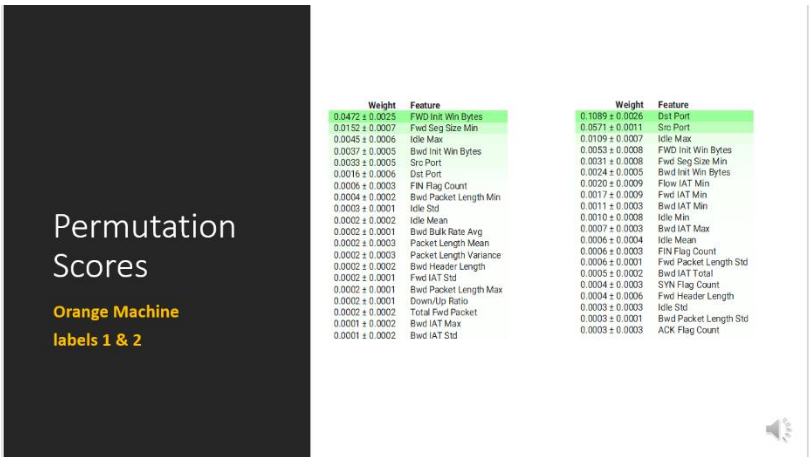
Labels 2

Visualisations

Does the addition of the decision tree assist in understanding the model?

End of Section 1

- Would the information presented so far be enough to instil trust in the machines?
- Please complete section one of the questionnaire provided.

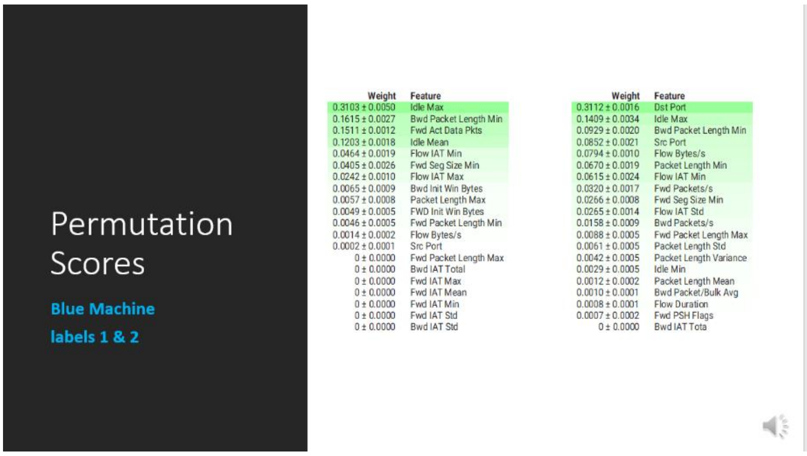


Permutation Scores

Orange Machine labels 1 & 2

Weight	Feature
0.0472 ± 0.0025	FWD Init Win Bytes
0.0152 ± 0.0007	Fwd Seg Size Min
0.0045 ± 0.0006	Idle Max
0.0037 ± 0.0005	Bwd Init Win Bytes
0.0033 ± 0.0005	Src Port
0.0016 ± 0.0006	Dst Port
0.0006 ± 0.0003	FIN Flag Count
0.0004 ± 0.0002	Bwd Packet Length Min
0.0003 ± 0.0001	Idle Std
0.0002 ± 0.0002	Idle Mean
0.0002 ± 0.0001	Bwd Bulk Rate Avg
0.0002 ± 0.0003	Packet Length Mean
0.0002 ± 0.0003	Packet Length Variance
0.0002 ± 0.0002	Bwd Header Length
0.0002 ± 0.0001	Fwd IAT Std
0.0002 ± 0.0001	Bwd Packet Length Max
0.0002 ± 0.0001	Down/Up Ratio
0.0002 ± 0.0002	Total Fwd Packet
0.0001 ± 0.0002	Bwd IAT Max
0.0001 ± 0.0002	Bwd IAT Std

Weight	Feature
0.1089 ± 0.0026	Dst Port
0.0571 ± 0.0011	Src Port
0.0109 ± 0.0007	Idle Max
0.0053 ± 0.0008	FWD Init Win Bytes
0.0031 ± 0.0008	Fwd Seg Size Min
0.0024 ± 0.0005	Bwd Init Win Bytes
0.0020 ± 0.0009	Flow IAT Min
0.0017 ± 0.0009	Fwd IAT Min
0.0011 ± 0.0003	Bwd IAT Min
0.0010 ± 0.0008	Idle Min
0.0007 ± 0.0003	Bwd IAT Max
0.0006 ± 0.0004	Idle Mean
0.0006 ± 0.0003	FIN Flag Count
0.0006 ± 0.0001	Fwd Packet Length Std
0.0005 ± 0.0002	Bwd IAT Total
0.0004 ± 0.0003	SYN Flag Count
0.0004 ± 0.0006	Fwd Header Length
0.0003 ± 0.0003	Idle Std
0.0003 ± 0.0001	Bwd Packet Length Std
0.0003 ± 0.0003	ACK Flag Count



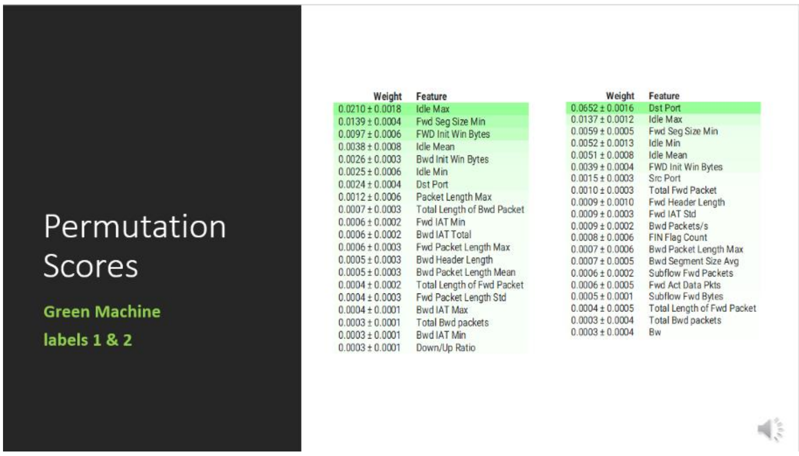
Permutation Scores

Blue Machine

labels 1 & 2

Weight	Feature
0.3103 ± 0.0050	Idle Max
0.1615 ± 0.0027	Bwd Packet Length Min
0.1511 ± 0.0012	Fwd Act Data Pkts
0.1203 ± 0.0018	Idle Mean
0.0464 ± 0.0019	Flow IAT Min
0.0405 ± 0.0026	Fwd Seg Size Min
0.0242 ± 0.0010	Flow IAT Max
0.0065 ± 0.0009	Bwd Init Win Bytes
0.0057 ± 0.0008	Packet Length Max
0.0049 ± 0.0005	FWD Init Win Bytes
0.0046 ± 0.0005	Fwd Packet Length Min
0.0014 ± 0.0002	Flow Bytes/s
0.0002 ± 0.0001	Src Port
0 ± 0.0000	Fwd Packet Length Max
0 ± 0.0000	Bwd IAT Total
0 ± 0.0000	Fwd IAT Max
0 ± 0.0000	Fwd IAT Mean
0 ± 0.0000	Fwd IAT Min
0 ± 0.0000	Fwd IAT Std
0 ± 0.0000	Bwd IAT Std

Weight	Feature
0.3112 ± 0.0016	Dst Port
0.1409 ± 0.0034	Idle Max
0.0929 ± 0.0020	Bwd Packet Length Min
0.0852 ± 0.0021	Src Port
0.0794 ± 0.0010	Flow Bytes/s
0.0670 ± 0.0019	Packet Length Min
0.0615 ± 0.0024	Flow IAT Min
0.0320 ± 0.0017	Fwd Packets/s
0.0266 ± 0.0008	Fwd Seg Size Min
0.0265 ± 0.0014	Flow IAT Std
0.0158 ± 0.0009	Bwd Packets/s
0.0088 ± 0.0005	Fwd Packet Length Max
0.0061 ± 0.0005	Packet Length Std
0.0042 ± 0.0005	Packet Length Variance
0.0029 ± 0.0005	Idle Min
0.0012 ± 0.0002	Packet Length Mean
0.0010 ± 0.0001	Bwd Packet/Bulk Avg
0.0008 ± 0.0001	Flow Duration
0.0007 ± 0.0002	Fwd PSH Flags
0 ± 0.0000	Bwd IAT Tota



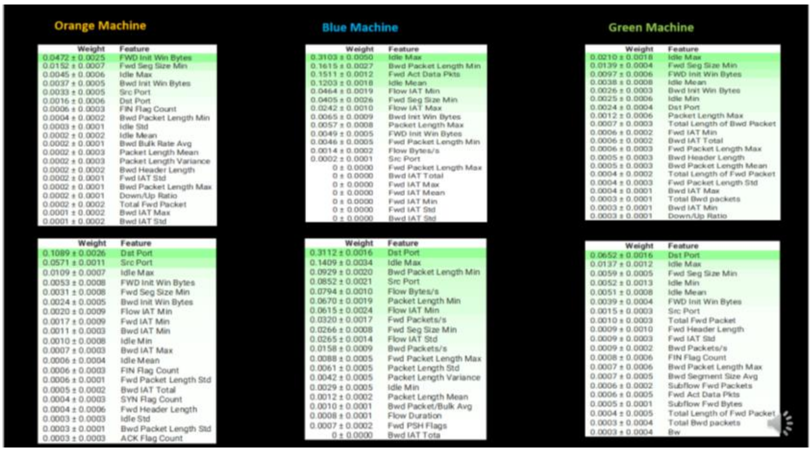
Permutation Scores

Green Machine

labels 1 & 2

Weight	Feature
0.0210 ± 0.0018	Idle Max
0.0139 ± 0.0004	Fwd Seg Size Min
0.0097 ± 0.0006	FWD Init Win Bytes
0.0038 ± 0.0008	Idle Mean
0.0026 ± 0.0003	Bwd Init Win Bytes
0.0025 ± 0.0006	Idle Min
0.0024 ± 0.0004	Dst Port
0.0012 ± 0.0006	Packet Length Max
0.0007 ± 0.0003	Total Length of Bwd Packet
0.0006 ± 0.0002	Fwd IAT Min
0.0006 ± 0.0002	Bwd IAT Total
0.0006 ± 0.0003	Fwd Packet Length Max
0.0005 ± 0.0003	Bwd Header Length
0.0005 ± 0.0003	Bwd Packet Length Mean
0.0004 ± 0.0002	Total Length of Fwd Packet
0.0004 ± 0.0003	Fwd Packet Length Std
0.0004 ± 0.0001	Bwd IAT Max
0.0003 ± 0.0001	Total Bwd packets
0.0003 ± 0.0001	Bwd IAT Min
0.0003 ± 0.0001	Down/Up Ratio

Weight	Feature
0.0652 ± 0.0016	Dst Port
0.0137 ± 0.0012	Idle Max
0.0059 ± 0.0005	Fwd Seg Size Min
0.0052 ± 0.0013	Idle Min
0.0051 ± 0.0008	Idle Mean
0.0039 ± 0.0004	FWD Init Win Bytes
0.0015 ± 0.0003	Src Port
0.0010 ± 0.0003	Total Fwd Packet
0.0009 ± 0.0010	Fwd Header Length
0.0009 ± 0.0003	Fwd IAT Std
0.0009 ± 0.0002	Bwd Packets/s
0.0008 ± 0.0006	FIN Flag Count
0.0007 ± 0.0006	Bwd Packet Length Max
0.0007 ± 0.0005	Bwd Segment Size Avg
0.0006 ± 0.0002	Subflow Fwd Packets
0.0006 ± 0.0005	Fwd Act Data Pkts
0.0005 ± 0.0001	Subflow Fwd Bytes
0.0004 ± 0.0005	Total Length of Fwd Packet
0.0003 ± 0.0004	Total Bwd packets
0.0003 ± 0.0004	Bw



Orange Machine

Weight	Feature
0.0472 ± 0.0025	FWD Init Win Bytes
0.0152 ± 0.0007	Fwd Seg Size Min
0.0045 ± 0.0006	Idle Max
0.0037 ± 0.0005	Bwd Init Win Bytes
0.0033 ± 0.0005	Src Port
0.0016 ± 0.0006	Dst Port
0.0006 ± 0.0003	FIN Flag Count
0.0004 ± 0.0002	Bwd Packet Length Min
0.0003 ± 0.0001	Idle Std
0.0002 ± 0.0002	Idle Mean
0.0002 ± 0.0001	Bwd Bulk Rate Avg
0.0002 ± 0.0003	Packet Length Mean
0.0002 ± 0.0003	Packet Length Variance
0.0002 ± 0.0002	Bwd Header Length
0.0002 ± 0.0001	Fwd IAT Std
0.0002 ± 0.0001	Bwd Packet Length Max
0.0002 ± 0.0001	Down/Up Ratio
0.0002 ± 0.0002	Total Fwd Packet
0.0001 ± 0.0002	Bwd IAT Max
0.0001 ± 0.0002	Bwd IAT Std

Weight	Feature
0.1089 ± 0.0026	Dst Port
0.0571 ± 0.0011	Src Port
0.0109 ± 0.0007	Idle Max
0.0053 ± 0.0008	FWD Init Win Bytes
0.0031 ± 0.0008	Fwd Seg Size Min
0.0024 ± 0.0005	Bwd Init Win Bytes
0.0020 ± 0.0009	Flow IAT Min
0.0017 ± 0.0009	Fwd IAT Min
0.0011 ± 0.0003	Bwd IAT Min
0.0010 ± 0.0008	Idle Min
0.0007 ± 0.0003	Bwd IAT Max
0.0006 ± 0.0004	Idle Mean
0.0006 ± 0.0003	FIN Flag Count
0.0006 ± 0.0001	Fwd Packet Length Std
0.0005 ± 0.0002	Bwd IAT Total
0.0004 ± 0.0003	SYN Flag Count
0.0004 ± 0.0006	Fwd Header Length
0.0003 ± 0.0003	Idle Std
0.0003 ± 0.0001	Bwd Packet Length Std
0.0003 ± 0.0003	ACK Flag Count

Blue Machine

Weight	Feature
0.3103 ± 0.0050	Idle Max
0.1615 ± 0.0027	Bwd Packet Length Min
0.1511 ± 0.0012	Fwd Act Data Pkts
0.1203 ± 0.0018	Idle Mean
0.0464 ± 0.0019	Flow IAT Min
0.0405 ± 0.0026	Fwd Seg Size Min
0.0242 ± 0.0010	Flow IAT Max
0.0065 ± 0.0009	Bwd Init Win Bytes
0.0057 ± 0.0008	Packet Length Max
0.0049 ± 0.0005	FWD Init Win Bytes
0.0046 ± 0.0005	Fwd Packet Length Min
0.0014 ± 0.0002	Flow Bytes/s
0.0002 ± 0.0001	Src Port
0 ± 0.0000	Fwd Packet Length Max
0 ± 0.0000	Bwd IAT Total
0 ± 0.0000	Fwd IAT Max
0 ± 0.0000	Fwd IAT Mean
0 ± 0.0000	Fwd IAT Min
0 ± 0.0000	Fwd IAT Std
0 ± 0.0000	Bwd IAT Std

Weight	Feature
0.3112 ± 0.0016	Dst Port
0.1409 ± 0.0034	Idle Max
0.0929 ± 0.0020	Bwd Packet Length Min
0.0852 ± 0.0021	Src Port
0.0794 ± 0.0010	Flow Bytes/s
0.0670 ± 0.0019	Packet Length Min
0.0615 ± 0.0024	Flow IAT Min
0.0320 ± 0.0017	Fwd Packets/s
0.0266 ± 0.0008	Fwd Seg Size Min
0.0265 ± 0.0014	Flow IAT Std
0.0158 ± 0.0009	Bwd Packets/s
0.0088 ± 0.0005	Fwd Packet Length Max
0.0061 ± 0.0005	Packet Length Std
0.0042 ± 0.0005	Packet Length Variance
0.0029 ± 0.0005	Idle Min
0.0012 ± 0.0002	Packet Length Mean
0.0010 ± 0.0001	Bwd Packet/Bulk Avg
0.0008 ± 0.0001	Flow Duration
0.0007 ± 0.0002	Fwd PSH Flags
0 ± 0.0000	Bwd IAT Tota

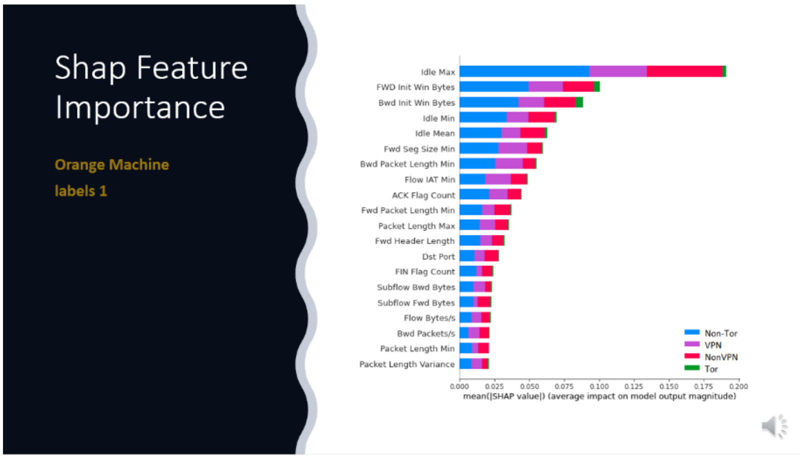
Green Machine

Weight	Feature
0.0210 ± 0.0018	Idle Max
0.0139 ± 0.0004	Fwd Seg Size Min
0.0097 ± 0.0006	FWD Init Win Bytes
0.0038 ± 0.0008	Idle Mean
0.0026 ± 0.0003	Bwd Init Win Bytes
0.0025 ± 0.0006	Idle Min
0.0024 ± 0.0004	Dst Port
0.0012 ± 0.0006	Packet Length Max
0.0007 ± 0.0003	Total Length of Bwd Packet
0.0006 ± 0.0002	Fwd IAT Min
0.0006 ± 0.0002	Bwd IAT Total
0.0006 ± 0.0003	Fwd Packet Length Max
0.0005 ± 0.0003	Bwd Header Length
0.0005 ± 0.0003	Bwd Packet Length Mean
0.0004 ± 0.0002	Total Length of Fwd Packet
0.0004 ± 0.0003	Fwd Packet Length Std
0.0004 ± 0.0001	Bwd IAT Max
0.0003 ± 0.0001	Total Bwd packets
0.0003 ± 0.0001	Bwd IAT Min
0.0003 ± 0.0001	Down/Up Ratio

Weight	Feature
0.0652 ± 0.0016	Dst Port
0.0137 ± 0.0012	Idle Max
0.0059 ± 0.0005	Fwd Seg Size Min
0.0052 ± 0.0013	Idle Min
0.0051 ± 0.0008	Idle Mean
0.0039 ± 0.0004	FWD Init Win Bytes
0.0015 ± 0.0003	Src Port
0.0010 ± 0.0003	Total Fwd Packet
0.0009 ± 0.0010	Fwd Header Length
0.0009 ± 0.0003	Fwd IAT Std
0.0009 ± 0.0002	Bwd Packets/s
0.0008 ± 0.0006	FIN Flag Count
0.0007 ± 0.0006	Bwd Packet Length Max
0.0007 ± 0.0005	Bwd Segment Size Avg
0.0006 ± 0.0002	Subflow Fwd Packets
0.0006 ± 0.0005	Fwd Act Data Pkts
0.0005 ± 0.0001	Subflow Fwd Bytes
0.0004 ± 0.0005	Total Length of Fwd Packet
0.0003 ± 0.0004	Total Bwd packets
0.0003 ± 0.0004	Bw

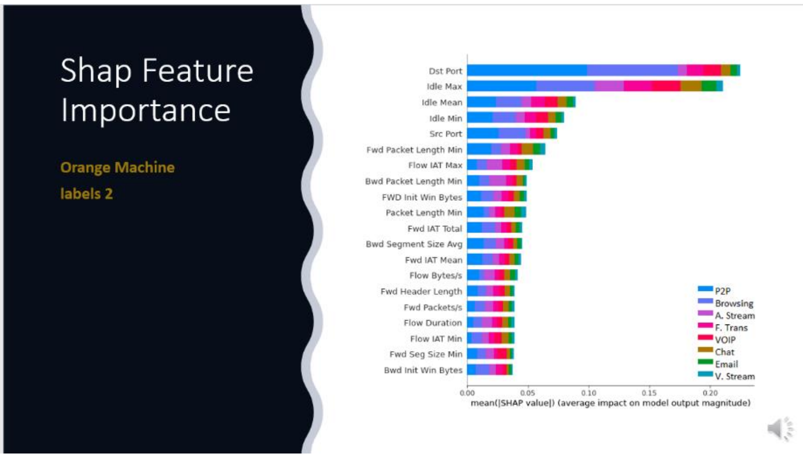
End of Section 2

- Have your original thoughts changed with the introduction of permutation scores?
- Please complete section two of the questionnaire provided.



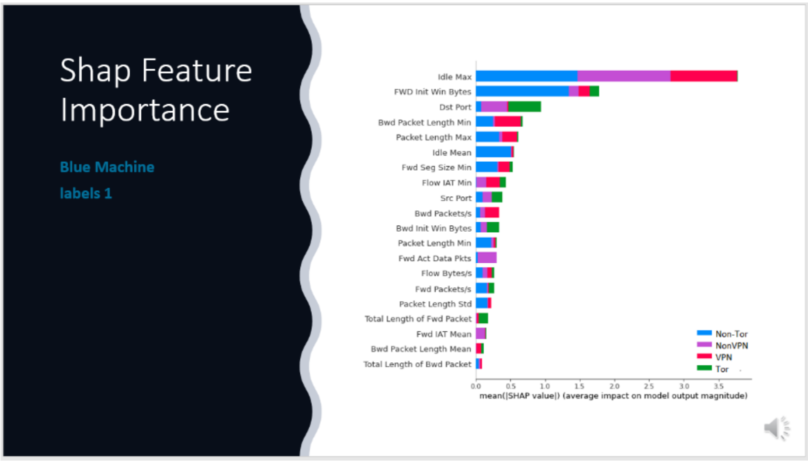
Shap Feature Importance

Orange Machine labels 1



Shap Feature Importance

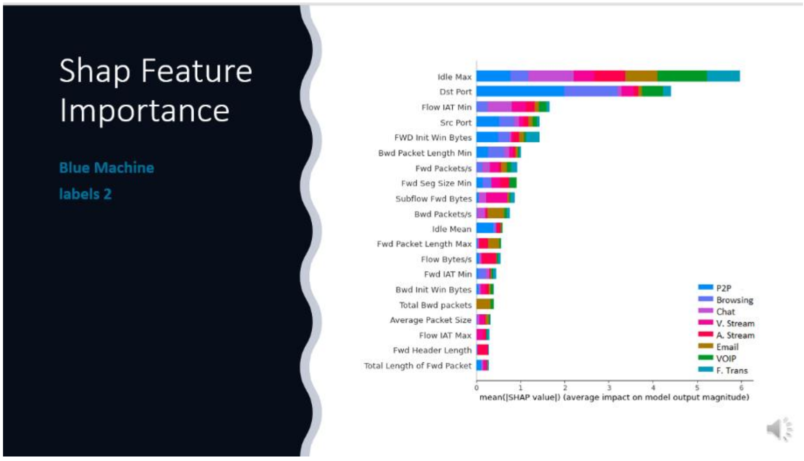
Orange Machine labels 2



Shap Feature Importance

Blue Machine

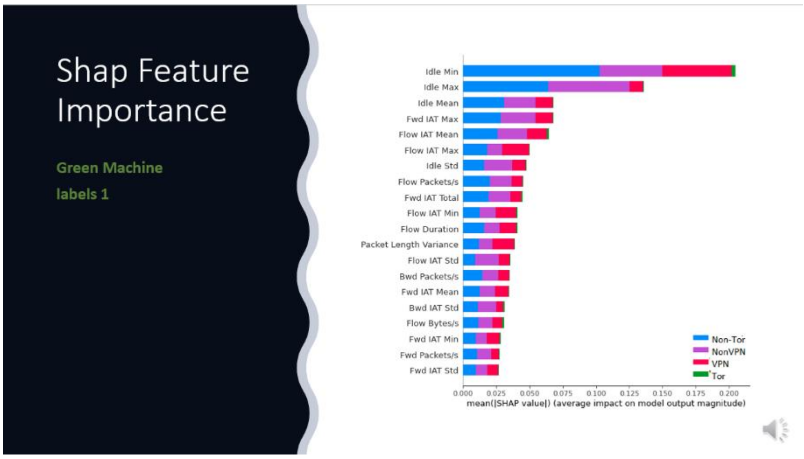
labels 1



Shap Feature Importance

Blue Machine

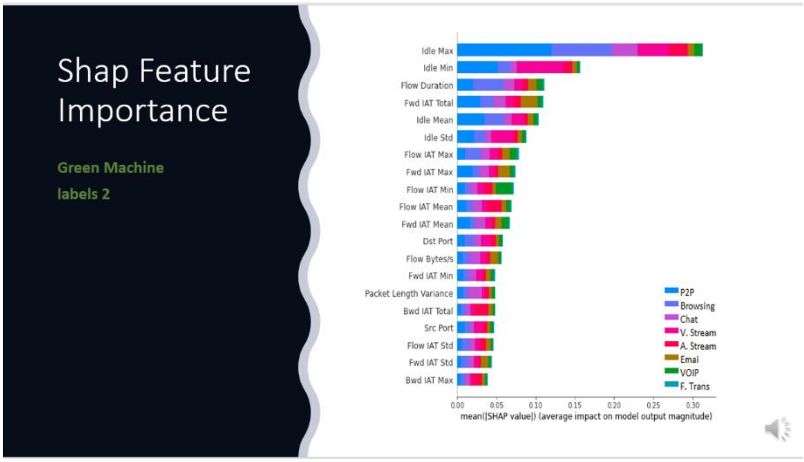
labels 2



Shap Feature Importance

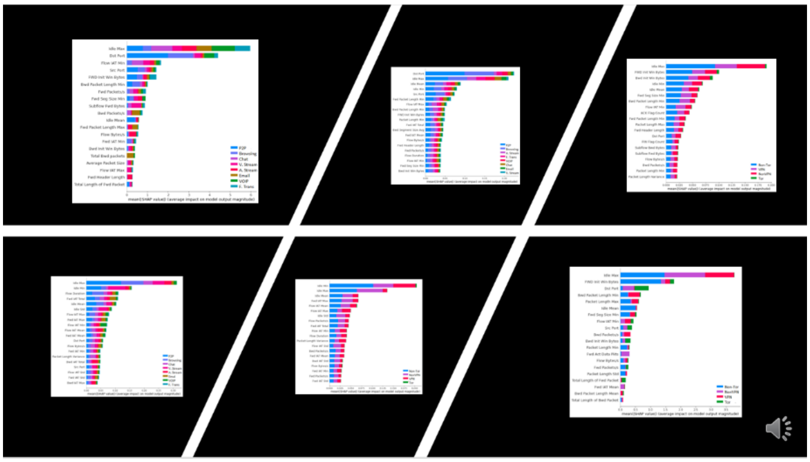
Green Machine

labels 1



Shap Feature Importance

Green Machine labels 2



End of Section 3

- What are your initial thoughts of Shap Feature Importance?
- Please complete section three of the questionnaire provided.

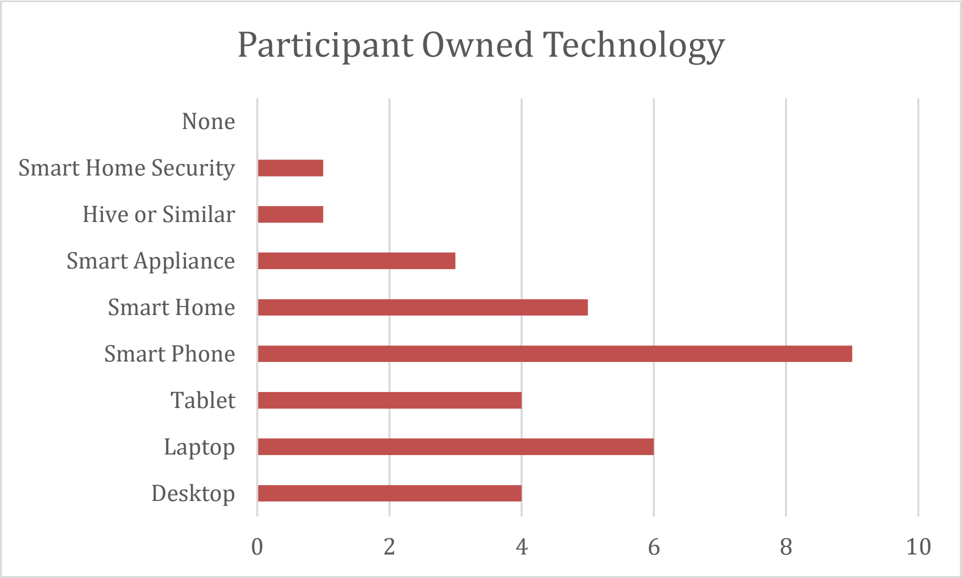
Chapter 4

Results

4.1 Pre-Study Questionnaire.

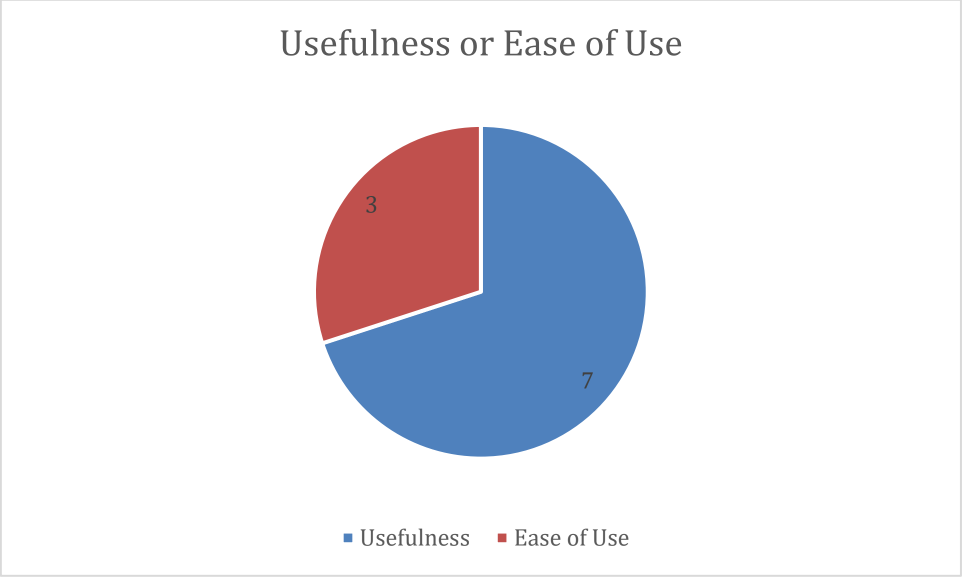
The aim of this questionnaire was to collate information regarding participants initial thoughts to the adoption of new technology. The only personal information collected was age and occupation. It was hoped that a good age range between participants could be achieved even with the small sample size. This was accomplished with a scale between 27 and 55 years of age of a sample size of ten. Morris and Venkatesh [56] found that compared to older workers, younger worker attitudes towards the acceptance of new technology revolved around the use. In contrast, the older workers seemed to be manipulated by subjective norm and perceived behavioural control, however, the effect of the subjective norm did diminish over time. Wang et al [73] state that adoption of new technology can be broken down into three stages; first is preadoption which includes self-management, self-image and negative conceptions of the technology that is being introduced. Second is adoption which includes adoption barriers and usage. Lastly postadoption which includes refusal to adapt to the advancements of the new technology. This is relevant to this study as classification of data that utilises encryption is relatively new. Most people who use application or the darknet do so for added privacy. If people tend to be negative towards the adoption of new technology, then it is going to be difficult to implement new applications even if they are for security and protection.

Q - Which of these do you own?



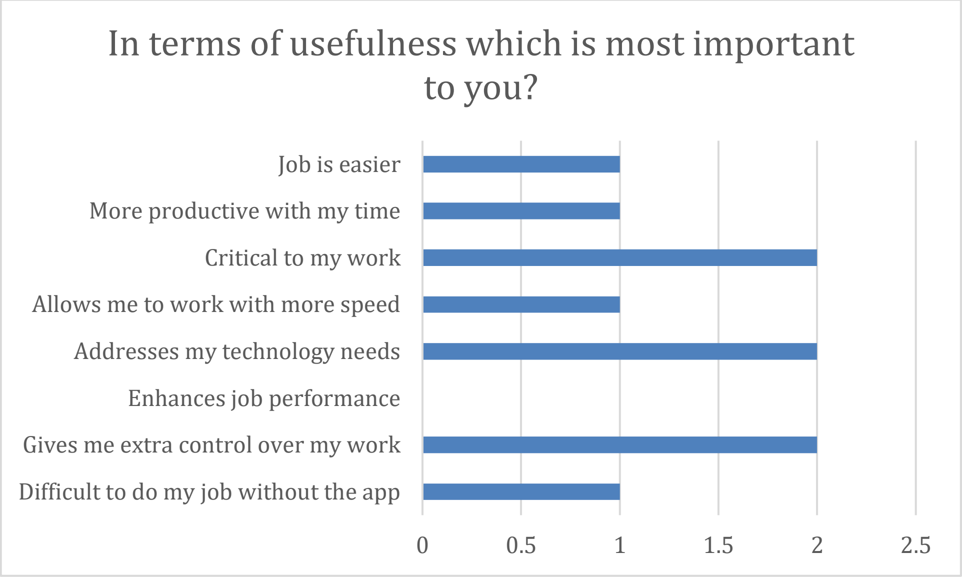
The revolution of technology has changed our lives, we are constantly being stimulated by something. As we can see above all participants in this study own one or more pieces of technology with only one reporting not to own a smartphone. Easy access devices such as tablets and smartphones allow easy access to Internet facilities and entertainment for all ages moving towards the need for low computer literacy skills due to their low maintenance and limited training requirements for use. However, while the uptake in use of these devices can aid learning and communication for both children [11] and adults. Extensive use of technology applications has proven to be disruptive to behaviour, social development and even create sleep disturbances [29], leading to many recommendations to reduce screen time using goal driven tactics. But with more people using Internet facilities and less training required to access these, are users leaving themselves open to security issues due to a lack of understanding of privacy and access data.

Q – What is more important to you in terms of adopting new technology?



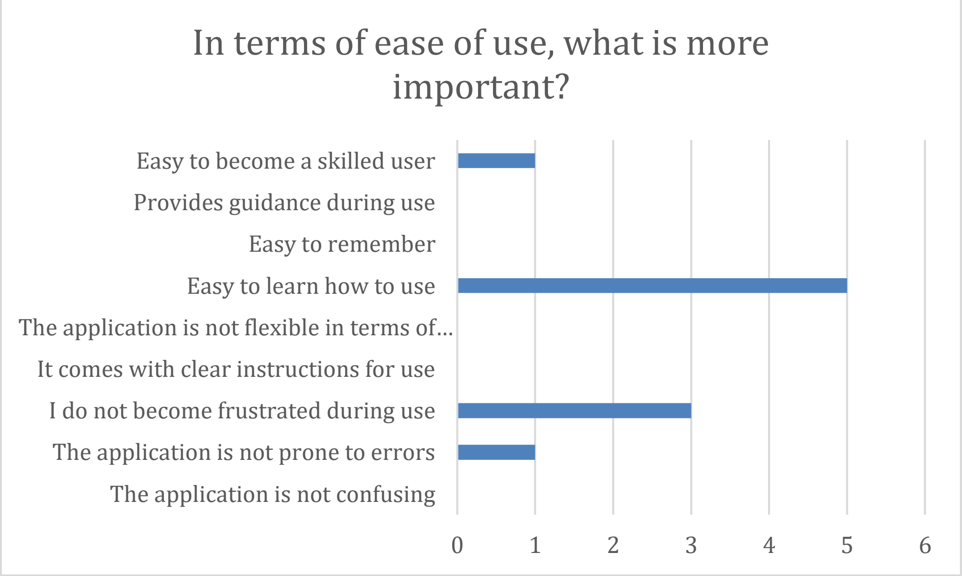
The results here show that an overwhelming majority of participants picked ease of use over usefulness when deciding whether to adapt to technology, this adds further backing to the question above. One model to examine behavioural intention in online settings is the Theory of Planned Behaviour (TPB) however, Hansen et al [35] go further to state that ease of use can moderate the effect of perceived behavioural control. Any perceived risk and trust are significant in user decision making and risk taking propensity influences a user's behavioural intention. Meaning that if a user finds something easy to control and understand leading to autonomous use then it is likely to build trust in that model, increasing the chances of adoption.

Q – In terms of usefulness which is most important to you?



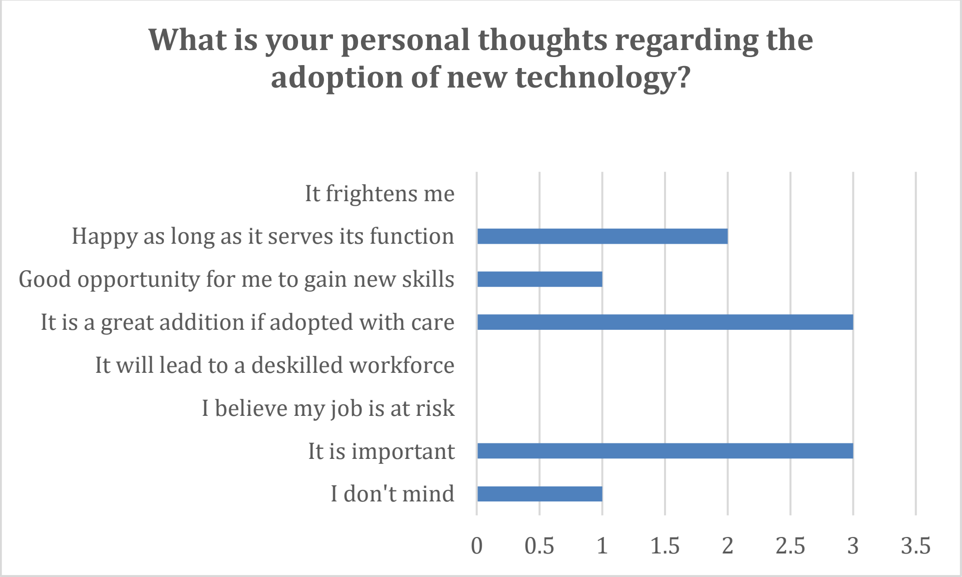
This shows an even distribution of opinion, interestingly there is zero response regarding the enhancement of job performance, but 20% of respondents pick control over their work as important. It would be interesting to expand on this to ascertain why these responses have been picked. Technological needs and criticality to task, also go hand in hand with control but could also show that participants acceptance of technology is out of necessity rather than choice.

Q – In terms of ease of use which is most important to you?



With a 50% response rate to ‘easy to learn’ how to use, there is a clear pattern emerging around ease of use being just that. That there is little to no training required to adapt and use new technology and users can become affiliated with it quickly. This has been noted by Cabero-Almenara et al [14] with their study regarding the adoption of augmented-reality technology by university students believing the enhancement in uptake is due to ease of use and access to the technology. It also allows for safer training with artificial scenarios, it promotes ubiquitous learning and enrichment of available information to all.

Q - What are your personal thoughts around the adoption of new technology?

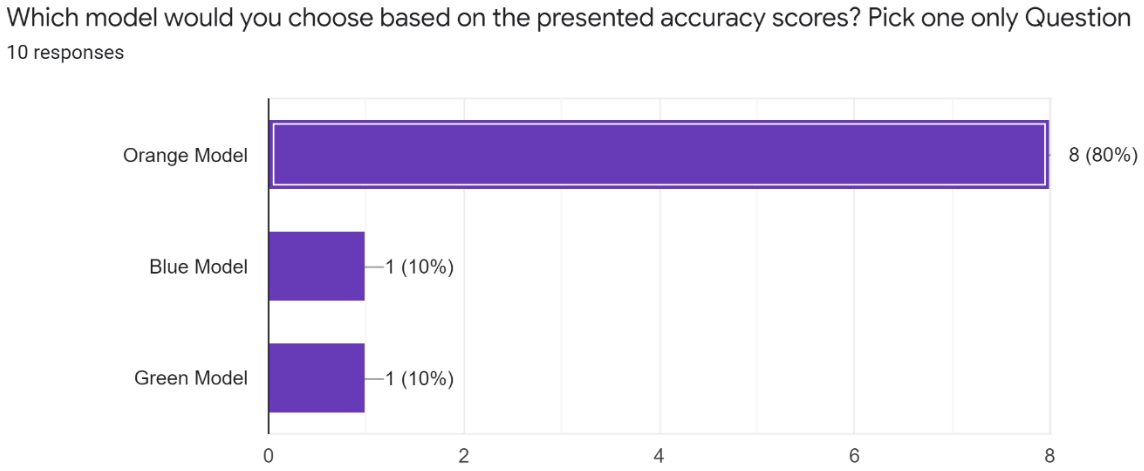


The final question gave participants the opportunity to express their opinion on the adoption of new technology. There is a positive response here if the technology is fundamental to its application and is implemented with care and attention.

4.2 Presentation and Questionnaire Results

Technology is irrefutably changing our working lives, the participants occupations in this study include Performance Coach, Nurse, Personal Trainer, Finance Analyst, Engineer, Underwriter, Hotel Manager and retired. All of which are impacted differently by technology over time. Examples such as going from paper only files to completely digital, to being able to produce interactive 3D plans. Most were positive experiences which meant that jobs had become more streamlined. However, there was some doubts over issuing complete autonomy to an algorithm. This is backed by Schneider and Leyer [65] who also found that participants with low situation awareness were more likely to delegate to an algorithm. This became apparent in this study also which will be discussed below.

This part of the study was completed by participants during the presentation of the video. They were presented with three different network classifiers, the Orange Model, Green Model and Blue Model. The orange model represents a Neural Network Classifier, the green model a Random Forest Classifier and finally the blue model represents a Decision Tree Classifier. The Random Forest was not initially chosen, it replaced the planned SVM Classifier that had been designed, unfortunately, due to time constraints it had to be abandoned after 4 days of training the model had not completed. Hopefully, this is something that can be looked at in the future. The researcher also looked at using a Naïve Bayesian classifier, but the RAM bottomed out during training, again, this is something to hopefully attempt later.
Participants were not informed as to which colour represented which model, this was an attempt at preventing a pre-emptive choice of preference prior to the study. The following results were noted during the study:

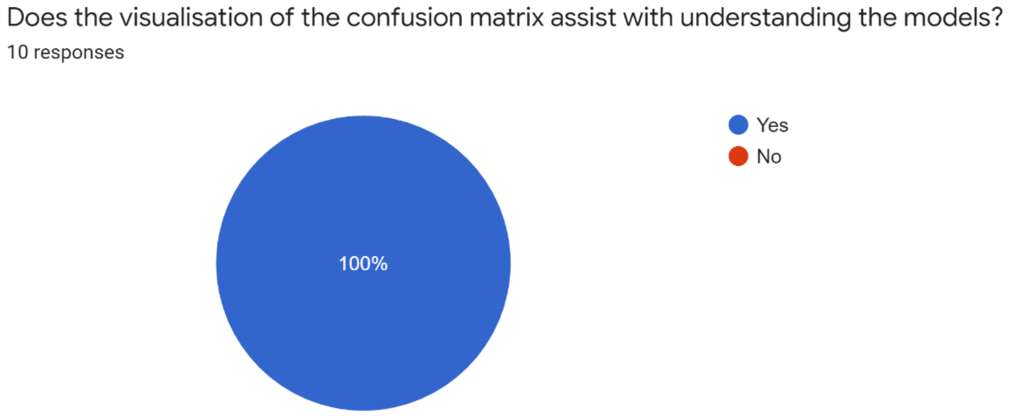


Which model would you choose based on the presented accuracy scores? Pick one only Question
10 responses

The accuracy of the models was as follows:

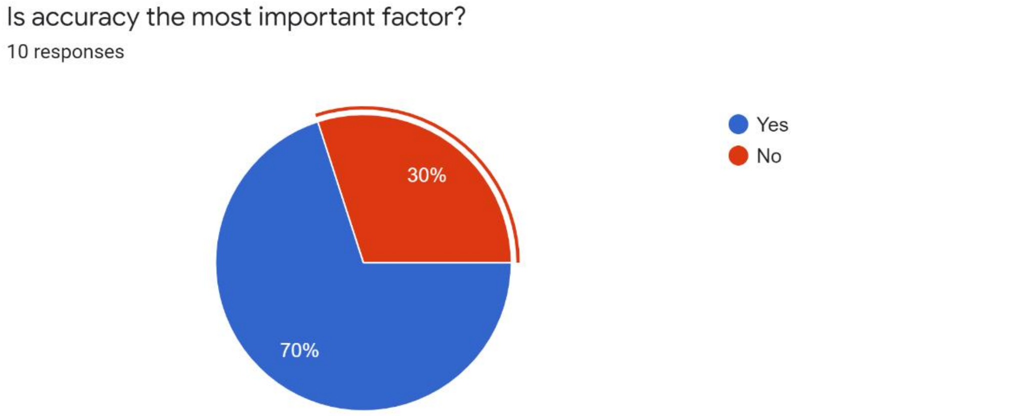
Orange machine – Label 1 score – 98.2%
Orange machine - Label 2 score – 88.8%
Blue Machine – Label 1 score – 98.2%
Blue Machine – Label 2 score – 84.6%
Green Machine – Label 1 score – 72.5%
Green Machine – Label 2 score – 61.0%

When presented with these figures it is unsurprising that 80% of participants picked the orange model as the best model based on accuracy alone. Participants were also presented with a visualisation of accuracy in the form of a confusion matrix. When asked why they had picked this model, responses ranged from ‘orange is my favourite colour’ to simply ‘had the highest accuracy score.’ A couple of responses also noted that the visualisation of the confusion matrix was ‘very confusing’ and ‘acted as a distraction’ these participants both confirmed that the visualisation did not provide a deeper explanation for them and would prefer to be shown just the percentage figures. One participant did state that it was easy to see the accuracy clearly on the orange matrix when compared to the others but that it made no difference to their choice.


Does the visualisation of the confusion matrix assist with understanding the models?
10 responses

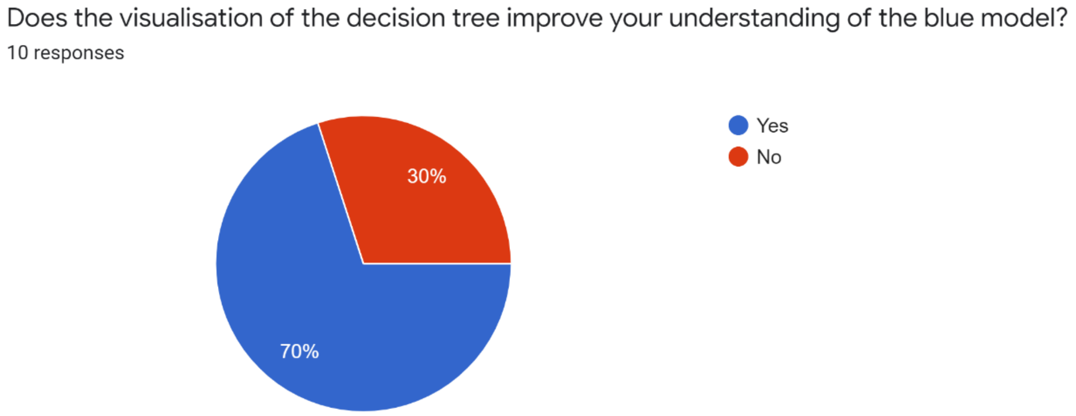
Despite these comments, 100% of respondents said that the visualisation of the matrices assisted with the understanding of the models.

Next participants were asked if accuracy is the most important factor:


Is accuracy the most important factor?
10 responses

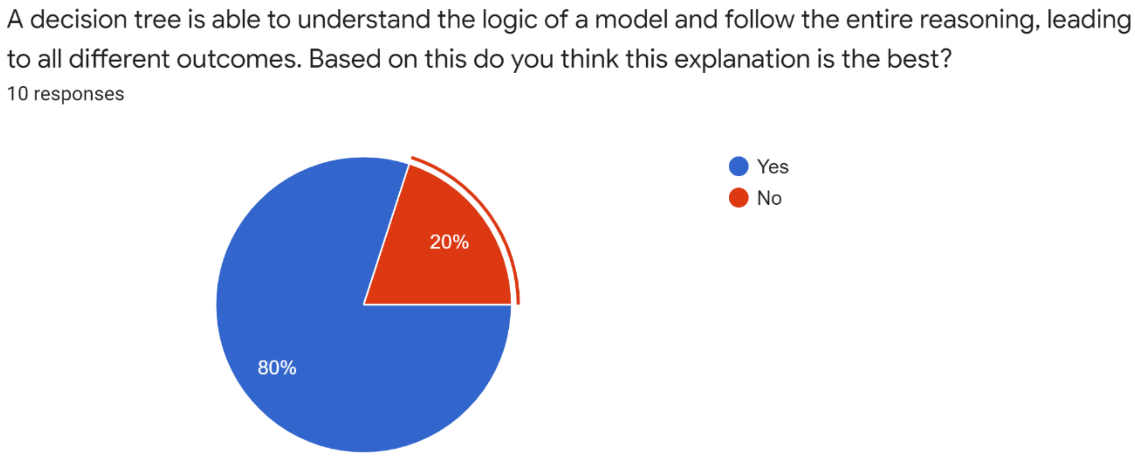
As we can see most participants picked yes to this question. Emphasis is put on the importance of accuracy especially if there is limited understanding of something. But this is not always the correct approach, when using classifiers accuracy may be high when identifying information such as key words or image detection, however, as we have seen above with Facebook's issue with COVID-19, this could also lead to removal of legitimate articles. This problem and example were explained to each participant at this point.

The blue model showed the ability to produce a decision tree:


Does the visualisation of the decision tree improve your understanding of the blue model?

10 responses

Seventy percent of responses determined that the decision tree visualisation aided in the understanding of the blue model. Participants were also asked:

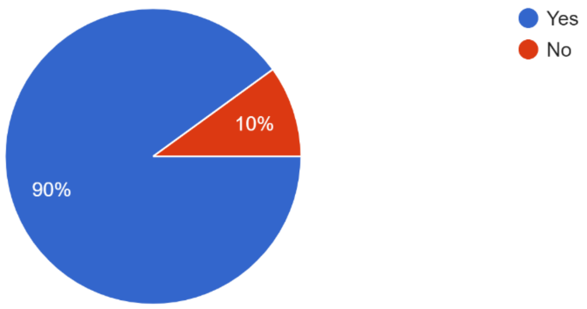

A decision tree is able to understand the logic of a model and follow the entire reasoning, leading to all different outcomes. Based on this do you think this explanation is the best?

10 responses

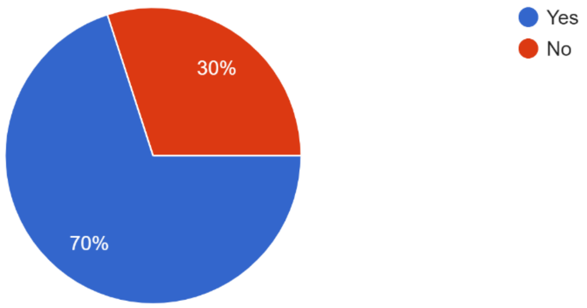
Most participants believed that the decision was understandable and simple to follow, based on this, at this point in the presentation they believed that this explanation was the best method.

Next in the presentation was global explanation method using permutation scores. Participants were shown individual scores for each model and then a summary screen at the end of the section. Here are the results.

Do you understand the function of permutation scores?
10 responses

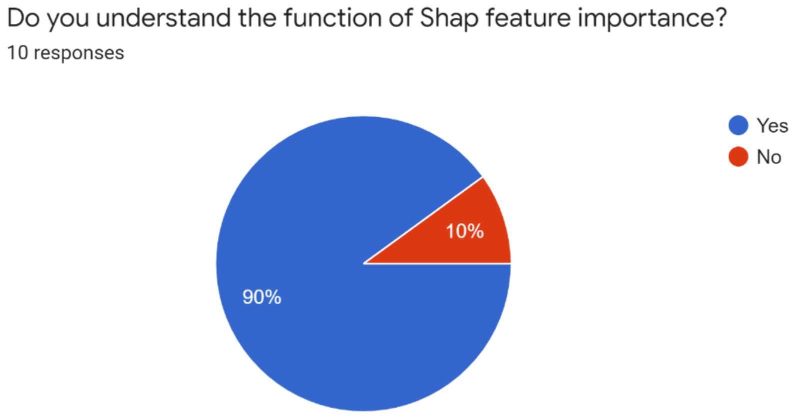


Does understanding this function help you understand the models?
10 responses



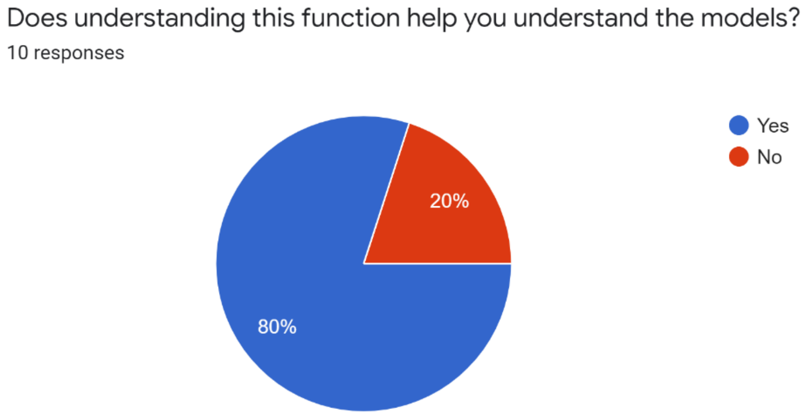
Ninety percent of participants stated that they understood permutation scores, but only 70% stated that it enhanced their understanding of the models. Some mentioned that the added detail developed their understanding, but that because they had no background knowledge in the subject, they found it difficult to appreciate the topic. Despite this the information was presented well and the researcher was able to help them gain some basic understanding. One participant was confused by the weighting element, another stated that this was helpful to gain more insight as they understood the weighting coming from a financial background. This is interesting to note, developers need to take backgrounds into account with certain design elements here.

The last description offered was local explanation using ShAP. These are the results:



Do you understand the function of Shap feature importance?

10 responses



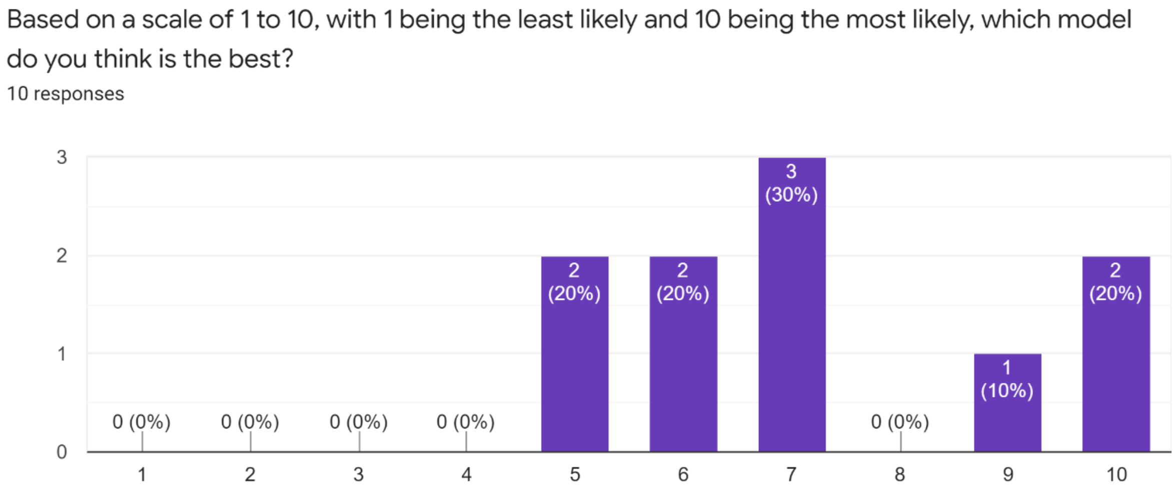
Does understanding this function help you understand the models?

10 responses

Again, 90% of participants understood this explanation and 80% felt that the function helped them understand the classifiers better. Participants stated that the data was presented in a clearer format which was easier to understand immediately, 'Output graphs provided strong visuals which were easy to read.' Simply the addition of colour made the visuals more appealing to look at meaning that they paid more attention to what was being presented and that the different colours made it easy to identify the breakdown of features. Most thought that this method was easier to understand for the lay person and offers the opportunity to pull apart the information into more detail, allowing for improved analysis.

Finally, participants were asked to give ratings on elements of the models and accuracy scores. First. they were asked to rate the models based on 1 – 10, 1 being poor and 10 being the best. Here is the summary:

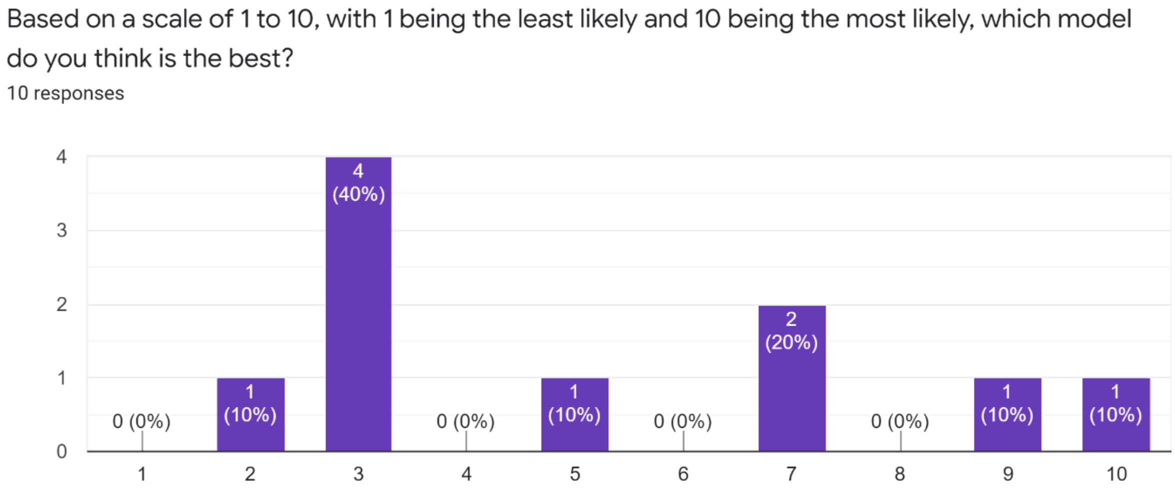
Orange Model



Based on a scale of 1 to 10, with 1 being the least likely and 10 being the most likely, which model do you think is the best?

10 responses

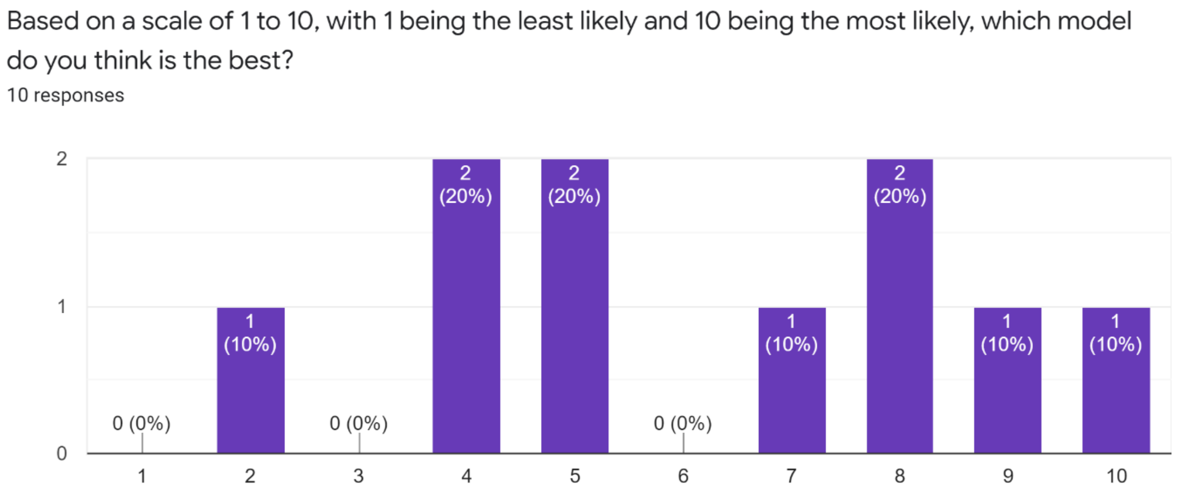
Green Model



Based on a scale of 1 to 10, with 1 being the least likely and 10 being the most likely, which model do you think is the best?

10 responses

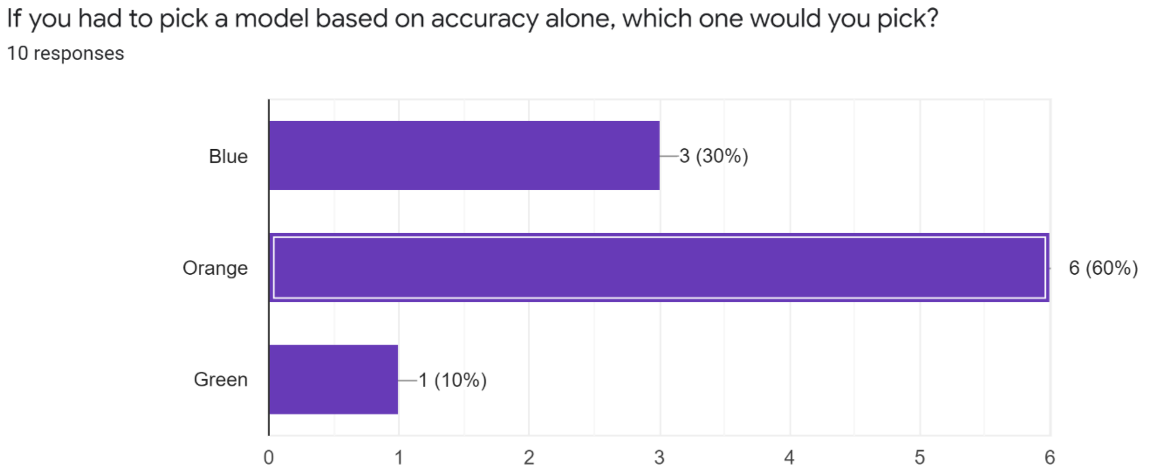
Blue Model



Based on a scale of 1 to 10, with 1 being the least likely and 10 being the most likely, which model do you think is the best?

10 responses

There seems to be a much more conflicted response than earlier in the study when participants were asked to pick a model based on accuracy alone. This seems to be backed by the next question when asked for a second time, which model they would pick based on accuracy. This shows that given an adequate explanation or example, people tend to think more about their response.

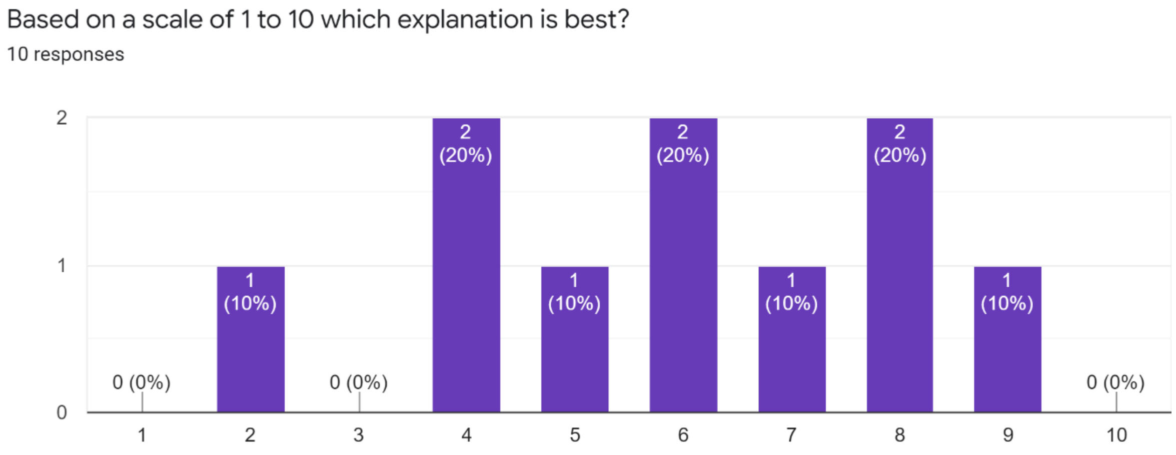


If you had to pick a model based on accuracy alone, which one would you pick?

10 responses

Following these participants were asked to rate each explanation method based on a scaled of 1 – 10. Here are the results

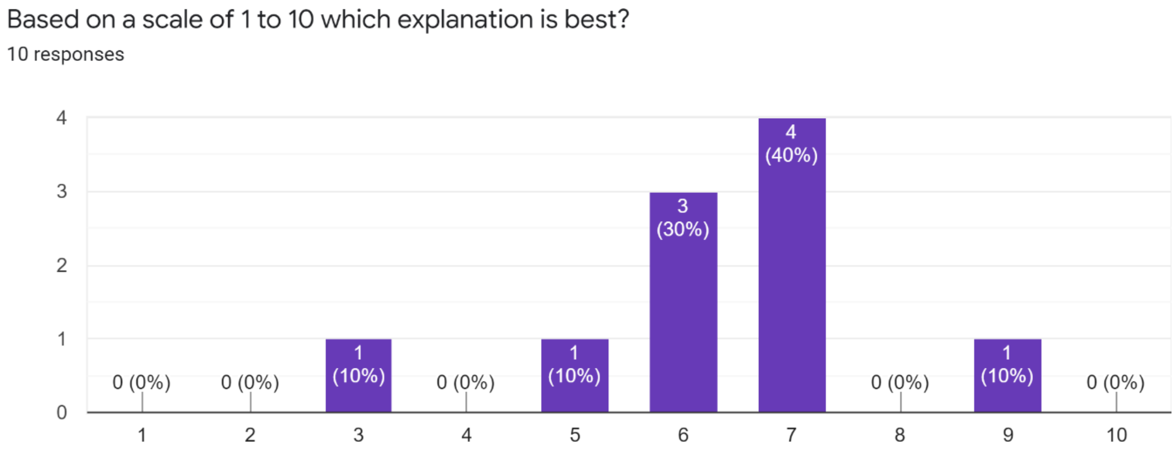
Decision Tree



Based on a scale of 1 to 10 which explanation is best?

10 responses

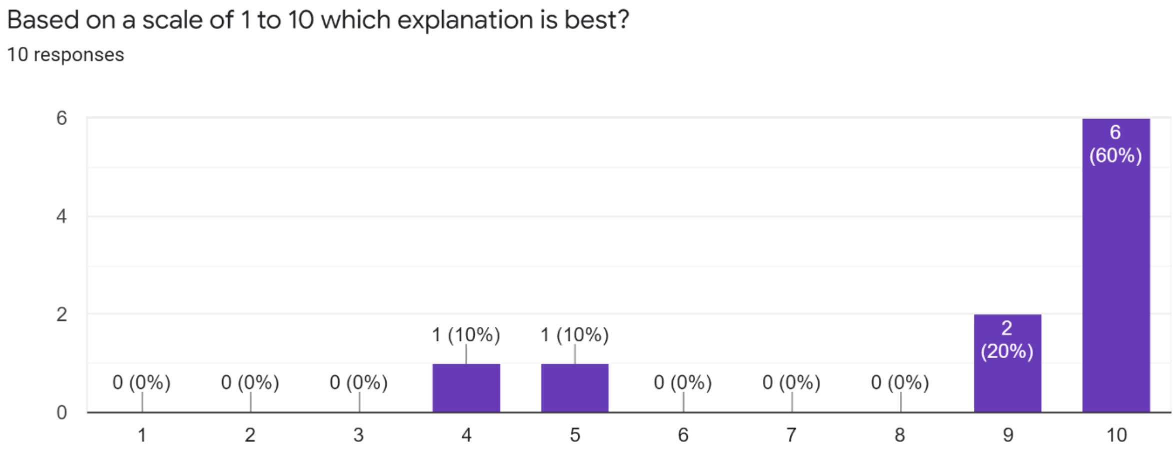
Permutation Score



Based on a scale of 1 to 10 which explanation is best?

10 responses

ShAP Feature Importance



Clearly ShAP feature importance is the clear favourite and is back by participant responses earlier in the questionnaire. This is quite surprising, initial expectations anticipated the decision tree explanation method to come out on top due to users being able to follow the decision process from the beginning through to a conclusion.

Finally, participants were asked to explain briefly why they has rated the explanation methods the way they had. Responses to this clearly promote the preference for ShAP feature importance, with participants stating that it 'was the best visualisation of the data,' 'offering a more simplistic fashion.' Some liked that it 'broke down all the data,' 'clearly representing feature importance,' 'supported the way they learn and needed less concentration time to understand.' However, one participant did make an argument for the blue model, believing that decision trees provided a better breakdown and would be easier for the majority to understand but that the orange model seemed to provide better ShAP results, with blue having wider +/- scores in the modelling.

Chapter 5

Conclusion

5.1 Discussion and Conclusion

Legitimacy is more important than accuracy. It is important to maintain a secure environment for Internet users, but as many of us use the Internet as a news and fact checker facility, it is important that any content that is removed is done so with accuracy, but legitimate posts are left alone. A wide range of public activities such as politics, journalism and civil engagement are conducted online, therefore the, the decisions that these platforms make have a substantial impact on public culture and the social and political lives of their users [28]. Regrettably, 'the black-box nature of content moderation on most platforms means that few good data are available about how these platforms make moderation decisions' [37]. With online communities thriving online it is important to be transparent as to why content has been identified as reprehensible and removed. One strategy for improving transparency is to provide feedback to the user. If a user is provided with the right explanation, at the right time, moderations system have the propensity to become a useful addition to Internet security and improve user experience of online communities [28]. Kim et al [38] argue that one of the most significant contributions to the field of A.I. is to design models that are transparent and accountable. This would allow the field to enhance the applicability and accountability for real world decision support. Trust us a key component when deploying big data-driven system solutions for decision making [9]. Users will not by into a system or accept the solution if there is no trust in it. Therefore, it can be assumed that by making a model with increased transparency is a fundamental step in building trust [38] and increasing the likelihood of adoption.

Participants were first presented with the accuracy of each model's performance. They were initially given each score as a percentage and then as a visualisation in the form of a confusion matrix. A confusion matrix is not a metric that is useful to evaluate a model but can provide some insight into predictions [79]. It provides richer detail behind accuracy by showing both accurate and inaccurate values. This study highlighted how difficult a confusion matrix can be to understand without any prior knowledge with one participant stating that ‘a confusion matrix certainly lives up to its name. In conclusion here, it is the belief that confusion matrices do add more depth than simple accuracy scores, but they are not utilised widely enough for broad understanding.

There is a growing need for machine learning models to not only provide accurate predictions but predictions that are also interpretable to assist with human decision making, this is especially important in crucial applications such as healthcare [78]. However, interpretability has different ambitions that are not always aligned even with the most generalised model architecture [43], this is also further compounded if users have strict and domain specific requirements for interpretability [60]. ‘Tree-based machine learning models such as random forests, decision trees and gradient boosted trees are popular nonlinear predictive models, yet comparatively little attention has been paid to explaining their predictions’ [45]. A decision tree algorithm can solve both regression and classification problems. There are some limitations decision trees, for instance a small change in the data can change the tree considerably causing instability. Calculations can be complex when compared to some other algorithms. They can be time consuming during training, thus making it expensive as the complexity absorbs more time and effort. It is inadequate to apply to regression and predicting continuous values. However, they require less pre-processing or normalisation of data. Missing values in the dataset do not affect the building of a decision tree, they are intuitive and easy to explain [22], much like a flow chart diagram. For this reason, it was assumed that this model would be picked as the preferred explanation by participants. It was a surprise that this was not the case. Only one participant out of the whole sample referred to decision tree understanding, demonstrating that it is in fact deemed complex.

Random forest eradicates the limitations of decision trees by reducing overfitting and increasing precision, it also generates predictions without the need for many configurations [049]. The accuracy and robustness of random forests [13] have been a popular and effective method in machine learning. However, they are not considered to be interpretable because they aggregate many decision trees, each of which is often quite large [59]. This again was demonstrated in the study, to keep things simple and comfortable to present to participants with limited knowledge of the field the visualisation was kept to a minimum and only highlighted one feature. This was explained to the participants, whilst many understood this, they found it difficult to understand and producing limitations when trying to instil trust is not ideal. However, since they are less time consuming than decision tree models and solve the problem of overfitting, in some circumstances this model is perfect for developers working for organisations that require accurate strategic decision making [049]. Therefore, it would be beneficial to spend more time on Random Forest transparency and increasing trust and adoption in wider society.

Neural network's main advantage is the ability to solve non-linear problems. 'This means that neural networks can generally be tested against a problem with an unknown shape even if other classes of machine learning algorithms have already failed' [8]. They do however, in comparison to other models require large datasets to be trained effectively. Due to this they often need substantial computational power to be trained which can be costly. Another problem arises if the dataset is too large as they do not scale well if the number of layers of neurons are increased. The order in which data is fed into a neural network can affect the outcome. Neural network architectures are not all the same, meaning that different architectures can solve the same problem and produce different conclusions. Despite its accuracy other algorithms should be considered if the dataset or computational power is insufficient [8].

Global explanation methods produce the average performance of a model's behaviour [51] for this study, permutation feature importance was chosen as the desired explanation visualisation. Permutation feature importance measures the increase in prediction error of a model after the feature values have been permutated. It provides a condensed, global insight into a model's behaviour. It automatically considers all interactions with other features, meaning it considers the main feature effect and the interaction effects on the model's performance. A major advantage is that permutation feature importance does not require retraining of the model for this method to be useful, you need the results of a fixed feature importance. Unfortunately, it is not clear as to whether test of training data should be used [51] which can be confounding especially with time constraints that were faced during this study. If features are correlated, then the outcome can be biased by unrealistic data instances. Participants found that the visualisation helped understand the concept of what was being presented but felt that the method did not go much further with explanation or increasing transparency. They could see which features were rated the highest due to the weight applied but did not understand the method behind this. Further explanation is required here, further development to the visualisations produced.

Local explanation methods justify individual predictions. As an explanatory tool ShAP method can facilitate both local and global interpretations [22]. It can produce an overall feature importance explanation right through to a singular feature clarification. This increases transparency of the model and allows for greater analysis and understanding [82]. It is the only explanation method with a solid theory containing efficiency and symmetry giving it a good foundation. ShAP allows for contrastive analysis meaning it can be utilised to compare a subset or even a single data point. There are some disadvantages however, such as it takes a long time computationally to produced results. The results can be easily misinterpreted, the feature value is not the difference of the predicted value after removing the feature from the model training but rather ‘the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value [51]. This method was most popular with participants picking it as the best interpretable and understandable technique of explanation. They believed it was the most simplistic method presented and appreciated different features being represented by different colours.

It would be desirable to invest more effort into problems, approaches, and architectures [47]. To aid in building trust and augment adoption of new technology.

5.2 Limitations and Future Work

- Time was the biggest constraint for this study. Chosen classifiers such as SVM were changed due to the time it was taking to train the model. Although frustrating, it is hoped that this method could be utilised at a later date.

- Small sample size, although the sample contained a good range of age and occupations, it would have been preferable to have been bigger, also containing some occupations from the computer science field to provide a comparison to opinions provided by those outside the field.

- COVID-19, most interactions with participants were conducted online via Zoom. It is believed a richer interaction may have occurred if the presentations had been in person.

Bibliography

[1] Acemoglu, D. and Restrepo, P. The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review 108*, 6 (2018), 1488-1542.

[2] Akiyoshi, R., Kotani, D. and Okabe, Y. Detecting Emerging Large-Scale Vulnerability Scanning Activities by Correlating Low-Interaction Honeypots with Darknet. *Ieeexplore.ieee.org*, 2018. https://ieeexplore.ieee.org/document/8377942.

[3] Al-Nabki, M., Fidalgo, E., Alegre, E. and Fernández-Robles, L. ToRank: Identifying the most influential suspicious domains in the Tor network. *Expert Systems with Applications 123*, (2019), 212-226.

[4] Alshammari, R. and Zincir-Heywood, A. Machine learning based encrypted traffic classification: Identifying SSH and Skype. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, (2009).

[5] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M. and Rudin, C. Learning Certifiably Optimal Rule Lists for Categorical Data. *arXiv.org*, 2016. https://arxiv.org/abs/1704.01701v4.

[6] Arntz, M., Gregory, T. and Zierahn, U. OECD Social, Employment and Migration Working Papers. *OECD Social, Employment and Migration Working Papers*, 189 (2016).

[7] Arras, L., Montavon, G., Müller, K. and Samek, W. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *arXiv.org*, 2017. https://arxiv.org/abs/1706.07206.

[8] Baeldung. Advantages and Disadvantages of Neural Networks. *Baeldung CS*, 2020. https://www.baeldung.com/cs/neural-net-advantages-disadvantages.

[9] Baesens, B., Bapna, R., Marsden, J., Vanthienen, J. and Zhao, J. Transformational Issues of Big Data and Analytics in Networked Business. *MIS Quarterly 40*, 4 (2016), 807-818.

[10] Baughman, D. and Liu, Y., 1995. Classification: Fault Diagnosis and Feature Categorization. *Neural Networks in Bioprocessing and Chemical Engineering*, pp.110-171.

[11] Blum-Ross, A. and Livingstone, S., 2016. *Families and screen time: Current advice and emerging research*. [online] Eprints.lse.ac.uk. Available at: <http://eprints.lse.ac.uk/66927/1/Policy%20Brief%2017-%20Families%20%20Screen%20Time.pdf> [Accessed 15 September 2021].

[12] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D. and Rinzivillo, S. Benchmarking and Survey of Explanation Methods for Black Box Models. *arXiv.org*, 2021. https://arxiv.org/abs/2102.13076.

[13] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5-32.

[14] Cabero-Almenara, J., Fernández-Batanero, J. and Barroso-Osuna, J., 2019. Adoption of augmented reality technology by university students. *Heliyon*, 5(5), p.e01597.

[15] Chakure, A. Decision Tree Classification. *Medium*, 2019. https://medium.com/swlh/decision-tree-classification-de64fc4d5aac.

[16] Chattopadhyay, A., Manupriya, P., Sarkar, A. and Balasubramanian, V. Neural Network Attributions: A Causal Perspective. *PMLR*, 2019. http://proceedings.mlr.press/v97/chattopadhyay19a.html.

[17] Crossman, A., 2019. *How a Pilot Study Can Improve Sociological Research*. [online] ThoughtCo. Available at: <https://www.thoughtco.com/pilot-study-3026449> [Accessed 14 September 2021].

[18] Dainotti, A., Pescape, A. and Claffy, K. Issues and future directions in traffic classification. *IEEE Network 26*, 1 (2012), 35-40.

[19] Davis, F. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly 13*, 3 (1989), 319.

[20] Dekker, F. and Salomens, A. Fear of robots at work: the role of economic self-interest. *Socio-Economic Review 15*, 3 (2017).

[21] Demertzis, K., Tsiknas, K., Takezis, D., Skianis, C. and Iliadis, L. Darknet Traffic Big-Data Analysis and Network Management for Real-Time Automating of the Malicious Intent Detection Process by a Weight Agnostic Neural Networks Framework. *Electronics 10*, 7 (2021), 781.

[22] Deshmukh, F. and Merchant, S. Explainable Machine Learning Model for Predicting GI Bleed Mortality in the Intensive Care Unit. *American Journal of Gastroenterology 115*, 10 (2020), 1657-1668.

[23] Dodel, M. and Mesch, G. Perceptions about the impact of automation in the workplace. *Information, Communication & Society 23*, 5 (2020), 665-680.

[24] Doody, O. and Doody, C., 2015. Conducting a pilot study: case study of a novice researcher. *British Journal of Nursing*, 24(21), pp.1074-1078.

[25] Erasmus, A., Brunet, T. and Fisher, E. What is Interpretability?. *Philosophy & Technology*, (2020).

[26] Fong, R. and Vedaldi, A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *Openaccess.thecvf.com*, 2017. https://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html.

[27] Garcia-Dorado, J., Finamore, A., Mellia, M., Meo, M. and Munafo, M. Characterization of ISP Traffic: Trends, User Habits, and Access Technology Impact. *IEEE Transactions on Network and Service Management 9*, 2 (2012), 142-155

[28] Gillespie, T. Platforms Intervene. *Social Media + Society 1*, 1 (2015), 205630511558047.

[29] Gilpin, A. and Taylor, G., 2018. *Changing behaviour: Children, Adolescents and screen time*. [online] Bps.org.uk. Available at: <https://www.bps.org.uk/sites/beta.bps.org.uk/files/Policy%20-%20Files/Changing%20behaviour%20-%20children%2C%20adolescents%2C%20and%20screen%20use.pdf> [Accessed 15 September 2021].

[30] GLT. Types of Neural Networks and Definition of Neural Network. *GreatLearning Blog: Free Resources what Matters to shape your Career!*, 2021. https://www.mygreatlearning.com/blog/types-of-neural-networks/.

[31] Graetz, G. and Michaels, G. Robots at Work. *The Review of Economics and Statistics 100*, 5 (2018), 753-768.

[32] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys 51*, 5 (2019), 1-42.

[33] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G. XAI—Explainable artificial intelligence. *Science Robotics 4*, 37 (2019), eaay7120.

[34] Hadlington, L., White, H. and Curtis, S., 2019. “I cannot live without my [tablet]”: Children's experiences of using tablet technology within the home. *Computers in Human Behavior*, 94, pp.19-24.

[35] Hansen, J., Saridakis, G. and Benson, V., 2018. Risk, trust, and the interaction of perceived ease of use and behavioral control in predicting consumers’ use of social media for transactions. *Computers in Human Behavior*, 80, pp.197-206.

[36] Hong, Y., Huang, C., Nandy, B. and Seddigh, N. Iterative-tuning support vector machine for network traffic classification. *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, (2015).

[37] Jhaver, S., Bruckman, A. and Gilbert, E. Does Transparency in Moderation Really Matter?. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 1-27.

[38] Kim, B., Park, J. and Suh, J. Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems 134*, (2020), 113302.

[39] Knocklein, O. Classification Using Neural Networks. *Medium*, 2019. https://towardsdatascience.com/classification-using-neural-networks-b8e98f3a904f.

[40] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications 10*, 1 (2019).

[41] Larson, J., Mattu, S., Kirchner, L. and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[42] Lashkari, A., Kaur, G. and Rahali, A. DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning | 2020 the 10th International Conference on Communication and Network Security. *Dl.acm.org*, 2020. https://dl.acm.org/doi/abs/10.1145/3442520.3442521.

[43] Lipton, Z. The Mythos of Model Interpretability. *Queue 16*, 3 (2018), 31-57.

[44] London, A. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report 49*, 1 (2019), 15-21.

[45] Lundberg, S., Erion, G. and Chen, H. et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2*, 1 (2020), 56-67.

[46] Malmqvist, J., Hellberg, K., Möllås, G., Rose, R. and Shevlin, M. Conducting the Pilot Study: A Neglected Part of the Research Process? Methodological Findings Supporting the Importance of Piloting in Qualitative Research Studies. *International Journal of Qualitative Methods 18*, (2019), 160940691987834.

[47] Manning, C. Computational Linguistics and Deep Learning. *Computational Linguistics 41*, 4 (2015), 701-707.

[48] Maqbool, M., Ramzan, M., Khan, S., Rehman, I. and Khan, T., 2021. A Pilot Study on Online-Education Supportive Tools in COVID-19 Context. *IT Professional*, 23(4), pp.63-68.

[49] Matsakis, L. and Martineau, P., 2021. *Coronavirus Disrupts Social Media's First Line of Defense*. [online] Wired. Available at: <https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/> [Accessed 8 September 2021].

[049] Mbaabu, O., 2019. *Introduction to Random Forest in Machine Learning*. [online] Engineering Education (EngEd) Program | Section. Available at: <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/> [Accessed 28 September 2021].

[50] Mittelstadt, B., Russell, C. and Wachter, S. Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, (2019).

[050] Mokyr, J., Vickers, C. and Ziebarth, N., 2015. The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different?. *Journal of Economic Perspectives*, 29(3), pp.31-50.

[51] Molnar, C. Interpretable Machine Learning. *Christophm.github.io*, 2021. https://christophm.github.io/interpretable-ml-book/shapley.html#advantages-16.

[52] Montavon, G., Samek, W. and Müller, K. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing 73*, (2018), 1-15.

[53] Moore, A. and Papagiannaki, K. Toward the Accurate Identification of Network Applications. *Lecture Notes in Computer Science*, (2005), 41-54.

[54] Moore, A. and Zuev, D. Internet traffic classification using bayesian analysis techniques. *ACM SIGMETRICS Performance Evaluation Review 33*, 1 (2005), 50-60.

[55] Moraffah, R., Karami, M., Guo, R., Raglin, A. and Liu, H. Causal Interpretability for Machine Learning - Problems, Methods and Evaluation. *ACM SIGKDD Explorations Newsletter 22*, 1 (2020), 18-33.

[56] MORRIS, M. and VENKATESH, V., 2006. AGE DIFFERENCES IN TECHNOLOGY ADOPTION DECISIONS: IMPLICATIONS FOR A CHANGING WORK FORCE. *Personnel Psychology*, 53(2), pp.375-403.

[57] Ooi, K. and Tan, G. Mobile technology acceptance model: An investigation using mobile users to explore smartphone credit card. *Expert Systems with Applications 59*, (2016), 33-46.

[58] Parsons, A., 2018. *Davos: Theresa May to urge tech firms to 'go further' over illegal content*. [online] Sky News. Available at: <https://news.sky.com/story/davos-theresa-may-to-urge-tech-firms-to-go-further-over-illegal-content-11221511> [Accessed 8 September 2021].

[59] Plumb, G., Molitor, D. and Talwalkar, A. Model Agnostic Supervised Local Explanations. *arXiv.org*, 2019. https://arxiv.org/abs/1807.02910v3.

[60] Rafique, H., Wang, T., Lin, Q. and Singhani, A. Transparency Promotion with Model-Agnostic Linear Competitors. *PMLR*, 2021. http://proceedings.mlr.press/v119/rafique20a.html.

[61] Rathi, P., 2020. *A Novel Approach to Feature Importance—Shapley Additive Explanations*. [online] Medium. Available at: <https://towardsdatascience.com/a-novel-approach-to-feature-importance-shapley-additive-explanations-d18af30fc21b> [Accessed 12 September 2021].

[62] Ribeiro, M., Singh, S. and Guestrin, C. "Why Should I Trust You?". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016).

[63] Rudin, C. and Radin, J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *1.2 1*, 2 (2019).
[64] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. and Müller, K. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, Berlin, 2019.

[65] Schneider, S. and Leyer, M., 2019. Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions. *Managerial and Decision Economics*, 40(3), pp.223-231.

[66] Sharma, S., Henderson, J. and Ghosh, J. CERTIFAI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (2020).

[67] Shukla, P. and Irondo, R. Main Types of Neural Networks and Their Applications— Tutorial. *Medium*, 2021. https://pub.towardsai.net/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e.

[68] Smith, A. and Shafer, G., 1976. A Mathematical Theory of Evidence. *Biometrics*, 32(3), p.703.

[69] Solawetz, J. The Train, Validation, Test Split and Why You Need It. *Roboflow Blog*, 2020. https://blog.roboflow.com/train-test-split/#:~:text=The%20train%2C%20validation%2C%20and%20testing%20splits%20are%20built,to%20look%20at%20and%20memorize%20the%20correct%20output.

[70] Stecanella, B. An Introduction to Support Vector Machines (SVM). *Monkeylearn.com*, 2017. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/.

[71] Su, X., Li, L., Shi, F. and Qian, H., 2018. Research on the Fusion of Dependent Evidence Based on Mutual Information. *IEEE Access*, 6, pp.71839-71845.

[72] Szabo, G., Szule, J. and Lins, B. et al. Capturing the real influencing factors of traffic for accurate traffic identification. *2012 IEEE International Conference on Communications (ICC)*, (2012).

[73] Wang, K., Chen, G. and Chen, H., 2018. Understanding Technology Adoption Behavior by Older Adults. *Social Behavior and Personality: an international journal*, 46(5), pp.801-814.

[74] Wang, Y. and Elhag, T. A comparison of neural network, evidential reasoning and multiple regression analysis in modelling bridge risks. *Expert Systems with Applications 32*, 2 (2007), 336-348.

[75] Williams, K. Written Statement: Enhancing digital resilience in education: An action plan to protect children and young people online. *Dera.ioe.ac.uk*, 2021. https://dera.ioe.ac.uk/37096/2/Datganiad%20Ysgrifenedig_%20Gwella%20cadernid%20digidol%20mewn%20addysg_%20Cynllun%20gweithredu%20i%20ddiogelu%20plant%20a%20p.pdf.

[76] Xiao, F., 2020. Generalization of Dempster–Shafer theory: A complex mass function. *Applied Intelligence*, 50(10), pp.3266-3275.

[77] Yang, C., Shi, X., Jie, L. and Han, J. I Know You'll Be Back. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (2018).

[78] Yang, B. and Liu, D. Research on Network Traffic Identification based on Machine Learning and Deep Packet Inspection. *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, (2019).

[79] Yildirim, S. Confusion Matrix — Explained. *Medium*, 2020. https://towardsdatascience.com/confusion-matrix-explained-34e4be19b3ec.

[80] Yu, X. and Guo, H. A Survey on IIoT Security. *2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)*, (2019).

[81] Zerilli, J., Knott, A., Maclaurin, J. and Gavaghan, C. Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?. *Philosophy & Technology 32*, 4 (2018), 661-683.

[82] Zhang, Y., Yang, D. and Liu, Z. et al. An explainable supervised machine learning predictor of acute kidney injury after adult deceased donor liver transplantation. *Journal of Translational Medicine 19*, 1 (2021).