Machine Learning based Predictive Modelling of Cybersecurity Threats Utilising Behavioural Data

Cybersecurity Threat Predictive Modelling

Ting Tin Tin[1], Khiew Jie Xin[2], Ali Aitizaz[3], Lee Kuok Tiung[4], Teoh Chong Keat[5], Hasan Sarwar[6]
Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia[1]
Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia[2]
School of IT, UNITAR International University, Petaling Jaya, Malaysia[3]
Faculty of Social Science and Humanities, Universiti Malaysia Sabah[4]
DigiPen Institute of Technology Singapore[5]
Department of Computer Science and Engineering, United International University, Bangladesh[6]

***Abstract*—With the rapid advancement of technology in Malaysia, the number of cybercrimes is also increasing. To stop the increase in cybercrimes, everyone, including normal citizens, needs to know how secure they are while using digital appliances. A system is developed to predict the risk of users based on their behaviour when they are online using real-life behavioural data obtained from a private university's 207 undergraduates. Five supervised machine learning methods are being tested which are: Regression Logistics, K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayesian Classifier with the aid of a tool, RapidMiner. The algorithms are used to construct, test, and validate three categories of cybercrime threat (Malware, Social Engineering, and Password Attack) predictive models. It was found that KNN model produces the highest accuracy and lowest classification error for all three categories of cybercrime threat. This system is believed to be crucial in alerting users with details of whether the consumer behaviour risk is high or low and what further actions can be taken to increase awareness. This system aims to prevent the rise in cybercrimes by providing a prediction of their risk levels in cybersecurity to encourage them to be more proactive in cybersecurity.**

Keywords*—*Cybersecurity threat; cybersecurity risk; predictive modeling; undergraduates; cybercrime

I. Introduction

Malaysia has entered the digital age, with online meetings and classes or cashless payments becoming more popular [1]. However, as the number of digital users has increased over the years, it may also lead to a surge in cybercrimes. Although most Malaysians have a good level of awareness of cyber threats and risks, only a few who act against it due to a low understanding in cybersecurity and the severity of cyber threats and attacks [2] [3]. This high number of cyberattacks has been estimated to cost the global economy USD 1 trillion in 2020, that is, 50% more than in the previous year [4]. According to researchers, the increase in cybercrimes is also happening in Malaysia [5] [6]. Malaysia's cybersecurity is currently slow to catch up with the pace of advancement, and people lack of knowledge in cybersecurity due to the consequences and impacts of Malaysia's organisation, in the private or public sector [5]. Cyberattacks go beyond the loss of money and reputation but remain a failure in finding a global systematic way to confront [7]. With numerous reports claiming that there is an increase in cybercrimes that are not only targeting important organisations and government but also normal citizens [7][8][9], there are various studies to warn digital users the don'ts and dos without certainly proclaiming how much precaution is needed to be considered safe in cyberspace.

The rise of cybercrimes in Malaysia has caused a lot of damage not only in terms of financial and reputation. However, as a normal citizen without any background knowledge in cybersecurity, it could be difficult for him to know and keep up to date with the latest cybersecurity news and may not even know where to start. One would need to read and listen to stories of victims of cybercrimes and learn from their mistakes to know the risks, but this is not enough because the sources of stories are limited as they were usually from the same social circle. This method of learning may be inaccurate and insufficient, as technology is advancing rapidly and may not be up to date with the latest cybersecurity methods. This concludes that there are no concrete means to prove one's knowledge of cyber risks in the current cybersecurity measures.

There are companies that offer cybersecurity services for companies to predict cyber threats and attacks using artificial intelligence (AI) and machine learning. Having that said, there is no need for normal citizens to hire a company just to know their risk levels in cybersecurity. Thus, the competition is scoped down to simple websites asking visitors a sample of questions to predict their awareness, as it is more non-tech-savvy friendly. These websites, however, do not have official databases and are opinion-based; no research is done in the prediction of results, but rather in a pop quiz-like structure. In general, these websites do not describe risks based on varied behaviours.

Furthermore, very little research was done among Malaysians and some existing research was outdated. Therefore, this study aims to fill this gap by helping researchers with cybersecurity prediction based on user behaviour. Hence, the system would predict the user's risks based on real-life data sets and can give users an idea of which aspect of cyber risk is greater rather than only scores.

Predictive modelling of cybersecurity threats predicts the risks of a user while using a digital device such as a mobile phone, laptop and personal computer by using machine learning algorithms tested and validated by user behaviour data acquired from undergraduates in Malaysia. Therefore, the objective of this study is to identify the factors that affect users’ security awareness (in terms of malware attacks, social engineering, and password attacks); build a predictive model of cybersecurity threats for undergraduates using the Internet in Malaysia; and develop a website to implement the predictive model.

This project has two main parts, data modelling and website deployment. Python is used to programme the machine learning part of the system, coded using Jupyter, with the dataset, while the website serves as a medium for users to predict their risks, who can access the Python files to make predictions. The results will then be displayed on the website. Web pages are structured using HyperText Markup Language(HTML) and the layout is formatted using Cascading Style Sheet(CSS) with JavaScript to give the web pages a final touch to make it more appealing in an integrated development environment (IDE), Visual Studio Code (VSC).

Young adults have a low understanding of the basics of cybersecurity as they are unfamiliar with common cyber threats [3]. The findings suggest that exposing cybersecurity knowledge at a young age can ensure healthy habits online, reducing the chance of cyber attacks and threats [2]. As gender does not affect prediction results because the prediction is entirely based on behaviours, this expands the system’s target of the system to both genders.

This paper is constructed in four sections: Section II about literature review of current research in cybercrime prediction; Section III describes the methodology used in this present study; Section IV presents the study result and discussion; and Section V concludes the present study with limitation and future works.

II. Literature Review

The digital economy dominates Malaysia in business transactions, as more than 40% of these transactions are made digitally. Research found that the future of Malaysia will depend on the digital economy; therefore, the digital space in Malaysia needs to be trusted to allow parties including enterprises, customers, public sectors and individuals to have a reliable digital space [1][10]. According to the Malaysia Computer Emergency Team, there is a visible increase in cyberattacks from January 2022 to July 2022 [11]. One of the recommendations to reduce the risk of cyber threats and attacks is the need to increase public awareness of risks, threats, and vulnerabilities in cyberspace. Therefore, increasing awareness of cyber threats is one of the objectives of the system and is achieved by providing a prediction of user risk in cyberspace using predictive modelling with machine learning techniques.

Since there are various types of cyber threats, three threats, namely malware attack, social engineering, and password attacks, are selected due to the high likelihood of these threats against individuals. Other threats include advanced persistent threats (APT), where attackers gain unauthorised access to a network and try to become a part of the network to prevent detection for an extensive period of time, and Man-in-the-middle attack (MitM), where attacker intercepts users when they are remotely accessing a system over the Internet. APT requires attackers to possess a high knowledge of the victim, and therefore these attacks are usually launched towards nation states, large organisations, companies, or very important people. [12] As the target of this project is young undergraduate Internet users, they are less likely to connect their device remotely online, making MitM not in scope.

There are no equivalent research studies to the proposed project, but similar studies have been found that predict the cyber risk of software [13] [14]. Both projects use machine learning techniques to identify weak points or vulnerabilities in the system and the risk that the software becomes infected or corrupted at a certain time. Zhang et al. [13] built the predictive model with data from the National Vulnerability Database (NVD), which is a public data source for reported software vulnerabilities. They tested the data with various approaches for predictive modelling to find the best techniques for their prediction model, as their study results show that the current approach is not accurate except for a few vendors. Bilge et al. [14] on the other hand, got their data from 18 enterprises for a year, which contains information about binaries appearing on machines with fully and semi-supervised machine learning. Semi-supervised machine learning is a technique that uses machine learning machine learning that uses both supervised, where labelled data is used, and unsupervised, where unlabelled data are used [15].

Other work closely related to cyber risk prediction is cyberattack predictions and cyberattack detection. The prediction is usually overlooked by the research community opposed to cyberattack detection. Ben Fredj explored the prediction of cyberattacks using a deep learning approach [16]. It is a subgroup of the machine learning approach in which multiple layers of neural networks are used to build the model [15], which simulates how neuronal nerves work in a human brain. In addition to that, there is a study that surveyed not only machine learning approaches, but also data mining approaches in terms of prediction and prediction methods used in cybersecurity [17]. Regarding cyberattack detection [18][19][20], most studies use Deep Learning (DL) to model their data to detect which attacks will occur in given situations and the rate of these attacks (Table I).

TABLE I. SUMMARY OF RELATED WORKS

Study	Algorithm	Features/Factors	Reference
Software cyber risk prediction	Supervised Machine Learning	Identifying software vulnerabilities	Zhang et al., 2015; Bilge et al., 2017
Cyberattacks prediction	Deep Learning, Data Mining	Predicting cyberattacks	Ben Fredj et al., 2020; Husák et al., 2018
Cyberattacks detection	Deep Learning	Detecting cyberattacks	Berman et al., 2019; Moustafa et al., 2019; Aldweesh et al., 2020

A. Malware

Malware means malicious software which refers to any software that intrudes on a system developed by cyber-attackers. This software can penetrate the device of a user ranging from viewing to modifying private data, such as user's personal photos, operating systems, and other data that the attacker can find on the victim's device [21]. Malware types include, but are not limited to, viruses, spyware, backdoor, and keyloggers, each with different threats and dangers. Viruses can attach themselves, using macros, to Microsoft Office software such as Words. Therefore, it infects the victim's computer when it is opened or viewed. Students will use Words frequently for various reasons such as completing assignments or recording notes, which pose a high possibility of becoming a victim. Other great possibilities include downloading free software online to avoid purchasing.

Malware is a programme that is inserted into a system with the intention of compromising the confidentiality, integrity, or availability of the victim's data, applications, or operating system, or otherwise annoying or disrupting the victim. Therefore, measuring risk in malware infection can be simplified to the ability to prevent malware from entering the system and the ability to mitigate threats if prevention fails. First, the ability to prevent malware can be measured by how many techniques the user knows about how a malware can enter a system and the depth of understanding of these techniques (MW1). Second, the ability to mitigate threats can be measured by how quickly the user can detect that malware has already entered the system, identified the source of the malware, and remove malware and its techniques (MW2) [22] [23]. Therefore, the system should collect the user response for the following regarding user's behaviour to avoid different malware threats:

M1. Is antivirus software, firewall, and anti-spyware available on the user computers? (MW1, MW2)

M2. What is the user's confidence level of antivirus software in their computers? (MW1, MW2)

M3. How inclined is the user to download materials from unsecure sites? (MW1)

M4. How inclined is the user to download freeware on the Internet? (MW1)

M5. How inclined is the user to scan removable drives before using them on computers? (MW1)

M6. How inclined is the user to apply security patches as soon as possible? (MW2)

M7. Is the user able to sense that something is wrong if the computer runs oddly slow? (MW2)

B. Social Engineering

The art of persuading people to breach information systems is known as social engineering. Instead of launching technical assaults on systems, social engineers use influence and persuasion to persuade people with access to information to reveal secret information or even carry out hostile actions. Most successful attacks on systems are rarely required to find technical vulnerabilities; hacking the human is usually sufficient [24]. Social engineering is the most successful when combined with other methods, such as phishing [25]. Phishing is the act of sending links that link victims to their website that do what cybercriminal programmes to do. For example, attackers send links decorated with official names and formatting to make them appear to come from a legitimate source to play mind tricks and get victims to click on the link. In addition to sending links, attackers can act as an advertiser trying to advertise a product and ask the victim to scan a QR code (quick response) that links to their malicious attack. Both situations are likely to occur amongst anyone.

Social engineering is a method of tricking victims to help compromise their own system. Therefore, user measurement of the risks in social engineering attacks can be simplified into the level of understanding of social engineering techniques and the ability to respond to these techniques correctly. First, how many social engineering techniques can the user know that can be used to measure the level of understanding (SE1). Second, whether the user knows how to respond to these techniques can be used to measure the ability to respond (SE2) [23] [26]. Therefore, the system should collect the user response for the following:

S1. Is the user interested in learning social engineering issues? (SE1, SE2)

S2. Does the user establish a trusted relationship with strangers on-line? (SE1, SE2)

S3. How inclined is the user to click hyperlinks in email messages? (SE1, SE2)

S4. How inclined is the user to check the authorisation of the interlocutor? (SE1)

S5. How inclined is the user to check URL spellings? (SE1)

S6. Does the user trust any benefit winning emails, calls, or SMS? (SE1)

S7. Does the user trust in any information online? (SE1)

S8. Is the user aware of the latest scam and phishing techniques? (SE1)

S9. Does the user feel intimidated by questions by any interlocutor? (SE2)

S10. How inclined is the user to provide details to authorities? (SE2)

S11. How inclined is the user to respond to calls, SMS, or email from strangers? (SE2)

C. *Password Attack*

Password attacks occur when attackers attempt to gain access to a victim's system using the victim's password. This attack is different from the above two threats, as this threat attacks through the 'front door' rather than in secret or stealthily by guessing and trying repetitively until it is correct. User passwords are easy to guess, since they are related to the victim or the password is an actual word or phrase [27], which can be easily obtained using social engineering techniques. As Malaysia is moving toward a digital era, account creation can be common and logging in or signing up requires a password. Other techniques of password attacks include, but are not limited to, brute force, where the attackers try every possible password combination, or dictionary attack, where attacks steal the encrypted data during transmission containing the victim's password and decrypt it using their encryption library.

Password attack is a method to legitimately enter the victim's system through victim passwords. Therefore, to measure the risks of users in password attacks, it can be measured by how securely users keep their passwords and complexity [23] [28]. Therefore, the system should collect the user response for the following:

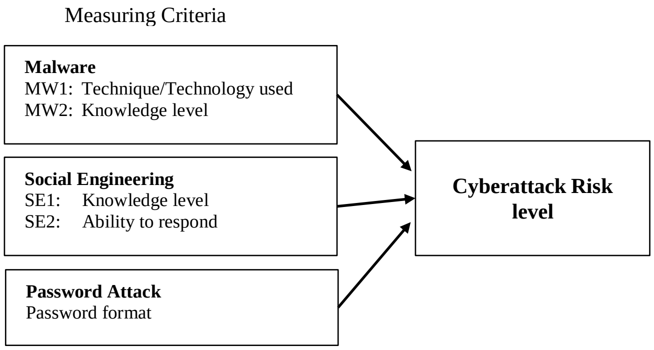
P1. Does the user's password follow a keyboard pattern?
P2. Does the user share passwords with other people?
P3. Does the user create different passwords for different applications?
P4. Is the user's password consisting of lowercase, uppercase, numbers, special characters?
P5. Is the user password longer than 8 characters?
P6. Is the user's password created based on personal/ information?
P7. Does the user change the password?
P8. Does the user use the 'Recall password' option?
P9. Does the user write the password?
P10. Does the user use 'hint' to recover forgotten passwords?
P11. Does the user check for a padlock symbol on browsers?

D. *Conceptual Framework for Measuring Ability to Avoid Cyberattacks*

Based on the literature review, Fig. 1 summarises the measurement criteria of ability to avoid cyberattacks of three categories of threats – malware, social engineering, and password attack.

E. *Web Projects*

Moving from research-based projects to web projects, three web services, namely ProProfs, W3Schools, and the Federal Trade Commission (FTC), are being compared as follows. This predictive modelling is built for young people in Malaysia, which is the scope that is not covered by these three websites.



Measuring Criteria

Malware
MW1: Technique/Technology used
MW2: Knowledge level

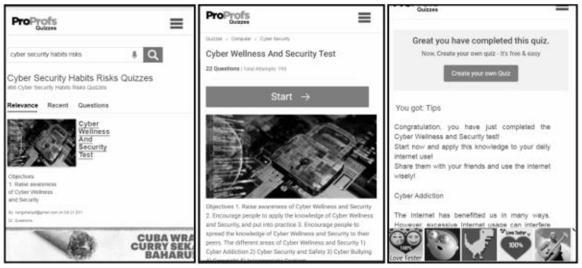
Social Engineering
SE1: Knowledge level
SE2: Ability to respond

Password Attack
Password format

Cyberattack Risk level

Fig. 1. Conceptual framework of measuring criteria for cyberattack risk level.

First, ProProfs is a website that allows any user to create quizzes and post them online on the ProProfs website itself (Fig. 2). Therefore, this website has a variety of quizzes from different domains, which, of course, includes cybersecurity. However, most of the questions of these quizzes are focused on cybersecurity as a course instead of a test for user risks on-line. The questions asked are technical and not suitable for general users who do not consider cybersecurity as their focus. On the ProProfs website, a quiz is found that tests for users' cyber health and security, but it seems to have the same results for all responses entered. Since it is available to everyone, most of these quizzes do not have concrete backing of data to support the claims of the results.



cyber security habits risks

Cyber Security Habits Risks Quizzes

Cyber Wellness And Security Test

Objectives
1. Raise awareness of Cyber Wellness and Security

Cyber Wellness And Security Test

22 Questions

Objectives 1. Raise awareness of Cyber Wellness and Security 2. Encourage people to apply the knowledge of Cyber Wellness and Security, and put into practice 3. Encourage people to spread the knowledge of Cyber Wellness and Security to their peers. The different areas of Cyber Wellness and Security 1) Cyber Addiction 2) Cyber Security and Safety 3) Cyber Bullying

Great you have completed this quiz.
Now, Create your own quiz - It's free & easy

You got: Tips

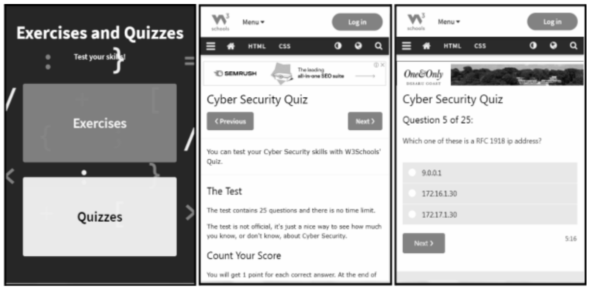
Congratulation, you have just completed the Cyber Wellness and Security test! Start now and apply this knowledge to your daily internet use! Share them with your friends and use the internet wisely!

Cyber Addiction

The Internet has benefitted us in many ways. However, excessive Internet usage can interfere

Fig. 2. Screenshots of the ProProfs website.

Next, W3School is a free educational website to learn Python coding (Fig. 3). However, this website is controlled by two entities namely Refsnes Data and W3schools Network instead of a central point for everyone to submit their viewpoints. As mentioned, this website is built for educational purposes and the cybersecurity quiz is one of the many quizzes found, which is also for people who want to make a revision of cybersecurity courses. Therefore, it does not inform users about the cyber risks that could occur to users.



Cyber Security Quiz

You can test your Cyber Security skills with W3Schools' Quiz.

The Test

The test contains 25 questions and there is no time limit.

The test is not official, it's just a nice way to see how much you know, or don't know, about Cyber Security.

Count Your Score

You will get 1 point for each correct answer. At the end of

Cyber Security Quiz

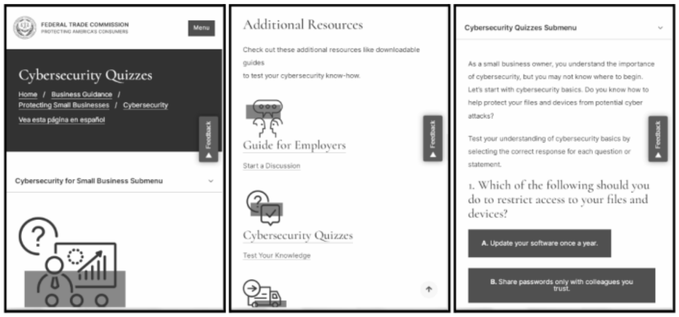
Question 5 of 25:

Which one of these is a RFC 1918 ip address?

○ 9.0.0.1
○ 172.16.1.30
○ 172.17.1.30

Fig. 3. Screenshots of W3Schools' website.

The Federal Trade Commission is an official website of the United States (USA) government that is built to protect American consumers (Fig. 4). It contains cybersecurity quizzes for small businesses to help guide them. The topics in the cybersecurity quizzes are the basics of cybersecurity, physical security, ransomware, phishing, vendor security, and secure remote access. In addition to quizzes, it also provides other means of guidance, such as but not limited to downloadable publications and videos of cybersecurity, which are all accessible in the additional resource's subsection of the page. Table II summarises the three websites in terms of owner, target users, location, and content(s).



FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

Cybersecurity Quizzes

Home / Business Guidance / Protecting Small Businesses / Cybersecurity

Vea esta página en español

Cybersecurity for Small Business Submenu

Additional Resources

Check out these additional resources like downloadable guides
to test your cybersecurity know-how.

Guide for Employers

Start a Discussion

Cybersecurity Quizzes

Test Your Knowledge

Cybersecurity Quizzes Submenu

As a small business owner, you understand the importance of cybersecurity, but you may not know where to begin. Let's start with cybersecurity basics. Do you know how to help protect your files and devices from potential cyber attacks?

Test your understanding of cybersecurity basics by selecting the correct response for each question or statement.

1. Which of the following should you do to restrict access to your files and devices?

A. Update your software once a year.

B. Share passwords only with colleagues you trust.

Fig. 4. Screenshots of FTC's website.

TABLE II. COMPARISON OF 3 WEB SERVICES WITH THE PROPOSED PROJECT

	ProProfs	**W3Schools**	**FTC**
Owner(s) of Content	Anyone	Refsnes Data and W3schools Network	US Government
Target Users	Not specific	Learner	Small business
Location	Not specific	Not specific	US
Content(s)	Quizzes	Quizzes Guidance	Quizzes Guidance

III. METHODOLOGY

Questionnaire items are modified to avoid multiple-choice types of questions. Its purpose is to overcome the limitations of multiple-choice questions, which are the excessive words that make users feel more like an exam and will try to give a 'correct answer' instead of their genuine online behaviour. The data entry designs are shown in Table III.

In this study a private university in Malaysia in the age group of 15 to 30 years constituted the population. The sampling plan implemented in this investigation is the simple random sampling method (SRS). A total of 207 undergraduates participated in the study.

TABLE III. MODIFIED QUESTIONS

ID	**Question**	**Response Type**	**Measured Questions**
	Malware		
L1	Is your device's operating system (OS) up-to-date?	5 likert scale	M6, M2
L2	Do you scan removable drives?		M5
L3	Do you download freeware online?		M3, M4
L4	Do you feel something is wrong if your device is running slow?		M7, M2, M4
L5	Is your device protected by any cybersecurity measures?		M1, M4
	Social Engineering		
E1	Are you interested in learning about social engineering issues?	5 likert scale	S1,S8
E2	Do you establish a trusted relationship with strangers online?	5 likert scale	S2, S11, S8
E3	Do you click on links in emails?	5 likert scale	S3, S7, S8
E4	Do you check the authorisation of the authorities?	5 likert scale	S4, S7, S8
E5	Which link is the right URL to the Google website? www.google.com; google.com; https://google.com; g00gle.com; http://google.com		S5, S8
E6	Do you feel intimidated by questions from any authority?	5 likert scale	S9, S11
E7	Do you provide details to the authorities?	5 likert scale	S10, S8
	Password Attacks		
A1	Create a password that you will use.	Open ended	P1, P4, P5
A2	Is the password created based on personal/ information?	5 likert scale	P6
A3	Do you change your password?		P7
A4	Do you use password management features?		P8, P10
A5	Write the password?		P9
A6	Do you share passwords?		P2
A7	Do you check for a padlock symbol on browsers?		P11
A8	Do you create different passwords for different applications?		P3

Participants responded to the questionnaire based on a 5-point Likert scale, which divides into 5 categories (strongly agree, agree, neither agree, disagree, disagree, strongly disagree). The questions are then analysed to determine whether they are good or bad practises. For questions classified under good practises, the mark is allocated accordingly based on the options (“Strongly agree”-5, “Agree” - 4, “Neutral” - 3, “disagree”- 2, “Strongly disagree” - 1) while for questions classified under bad practises, the mark allocated for each option is the opposite of good practises (“Strongly agree”-1, “Agree” - 2, “Neutral” - 3, “disagree”- 4, “Strongly Disagree” - 5). The responses to every question may vary; however, they generally have the same meaning. The scores for each question for each category are summed up as a total score. Thus, the highest scores attainable on the questionnaire for Malware, Social Engineering, and Password Attack are 25, 35, 40 respectively (best cybersecurity practises implemented), and the lowest scores are 5, 7, 8 respectively (worst cybersecurity practises implemented). Data are then statistically transformed into maximum scores of 35, 55, 55 and lowest 7, 11, 11 respectively.

Questions are either the main question itself, thus not needing to be processed, or are paired with other questions. Questions that are paired with others are calculated using the mean of all questions related to it, except questions A1 and A4. For example, questions L1 and L4 also have value for question M2. Therefore, the value of M2 to be given to the model is the mean value of L1 and L4.

For question A1, the input text will be used to measure 3 parameters.

1) For input text that follows a keyboard pattern, will be marked as low score, while a text that does not will be marked as a high score.

2) The length of the text will determine the score for P4. To achieve the best score (5), users must have an input text of more than 16 characters, while less than 4 characters will be marked as low score(1).

3) The number of character types will determine the score for P5. Text input will be marked as the best score (5) if it contains all types of character (lower case, upper case, numbers, special characters) and the lowest score(one) if it only contains one type of character.

Regarding questions A4, P8 and P10, they point to similar features that most applications provide, which are ‘remember password’ and ‘forget password’. Thus, both scores will be equal. The model will receive the user's behaviour in cyberspace as input to determine its awareness and then predict the user’s risk of cyber threats (Fig. 5).

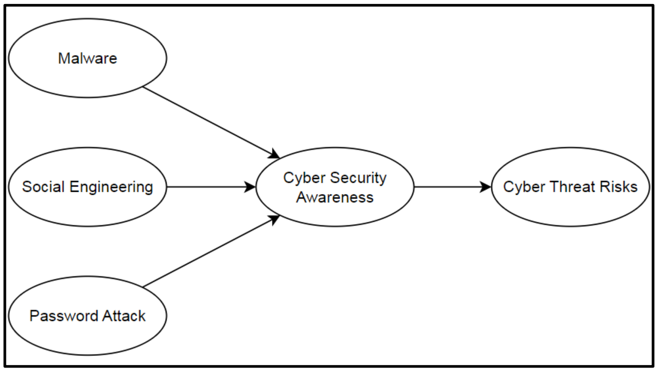


Fig. 5. Predictive Modelling components of Cyber Threats.

IV. Result and Discussion

A. Model Performance

Based on several related works that have been studied, a supervised machine learning method has been selected for the predictive model (Fig. 6). Five supervised machine learning methods are being tested, Regression Logistics, K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayesian Classifier with the aid of a tool, RapidMiner. K-fold cross-validation, where the dataset is divided into 5 groups with each group being the test data set after training the machine with other groups, is used to assess every method above. Of the above five, KNN is selected as the machine learning methodology, as it has the highest accuracy among the other methods (Table IV). KNN is an algorithm that calculates the distance between the new data point and the nearest available data point, where k is a positive integer. The new point is then classified according to which class has the most data points closest to the new data point. The contingency table or confusion matrix is used to help display the accuracy of all the above-mentioned methods. The accuracy is calculated with formula 1 and simplified with formula 2 into percentage (%).

Formula 1:

$$\frac{TruePositives}{TotalPredictedYes} + \frac{TrueNegatives}{TotalPredictedNo} = Accuracy$$

Formula 2:

$$TruePositives + TrueNegatives = Accuracy$$

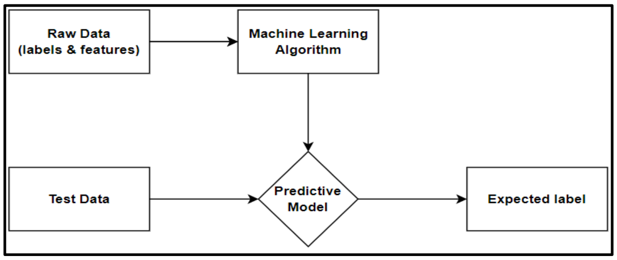


Fig. 6. Supervised machine learning model.

TABLE IV. SUMMARY OF 5 MODELS FOR EACH CATEGORY OF CYBERATTACKS

Model	Malware		Social Engineering		Password Attack		Average Accuracy (%)
	Accuracy (%)	Classification Error (%)	Accuracy (%)	Classification Error (%)	Accuracy (%)	Classification Error (%)	
Naïve Bayesian Classifier	79.5	20.5	89.8	10 2	91.4	8.6	86.9
Regression Logistics	79.5	20.5	88.2	11.8	91.4	8.6	86.4
KNN	92.9	7.1	93.8	6.2	97.6	2.4	94.8
DT	83.0	17	91.5	8.5	79.5	20.5	84.7
SVM	83.0	17	91.5	8.5	81.2	18.8	85.2

B. Model Fit

Python has been selected as the programming language for the machine learning part of the system. Python is selected because it has built-in libraries and frameworks suitable for data science. The libraries used for this project are pandas, NumPy, and Scikit-learn. Pandas library is used to read data tabulated in excel sheets, NumPy is used to process the data into machine learning parameters for the model to train, and Scikit-learn is used to implement machine learning models. A built-in Python module, pickle, is used to save the model as a non-readable binary file to be placed in the server and accessed by the webpage. Two parameters are needed to train the model, the first being the data to test, while the second being the K values.

To get the first parameter, the data is loaded into memory with pandas extracting data from excel sheets. Numpy is then used to convert the data into arrays. These arrays are then divided into training and test data with a ratio of 5: 1 (80% training, 20% testing). Training data will be used to fit the model while test data are used to measure the accuracy of the model.

The next parameter is to find the best K-value for the model. For this, another two-array list is created, namely a set of K values, from 3 to 30, and an empty list to store the results. The model is trained 28 times, and its result score is stored in the empty array list. The least K value with the highest accuracy is then selected as the K value. A lower value of K means that the classification is close to the original value and will not include further away data points, thus increasing precision. K values are evaluated as shown in Fig. 7, 8, and 9. The K values are 4, 5 and 3 for malware attack, social engineering, and password attack model, respectively. After splitting the data set and finding the optimal K values, the model is ready to be saved as a binary file using pickle.

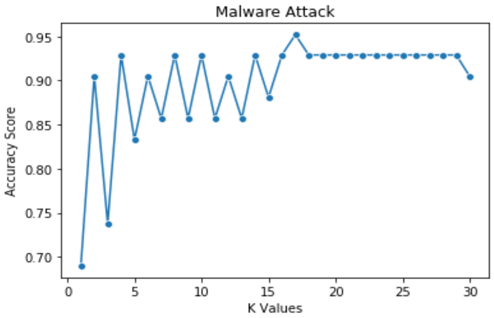


Fig. 7. Accuracy score of K values for malware attacks.

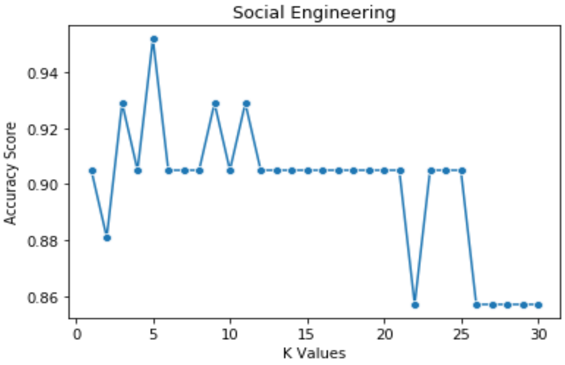


Fig. 8. Accuracy score of K values for social engineering.

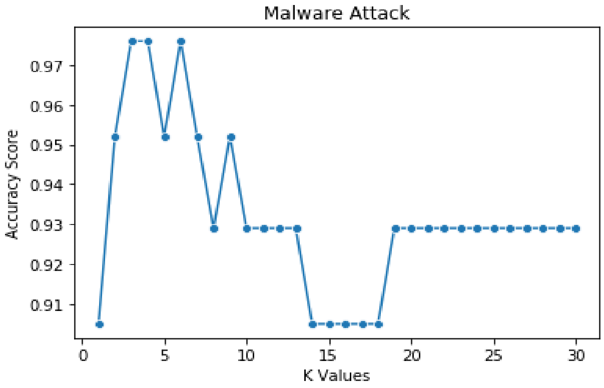


Fig. 9. Accuracy score of K values for malware attacks.

C. Validity and Reliabitliy Consideration

Prior to this actual study, the instrument (questionnaire) was pilot tested with a group of 30 students from the same research site. The researchers ensure that the participants for the pilot test do not participate in the actual study. Data collected from the pilot test were measured for reliability using the Cronbach alpha reliability coefficient, formula 3.

Formula 3:

$$a = \frac{k}{k-1}\left(1 - \frac{\Sigma Vi}{Vt}\right)$$

Where a is the reliability coefficient, k is the number of questions, V_i is the variance of the responses of each question, and V_t is the variance of the total score of each respondent. The reliability measurement of the questionnaire is shown in Table V where all categories show good reliability.

TABLE V. RELIABILITY MEASUREMENT

Cyber Security Categories	Number of Items	Reliability Coefficient
Malware	5	84.3%
Social Engineering	7	81.0%
Password Attack	8	80.2%

D. Demographic Variable Effect

This study is specifically aimed at the demographics of users (age, geography). The variable (gender) is not used in the prediction; however, this variable is still being tested by independent samples T-Test to compare the means of two independent groups (male and female) to prove that this variable does not affect the accuracy of the prediction (Table VI). Data are statistically analysed using the SPSS programme. A total of 207 Malaysians participated in the study, of whom 98 of the respondents are male and 109 of the respondents are female.

TABLE VI. VALUES OF THE T-TEST FOR DIFFERENCES IN THE LEVEL OF CYBER SECURITY BEHAVIOUR BY GENDER IN THE ASPECT OF MALWARE, SOCIAL ENGINEERING, AND PASSWORD ATTACK

	Group statistics		T-Test	
	Mean	Std Deviation	t-value	Sig.
Malware Male Female	33.60 33.57	5.49 4.76	0.05	0.963
Social engineering Male Female	34.28 34.99	5.27 4.76	-1.02	0.306
Password Attack Male Female	33.46 33.25	5.79 5.96	0.26	0.796

Based on the three tables above, there are no significant gender differences in the level of cybersecurity behaviour in all three aspects (malware, social engineering, password attack); thus, gender does not affect the accuracy of the prediction of the model.

E. Comparing Proposed Website with Existing Websites

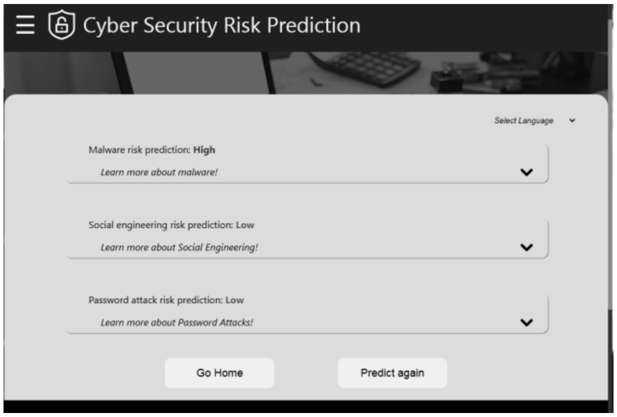
1) Similar existing website: Three previously mentioned websites are studied for their functionalities to choose those that are applicable for this project. All of them follow the same flow, that is, to introduce what and how important cybersecurity is and a navigation panel or menu which links to other functionalities.

The Proprofs website allows users to view the list of questions of the selected quiz, contact the author of the quiz, take the quiz, edit the settings of the webpage, to search for other quizzes from any domain, to share the selected quiz, in embedding the quiz to another website. The Proprofs website also allows users to create quizzes with the precondition of having an account with Proprofs, thus needing users to log in prior to creating quizzes [29].

The W3schools website allows users to view an introduction of cybersecurity, search for other services that w3school provides, log into a w3school account, take the quiz, quick link to access other tutorials, change the theme of the website, translate the website to another language. The W3schools website also allows users to subscribe to their services under the condition that the user has an account with w3schools, which requires that the users log in [30].

The FTC website allows users to translate the website into another language, report fraud, sign up for FTC newsletters, search for other documents in a legal library, give feedback, view Introduction to Cybersecurity, print the website, take the risk assessment, and access to other services [31].

Based on the study above, all websites have similar design and functions; therefore, to be consistent with the existing websites, the Predictive Modelling of Cyber Security Threats website should display an introduction to cybersecurity and explanation of its importance (Fig. 10). The navigation menu should also be added to this website for easy navigation between other web pages, which includes displaying the information page and the methodology page. Users can also choose to share this website. Feedback from user functions should also be included so that this website can interact with users for future improvements. Lastly, the website should allow users to assess their cyber risks. Other functions such as searching, log-in or log-out, and printing are omitted in this website, as they serve no purpose for their functions on this website. For example, this website does not need a search function, as this project has only cyber risk prediction as its focus. The comparison is summarised in Table VII.



Malware risk prediction: **High**

Learn more about malware!

Social engineering risk prediction: Low

Learn more about Social Engineering!

Password attack risk prediction: Low

Learn more about Password Attacks!

Fig. 10. Screenshot of the project website prediction result page.

TABLE VII. FEATURES OF THE ABOVE 3 WEBSITES AND PROPOSED SYSTEM

Features	Proprofs	w3schools	FTC website	Proposed website
Login / Logout	✓	✓		✓
Subscription	✓	✓		✓
Display cyber info	✓	✓	✓	✓
Quiz / Assessment	✓	✓	✓	✓
Share website	✓			✓
Guides		✓	✓	✓
Webpage translation			✓	✓
Feedback			✓	✓
Display methodology				✓
Machine learning				✓
FAQ				✓

V. CONCLUSION

This project uses five machine learning algorithms (Regression Logistics, K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayesian Classifier) to predict the risk of cyber threats in the aspects of malware attack, social engineering, and password attacks among Internet users based on their online behaviour. During the development of this present study, it was also found that gender does not play a role in the perception of cybersecurity in Malaysia. KNN predictive model produced the highest accuracy and the lowest classification error. Therefore, KNN model is further improved using Python.

Given the absence of previous studies utilizing machine learning techniques for predicting users' cyberattack risk levels, this present study introduces a conceptual framework that includes measurement criteria for assessing risk levels. Most of the previous studies are predicting the cyberattacks of organisation websites or companies' networks instead of individual risk level. This study serves as guidance for future researchers to continue the study in other cyberattacks such as MitM. New behaviours can also be incorporated to investigate cyber risks. Furthermore, this present study only focused on a data set of young people, since all participants in this project were in the age group of 15 to 30. More efforts are needed in this domain, as predicting human behaviour is a complex task [10]. Techniques to detect potential cyberattacks are crucial to ensure a safe world of the Internet for global users.

REFERENCES

[1] Mat, B., Pero, S., Wahid, R., and Sule, B., 2019. Cybersecurity and the digital economy in Malaysia: trusted law for customer and enterprise protection. International Journal of Innovative Technology and Exploring Engineering, 8(3), pp.214-220.

[2] Zulkifli, Z., Molok, N.N.A., Abd Rahim, N.H. and Talib, S., 2020. Cyber security awareness among secondary school students in Malaysia. Journal of Information Systems and Digital Technologies, 2(2), pp.28-41.

[3] Fatokun, F.B., Hamid, S., Norman, A. and Fatokun, J.O., 2019. The impact of age, gender, and educational level on the cybersecurity behaviors of tertiary institution students: an empirical investigation on Malaysian universities. Journal of Physics: Conference Series, 1339(1), p. 012098. https://doi:10.1088/1742-6596/1339/1/012098

[4] Cremer, F., Sheehan, B., Fortmann, M., Kia, A.N., Mullins, M., Murphy, F. and Materne, S., 2022. Cyber risk and cybersecurity: a systematic review of data availability. The Geneva Papers on Risk and Insurance-Issues and Practice, pp.1-39. https:// doi.org/10.1057/s41288-022-00266-6

[5] Teoh, C.S., Mahmood, A.K. and Dzazali, S., 2018. Cyber security challenges in organizations: a case study in Malaysia. 2018 4th International Conference on Computer and Information Sciences, pp. 1-6.

[6] Abdullah, F., Mohamad, N.S. and Yunos, Z., 2018. Safeguarding Malaysia's cyberspace against cyber threats: contributions by cybersecurity Malaysia. OIC-CERT Journal of Cyber Security, 1(1), pp.22-31.

[7] Singh, M.M., Frank, R. and Zainon, W.M.N.W., 2021. Cyber-criminology defense in pervasive environment: a study of cybercrimes in Malaysia. Bulletin of Electrical Engineering and Informatics, 10(3), pp.1658-1668.

[8] Khan, S., Khan, N. and Tan, O., 2020. Efficiency of legal and regulatory framework in combating cybercrime in Malaysia. In Understanding Digital Industry, pp. 333-336. Routledge.

[9] Isa, M.Y.B.M., Ibrahim, W.N.B.W. and Mohamed, Z., 2021. The relationship between financial literacy and public awareness on combating the threat of cybercrime in Malaysia. The Journal of Industrial Distribution & Business, 12(12), pp.1-10.

[10] Sulaiman, N. S., Fauzi, M. A., Hussain, S., & Wider, W., 2022. Cybersecurity behavior among government employees: The role of protection motivation theory and responsibility in mitigating cyberattacks. Information, 13(9), 413. MDPI AG. http://dx.doi.org/10.3390/info13090413

[11] MyCERT, 2022. MyCERT Incident Report 2022.

[12] Cassetto, O., 2023. Cybersecurity threats: Types and challenges, Exabeam. Available at: https://www.exabeam.com/information-security/cyber-security-threat/

[13] Zhang, S., Ou, X. and Caragea, D., 2015. Predicting cyber risks through national vulnerability database. Information Security Journal: A Global Perspective, 24(4-6), pp.194-206. https://doi.org/10.1080/19393555.2015.1111961

[14] Bilge, L., Han, Y. and Dell'Amico, M., 2017. Riskteller: Predicting the risk of cyber incidents. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp. 1299-1311. https://doi.org/10.1145/3133956.3134022

[15] Zhou, Z.H., 2021. Machine learning. Springer Nature.

[16] Ben Fredj, O., Mihoub, A., Krichen, M., Cheikhrouhou, O. and Derhab, A., 2020. CyberSecurity attack prediction: a deep learning approach. 13th International Conference on Security of Information and Networks, pp. 1-6. https://doi.org/10.1145/3433174.3433614

[17] Husák, M., Komárková, J., Bou-Harb, E. and Čeleda, P., 2018. Survey of attack projection, prediction, and forecasting in cyber security. IEEE Communications Surveys & Tutorials, 21(1), pp.640-660. https://doi.org/ 10.1109/COMST.2018.2871866

[18] Berman, D.S., Buczak, A.L., Chavis, J.S. and Corbett, C.L., 2019. A survey of deep learning methods for cyber security. Information, 10(4), p.122. https://doi.org/10.3390/info10040122

[19] Moustafa, N., Hu, J. and Slay, J., 2019. A holistic review of network anomaly detection systems: A comprehensive survey. Journal of Network and Computer Applications, 128, pp.33-55. https://doi.org/10.1016/j.jnca.2018.12.006

[20] Aldweesh, A., Derhab, A. and Emam, A.Z., 2020. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems, 189, p.105124.

[21] Lutkevich, B., 2022. What is malware? definition, types, prevention - techtarget, Security. TechTarget. Available at: https://www.techtarget.com/searchsecurity/definition/malware

[22] Martens, M., De Wolf, R. and De Marez, L., 2019. Investigating and comparing the predictors of the intention towards taking security measures against malware, scams and cybercrime in general. Computers in Human Behavior, 92, pp.139-150. https://doi.org/10.1016/j.chb.2018.11.002

[23] Muniandy, L., Muniandy, B. and Samsudin, Z., 2017. Cyber security behaviour among higher education students in Malaysia. J. Inf. Assur. Cyber Secur, 2017, pp.1-13.

[24] Mann, I., 2017. Hacking the human: social engineering techniques and security countermeasures. Routledge.

[25] Abass, I.A.M., 2018. Social engineering threat and defense: a literature survey. Journal of Information Security, 9(04), p.257.

[26] Albladi, S.M. and Weir, G.R., 2020. Predicting individuals' vulnerability to social engineering in social networks. Cybersecurity, 3(1), pp.1-19. https://doi.org/10.1186/s42400-020-00047-5

[27] Tasevski, P. and Eurecom, F., 2015. Methodological approach to security awareness program. In CyberSecurity for the Next Generation Conference.

[28] Ye, B., Guo, Y., Zhang, L. and Guo, X., 2019. An empirical study of mnemonic password creation tips. Computers & Security, 85, pp.41-50. https://doi.org/10.1016/j.cose.2019.04.009

[29] ProProfs.com. Available at: https://www.proprofs.com/

[30] W3Schools.com. Available at: https://www.w3schools.com/

[31] Federal Trade Commision. Available at: https://www.ftc.gov/