SISTEM PENGESAHAN BERASASKAN LAMAN WEB BAGI MENGESAN SERANGAN PANCINGAN DATA

NUR HAMIMI BINTI MOHD RATHI

UNIVERSITI KEBANGSAAN MALAYSIA

SISTEM PENGESAHAN BERASASKAN LAMAN WEB BAGI MENGESAN SERANGAN PANCINGAN DATA

NUR HAMIMI BINTI MOHD RATHI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA KESELAMATAN SIBER

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2023

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

27 Mac 2023

NUR HAMIMI MOHD RATHI
P109215

PENGHARGAAN

Alhamdulillah, dengan izin Allah SWT, saya telah berjaya menyelesaikan tesis saya dengan baik. Segala puji bagi Allah S.W.T., tuhan sekalian alam, yang telah memberikan saya kekuatan dan kemampuan untuk menyempurnakan tesis ini.

Saya juga ingin mengucapkan terima kasih kepada penyelia projek saya iaitu Dr Ahmad Tarmizi Abdul Ghani yang telah mencurahkan ilmu, tunjuk ajar dan bimbingan yang sangat berharga selama proses penulisan tesis ini. Terima kasih atas kesabaran dan sokongan yang diberikan.

Saya juga ingin berterima kasih kepada pihak Fakulti Teknologi dan Sains Maklumat yang telah memberikan kesempatan kepada saya untuk mengejar pendidikan tinggi dan menyelesaikan tesis ini. Saya sangat berterima kasih atas sumber daya dan kemudahan akademik yang telah disediakan oleh pihak UKM.

Buat ibu bapa tercinta, ibunda Hajah Hamidah binti Ahmad dan ayahanda Mohd Rathi bin Ismail, terima kasih atas segala pengorbanan dan kasih sayang yang diberikan. Buat suami yang dikasihi, Mohammad Azill Azmi Sabni, dan anakanda tersayang Muhammad Adam, serta adik beradik dan seluruh keluarga tercinta. Terima kasih atas pengorbanan, kesabaran dan sokongan yang tidak putus selama proses penulisan tesis ini. Semoga hasil dari tesis ini dapat memberikan manfaat bagi masyarakat dan dapat digunakan sebagai rujukan yang baik dalam penelitian selanjutnya. Sekali lagi, terima kasih atas semua sokongan dan motivasi yang diberikan.

ABSTRAK

Serangan pancingan data, atau lebih dikenali sebagai *phishing* adalah sejenis serangan siber yang digunakan untuk menipu pengguna untuk memberikan informasi sensitif seperti kata laluan atau maklumat peribadi atau perbankan. Serangan ini biasanya dilakukan melalui emel atau mesej yang mengelirukan yang mengarahkan pengguna ke laman web yang palsu, dimana laman web palsu ini dibuat untuk menyerupai laman web asli bagi mengelirukan pengguna untuk memberikan data peribadi atau melakukan transaksi yang tidak sepatutnya. Oleh itu, adalah amat penting untuk berhati-hati dan mengenali ciri-ciri laman web palsu untuk mengelakkan jatuh ke dalam serangan *phishing*. Ini kerana, serangan ini boleh memberi kesan yang besar kepada pelbagai industri, terutamanya sektor perbankan, perniagaan dalam talian dan media sosial kerana ia boleh menyebabkan kerugian kewangan yang besar, kehilangan rahsia perniagaan, malahan juga boleh menjatuhkan reputasi sesebuah syarikat. Sesebuah entiti yang diserang juga boleh dikenakan denda jika terdapat pelanggaran privasi akibat dari serangan *phishing*. Kajian ini bertujuan untuk menganalisa keadaan semasa serangan pancingan data dan membangunkan sistem pengesahan berasaskan laman web yang boleh mengesan serangan pancingan data dengan tepat. Melalui tinjauan kesusasteraan, kajian ini mengungkai teknik dan taktik yang digunakan oleh penggodam, serta kaedah dan teknologi yang digunakan untuk mengesan dan mencegah serangan *phishing*. Kajian ini juga meninjau faktor-faktor yang menyebabkan serangan ini terjadi, dan mencadangkan satu sistem *anti-phishing* yang boleh mengesan laman web palsu Kemudian, keberkesanan dan ketepatan sistem pengesahan ini akan dinilai dan dianalisis. Sistem ini dibangunkan berdasarkan kaedah pembelajaran mesin atau machine learning, melalui algoritma Pokok Keputusan atau lebih dikenali sebagai Decision Tree. Ciri-ciri dari URL telah dipilih dan dimasukkan ke dalam sistem anti-*phishing*, dan seterusnya sistem ini akan mengesan sekiranya sesebuah laman web tersebut adalah asli atau sebaliknya. Sebanyak 100 URL digunakan sebagai sampel data bagi set ujian terhadap sistem ini, dah hasil ujian mendapati bahawa sistem ini mampu mengesan sebanyak 97% URL pancingan data. Hasil daripada kajian ini boleh digunakan untuk menambah pembangunan strategi yang lebih berkesan untuk mengesan dan mencegah serangan *phishing*, serta untuk mendidik individu dan organisasi tentang risiko dan cara untuk melindungi diri.

WEB BASED VERIFICATION SYSTEM FOR PHISHING DETECTION

ABSTRACT

Phishing attacks are a type of cyber attack used to deceive users into providing sensitive information such as passwords or personal or banking information. These attacks are usually carried out through misleading emails or messages that direct users to a fake website, which is made to look like a real website to trick users into providing personal data or conducting unauthorized transactions. Therefore, it is very important to be cautious and recognize the characteristics of fake websites to avoid falling into phishing attacks. These attacks can have a significant impact on various industries, especially the banking, online business, and social media sectors, as they can cause significant financial losses, loss of trade secrets, and even damage a company's reputation. An entity that is attacked can also be fined if there is a privacy violation resulting from a phishing attack. This study aims to analyze the current state of phishing attacks and develop a web-based verification system that can accurately detect phishing attacks. Through a literature review, this study reveals the techniques and tactics used by hackers, as well as the methods and technologies used to detect and prevent phishing attacks. The study also examines the factors that cause these attacks to occur and proposes an anti-phishing system that can detect fake websites. The effectiveness and accuracy of this verification system will then be evaluated and analyzed. This system is developed using machine learning methods, specifically the Decision Tree algorithm. URL characteristics have been selected and included in the anti-phishing system, which will then detect whether a website is real or fake. A total of 100 URLs were used as a sample data for testing the system, and the test results found that the system was able to detect 97% of phishing URLs. The results of this study can be used to enhance the development of more effective strategies to detect and prevent phishing attacks, as well as to educate individuals and organizations about the risks and ways to protect themselves.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

BAB I

PENGENALAN

1.1 LATAR BELAKANG KAJIAN

Zaman kini, kebanyakan institusi perniagaan amat mementingkan pengurusan risiko di dalam semua aktiviti kerja harian. Kebanyakan syarikat pada masa kini mengendalikan perniagaan mereka dengan menggunakan platform digital, termasuk aktiviti yang melibatkan pengurusan kewangan, jualan, pinjaman dan transaksi, dan kesemua aktiviti ini membuatkan sektor perniagaan menjadi salah satu pekerjaan yang paling berisiko jika di bandingkan dengan sektor-sektor lain di dalam industri yang sama. Laman e-dagang, atau lebih dikenali sebagai e-commerce, tumbuh bagaikan cendawan selepas hujan. Lazada, Shopee, Mudah.my, antara beberapa nama terkemuka yang menjalankan ativiti perniagaan dalam talian yang mempunyai jutaan pengikut dan pelanggan. Setiap entiti perniagaan menyimpan segala maklumat dan data secara sulit, merangkumi data-data transaksi serta maklumat peribadi pelanggan. Perkara ini telah memotivasi pihak-pihak yang tidak bertanggungjawab untuk cuba menggodam atau melakukan serangan dan 'rompakan' secara maya.

Kesemua laman e-dagang ini dipautkan ke laman perbankan internet untuk tujuan pembayaran, dan kaedah pembayaran atas talian ini membuka peluang yang besar kepada para penggodam untuk mencuti data dan duit pelanggan. Pelbagai kaedah lazimnya digunapakai oleh penjenayah siber bagi mendapatkan maklumat sulit dan data-data penting daripada pihak bank mahupun pemegang akaun, contohnya melalui serangan pancingan data, atau lebih dikenali sebagai "*phishing*". Serangan ini berlaku apabila seseorang cuba mendapatkan maklumat berharga dengan cara memperdaya mangsa-mangsa untuk berkongsi maklumat sensitif. Kebiasaannya, ia dilakukan melalui beberapa saluran, seperti e-mel, iklan, atau laman web yang menyerupai

antaramuka laman-laman web terkenal. Sekali imbas, laman web tersebut kelihatan sama seperti laman web rasmi milik syarikat terkemuka, namun hakikatnya, ia tidak lebih dari sekadar medium untuk pencurian data. Data-data yang disasarkan adalah maklumat peribadi yang bersifat sensitif, contohnya seperti kata laluan, log masuk dan nombor akaun bank. Penggodam kemudiannya memanfaatkan data-data yang dicuri ini dengan beberapa cara, antaranya dengan menjual akaun mangsa kepada pihak ketiga, dan adakalanya ia digunakan utuk kepentingan peribadi.

Dalam kebanyakan kes, mangsa lazimnya melaporkan mengenai kehilangan wang dari dalam akaun, dan terdapat juga kes dimana identiti mangsa juga diexploitasi dan disalahguna. Ketika masih di era awal 2000-an, serangan pancingan data ini lebih mudah dikenalpasti, iaitu dengan kesalahan ejaan yang jelas, bahkan segelintir penggodam menggunakan terjemahan langsung dari Terjemahan Google. Namun kini, dengan kecanggihan pelbagai aplikasi dan perisian komputer, taktik serangan pancingan data telah menjadi lebih terperinci dan tersusun, sekaligus menyukarkan para pengguna untuk membezakan antara laman web dan e-mel yang palsu dengan yang asal.

Kini, serangan pancingan data telah semakin tertumpu kepada aplikasi dan laman web yang menawarkan perkhidmatan pembayaran dalam talian dan juga perbankan. Tidak dinafikan, kemunculan internet perbankan dan e-dagang telah memberi impak yang baik terhadap bank dan juga entitu perniagaan di seluruh dunia, dimana mereka mampu menawarkan pelanggan mereka dengan aktiviti perbankan dan pembayaran yang lebih mudah dan fleksibel. Terdapat pelbagai ciri dan fungsi didatangkan sekali dengan perkhidmatan ini, termasuk aktiviti pemindahan dana, penguruskan akaun cek, pembayaran belian atas talian dan juga pelunasan bil. Selain itu, e-perbankan membolehkan pelanggan untuk mengakses akaun bank mereka melalui bank laman web tanpa perlu pergi ke cawangan bank masing-masing. Ini bermaksud, e-perbankan telah memberi akses pantas kepada pelbagai aktiviti kewangan, seperti sebagai pemindahan wang, pembayaran bil utiliti dan semakan pengurusan akaun. Dapat kita lihat di sini bahawa perbankan internet dan perkhidmatan e-dagang telah memberi manfaat kepada kedua-dua pihak, iaitu institusi yang terlibat dan juga pelanggan. Pihak bank dan entiti bisnes juga mendapat manfaat kerana perkhidmatan perbankan internet telah mengurangkan kos operasi mereka dari segi pengurangan

kemudahan fizikal yang melibatkan sumber manusia, kertas kerja, dan kakitangan sokongan.

Kepelbagaian aktiviti kewangan yang ditawarkan oleh perkhidmatan ini telah menarik minat penggodam untuk menjalankan aktiviti pancingan data mereka, kerana jelas, pulangan yang akan diperoleh lebih lumayan daripada menggodam laman-laman web yang lain. Pancingan data adalah tindakan penipuan untuk mengakses dan memindahkan dana daripada akaun perbankan dalam talian seseorang untuk tujuan keuntungan kewangan. Mereka merasakan bahawa mencuri data melalui laman e-perbankan dan e-dagang lebih berbaloi kerana kesemua aktiviti yang berlaku di dalam laman web ini berkaitan dengan wang. Menurut Alhuseen O. Alsayed dan Anwar L. Bilgrami (2017), banyak kajian penyelidikan membuktikan bahawa banyak keselamatan isu, seperti serangan pancingan data, telah digunakan oleh penggodam untuk menceroboh akaun pelanggan e-perbankan.

Kini, semua institusi perniagaan dan kewangan mengambil serius mengenai sebarang kegiatan penipuan ini dan secara aktif melaksanakan perbagai langkah keselamatan untuk melindungi pelanggan dari sebarang kejadian yang tidak diingini. Dari segi pancingan data, institusi-institusi ini haruslah melindungi data-data sulit pelanggan mereka daripada jatuh ke tangan pihak yang tidak bertanggungjawab. Malangnya, kekurangan perlindungan keselamatan dari kebanyakan laman web perniagaan dalam talian adalah kondusif untuk serangan pancingan data.

Selain dari e-dagang dan e-perbankan, terdapat satu lagi laman yang sering dijadikan sasaran oleh para penggodam, iaitu laman-laman media sosial. Facebook, Twitter dan Instagram tidak terkecuali, dimana platform-platform ini mengadaptasi sistem pembayaran dalam talian di dalam setiap applikasi mereka. Menurut kajian terbaru oleh Verizon, 33% daripada semua serangan pancingan data dilakukan melalui media sosial. Seperti yang diketahui umum, media sosial bersifat global dan dimuat turun oleh hampir semua orang. Keadaan ini merupakan peluang yang besar untuk para penggodam mengumpulkan maklumat mengenai sasaran mereka sebelum kemudiannya melancarkan taktik pancingan terhadap mangsa-mangsa mereka. Sebaik sahaja penggodam mempelajari segala yang mereka perlu daripada profil media sosial sasaran

mereka, maklumat tersebut biasanya akan digunakan untuk memanipulasi mangsa supaya menyerahkan lebih banyak data atau wang. Untuk tujuan ini, penjenayah siber membuat iklan yang disasarkan dan e-mel pancingan data yang mengandungi perisian hasad bagi mencuri maklumat peribadi mangsa.

Menariknya, melalui media sosial, teknik serangan ini bukan sahaja tertumpu kepada sistem pembayaran atas talian dan profil mangsa, bahkan ia telah pergi lebih jauh apabila mereka melakukan pancingan data terhadap banyak aplikasi lain dalam media sosial, seperti kuiz dan permainan di Facebook, iklan-iklan tawaran pekerjaan, dan juga tawaran khidmat pelanggan palsu.

1.2 PERNYATAAN MASALAH

Kajian oleh Nilay Yildirim dan Asaf Varol (2019) menunjukkan bahawa kadar pengguna dalam talian yang menggunakan telefon bimbit sekurang-kurangnya sekali seminggu untuk tujuan servis perbankan telah meningkat sekurang-kurangnya 5%, dan jumlah pengguna perbankan mudah alih secara global telah meningkat daripada 39% pada 2017 kepada 42% pada 2018. Ini jelas membuktikan bahawa kebergantungan pengguna terhadap servis kewangan dalam talian semakin bertambah, dan seiring dengan penambahan ini, kadar ancaman siber terhadap laman ini juga turut meningkat.

Kerugian yang disebabkan oleh insiden kebocoran data yang melibatkan sektor kewangan telah menjadi topik perdebatan hangat kerana banyak firma perbankan tidak pasti jika sistem mereka telah terjejas. Oleh itu, kebanyakan syarikat perniagaan akan membuat pelaburan yang tinggi dalam pembelian produk dan melanggan perkhidmatan keselamatan dari entity dan pihak ketiga untuk melindungi syarikat, data dan pekerja mereka daripada menjadi mangsa serangan siber. Cybersecurity Ventures, firma penyelidikan, menganggarkan bahawa kos perbelanjaan global untuk produk dan perkhidmatan keselamatan siber akan melebihi USD 1 trilion secara kumulatif antara 2017 dan 2021. Kos yang dilaburkan ini adalah kesan daripada kerosakan sistem, kecurian wang dan data, pemulihan system maklumat dan pelan tindakan pemulihan. Sekali imbas, insiden ini memerlukan kos yang tinggi, tetapi apabila kita melihat insiden serangan siber yang telah dilaporkan dan disiasat, sebenarnya penjenayah siber

inilah yang mendedahkan kesemua kelemahan yang wujud dalam institusi-institusi perniagaan yang menyebabkan semua insiden keselamatan siber terjadi.

Teknik yang paling banyak digunapakai oleh para penggodam adalah tentunya pancingan data. Kecanggihan teknologi masa kini memudahkan penggodam untuk merekabentuk laman web palsu yang kelihatan tidak ubah seperti web tulen. Terdapat banyak kes yang dilaporkan dimana penyerang menyamar sebagai pihak bank dan staf khidmat pelanggan dari syarikat-syarikat e-dagang atau bank bagi mendapatkan maklumat kepada soalan keselamatan yang telah ditetapkan oleh pengguna di dalam profil laman e-perbankan dan e-dagang mereka (R. Kiruthiga dan D. Akila, 2019). Ini dibuktikan oleh laporan dari Phishlabs 2019 yang menyatakan bahawa pancingan data meningkat sebanyak 40.9% dalam tahun 2018 dengan 83.9% serangan menyasarkan sistem perkhidmatan kewangan, e-mel, dan system pembayaran atas talian. Menurut kepada laporan itu lagi, jumlah laman pancingan data meningkat secara berterusan pada suku pertama bagi tahun 2018 dan kekal tinggi sepanjang suku kedua dan ketiga. (Suleiman Y. Yerima dan Mohammed K. Alzaylaee, 2020).

Berdasarkan kajian yang dilakukan oleh L. Karthika dan V. Perumal (2016), motif utama di sebalik serangan pancingan data ini, dari sudut pandangan penyerang mungkin boleh dikelaskan kepada tiga kategori :

1. Keuntungan kewangan

   Segala data dan maklumat sulit yang dicuri dari rangkaian e-dagang dan internet perbankan lazimnya dijual bagi mendapatkan pulangan wang yang lumayan.

2. Penyembunyian identiti

   Penggodam mungkin menceroboh dan menggunakan maklumat mangsa yang dicuri bagi menyembunyikan identiti mereka, yang kemungkinan besar adalah penjenayah siber yang dikehendaki.

3. Kemasyhuran dan nama

   Penyerang mungkin mengenali mangsa mereka melalui rakan sebaya atau komuniti sekeliling mereka. Menggodam profil e-perbankan mangsa yang dikenali mungkin memberi kepuasan kepada penyerang, dan merasakan bahawa

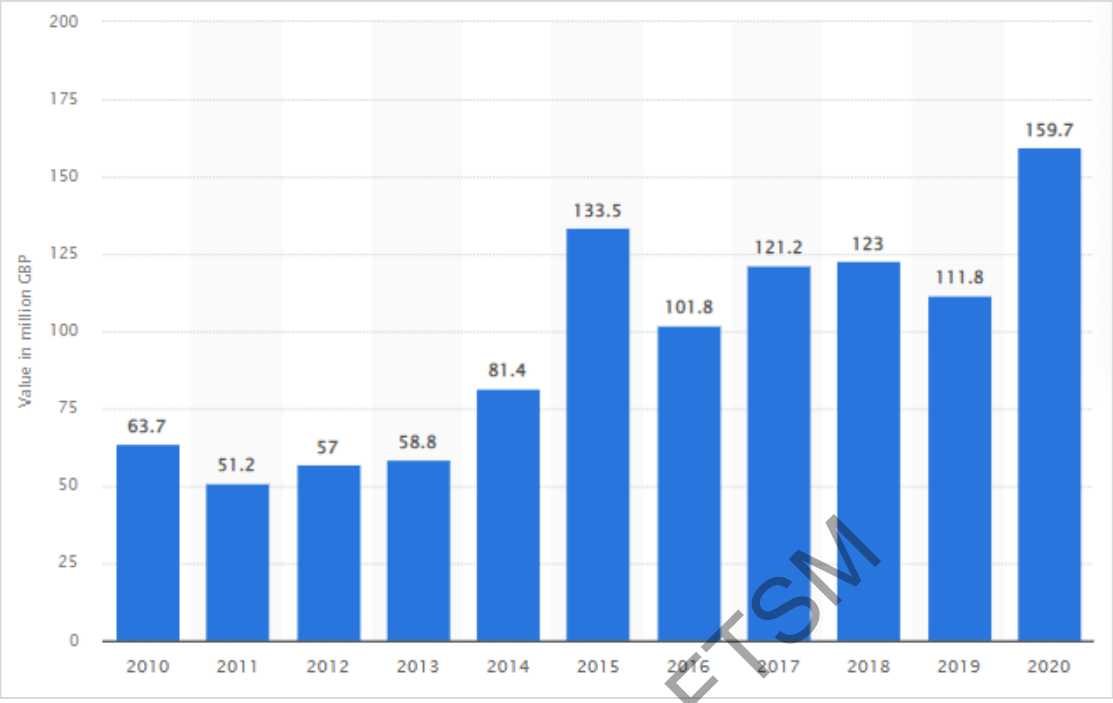
kemahiran menggodam mereka di iktiraf untuk terus menjalani jenayah dan aktiviti haram.

Penjenayah siber mungkin mempunyai akses kepada pelbagai jenis taktik dan pendekatan teknologi yang boleh digunakan untuk mencipta serangan pancingan data yang direka dengan baik. Ini termasuk memasang perisian hasad pautan di laman web palsu (M. A. Adebowale et. al, 2019).

Pelbagai kajian telah dilakukan secara berterusan untuk menghalang serangan pancingan data oleh komuniti yang berbeza di seluruh dunia. Serangan ini boleh dicegah dengan mengesan sumber laman web palsu, dan mewujudkan kesedaran kepada pengguna untuk mengenal pasti keaslian web yang tulen. Dalam usaha membendung ancaman ini, algoritma pembelajaran mesin telah menjadi salah satu teknik yang ampuh dalam mengesan laman web pancingan data (R. Kiruthiga dan D. Akila, 2019).

1.2.1 Statistik pancingan data

Sejak kewujudan laman-laman perniagaan dalam talian, statistik bagi serangan pancingan data menjadi semakin meningkat di seluruh dunia (Ammara Zamir et. al, 2020). Bagi menyokong dakwaan ini, terdapat banyak statistik dan laporan mengenai insiden kebocoran data yang berlaku kepada pelbagai organisasi dalam sektor perniagaan dan kewangan di seluruh dunia. Salah satunya adalah dari statistik yang dikeluarkan oleh Statista, yang menyatakan bahawa nilai kerugian penipuan perbankan dalam talian tahunan di United Kingdom (UK) telah meningkat dalam tempoh 10 tahun, dari 2010 hingga 2020 (dalam juta GBP), seperti yang ditunjukkan dalam Rajah 1.1 di bawah.

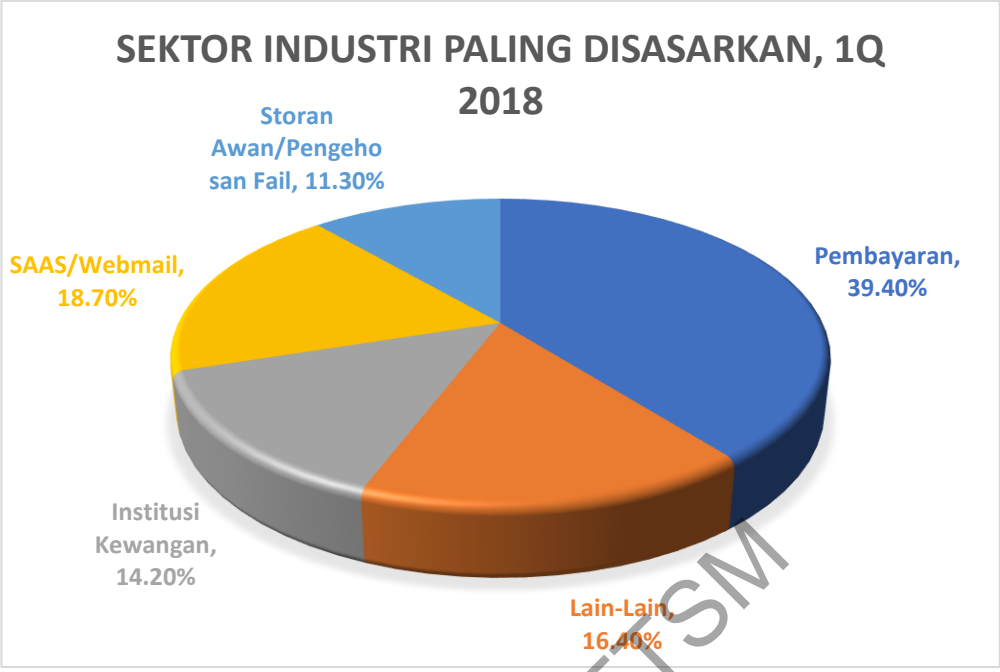


Rajah 1.1 Nilai kerugian penipuan e-perbankan tahunan di UK (2010 - 2020)

Laman Statista menerangkan, nilai tahunan kerugian hasil dari kes-kes penipuan perbankan dalam talian turun naik secara keseluruhan dalam tempoh pemerhatian, mencapai nilai kira-kira 159.7 juta pound British pada tahun 2020. Nilai kerugian penipuan perbankan dalam talian kedua terbesar di UK ditemui pada 2015, apabila jumlah nilai tahunan sebanyak 133.5 juta pound British direkodkan.

Penipuan perbankan dan pambayaran dalam talian telah menyumbang lebih daripada 150 bilion GBP kerugian sepanjang tahun 2019. Penipuan ini juga dikategorikan di dalam penipuan kewangan ketiga terbesar yang dilihat pada tahun 2020. Jenis penipuan kewangan lain termasuk kad pembayaran dan penipuan cek.

Menurut Laporan Trend Aktiviti Pancingan Data (2018) oleh Anti-Phishing Working Group (APWG), jumlah serangan pancingan data yang dikesan pada suku pertama tahun 2018 adalah 46% lebih tinggi daripada suku keempat 2017. Berdasarkan rajah 1.2 di bawah, sektor yang paling banyak disasarkan ialah perkhidmatan pembayaran, yang merupakan 39.4% daripada serangan pancingan data, diikuti dengan servis e-mel dengan 18.7%, institusi kewangan dengan 14.2% dan lain-lain sektor dengan 16.4%. (M. A. Adebowaleet. Al, 2019).



Rajah 1.2 Sektor yang paling disasarkan oleh penjenayah pancingan data

Ini dibuktkan oleh L. Karthika dan V. Perumal (2016) dengan nyatakan bahawa kira-kira 80% daripada kes pancingan data dilakukan terhadap perkhidmatan pembayaran dan kewangan. Jelas, ancaman pancingan data telah menjadi masalah serius kerana kerosakan yang meluas pada sasarannya, iaitu sistem pembayaran dalam talian dan institusi kewangan. Mengambil kes yang berlaku di Amerika Syarikat, kadar kerugian ekonomi mereka berkisar antara 61 juta USD hingga 3 billion USD akibat daripada serangan pancingan data. (Ali Aljofey et. al, 2020).

1.2.2 Kesan terhadap pengguna internet

Risiko pancingan data bukan sahaja boleh melumpuhkan sistem institusi perbankan, malah ia juga boleh mengakibatkan kesan yang amat teruk terhadap pengguna, seperti kehilangan wang, kebocoran maklumat peribadi, kecurian identiti, exploitasi privasi dan juga memberi impak ke atas gangguan emosi pengguna. Sebagai contoh, Apabila wang dan data dicuri, mangsa berkemungkinan mengalami tekanan perasaan dan seterusnya boleh mengakibatkan kemurungan, lebih-lebih lagi jika wang yang dicuri dalam jumlah yang besar, dan data mereka digunakan dijalan yang tidak sepatutnya (Rui Chen et. al, 2020). Berdasarkan kajian kesusastera yang telah dijalankan, berikut adalah impak utama serangan pancingan data :

1. Kehilangan wang

   Mangsa cenderung untuk kehilangan semua wang simpanan di dalam akaun bank mereka, jika profil e-dagang dan e-perbankan digodam dan diambil alih oleh penjenayah siber.

2. Kebocoran maklumat peribadi

   Segala data dan maklumat diri pengguna akan dicuri oleh pemancing data, termasuk semua informasi penting dan sulit, seperti alamat tempat tinggal, nombor akaun bank, senarai transaksi wang, nombor kad pengenalan dan nombor telefon yang berdaftar dengan pihak bank.

3. Kecurian identiti

   Penggodam boleh melakukan penipuan dan jenayah lain dengan menggunakan kesemua maklumat peribadi yang dicuri dari laman e-perbankan dan e-dagang.

Selain pengguna individu, serangan pancingan data ini juga boleh memberi kesan kepada sesebuah organisasi dan perniagaan dalam pelbagai cara, bergantung pada pelbagai faktor seperti saiz organisasi dan jumlah maklumat yang telah terjejas. Syarikat yang mengalami kebocoran data serius harus melaporkan insiden tersebut kepada pihak berwajib. Perkara ini seterusnya akan menjatuhkan reputasi syarikat tersebut, dan kepercayaan dari pekerja, rakan kongsi dan pelanggan akan hilang. berikut adalah dua impak terbesar serangan ini terhadap institusi perniagaan dan kewangan :

1. Kehilangan Harta Intelek

   Kesan yang terbesar yang akan dialami oleh sesebuah organisasi adalah kecurian harta intelek, yang boleh menjadi kerugian yang paling merosakkan. Kebocoran maklumat sulit milik syarikat dan kehilangan harta intelek boleh mengakibatkan kerosakan pada reputasi dan gangguan aktiviti operasi. Secara tidak langsung, nilai syarikat akan turun dengan mendadak, dan kesan ini memberi impak dan mengambil masa yang lama untuk dibaikpulih. Ini kerana, apabila penggodam mendapat akses kepada maklumat atau sistem sensitif milik sesebuah perniagaan atau institusi kewangan, mereka akhirnya boleh

mengambil fail penyelidikan, rahsia perdagangan, senarai pelanggan, formula dan pelan perniagaan akan datang.

2. Kerugian Kewangan

Satu lagi kesan perniagaan yang dijangkakan yang akan dialami oleh organisasi apabila mereka terlibat dalam penipuan pancingan data ialah sejumlah besar kerugian kewangan. Selain kerugian sebenar dalam akaun syarikat, syarikat boleh menjangkakan lebih banyak defisit kewangan apabila mereka terpaksa membayar denda kepada pihak-pihak yang terlibat, seperti membayar balik kepada pelanggan yang terjejas.

Secara keseluruhannya, serangan ini mempunyai kesan negatif bukan sahaja keatas individu dan sektor perniagaan, malahan juga ke atas ekonomi melalui kerugian kewangan yang dialami oleh perniagaan dan pengguna, sekaligus menjejaskan keyakinan pengguna dalam perdagangan dalam talian.

1.2.3 Implikasi terhadap institusi perniagaan dan kewangan

Semua risiko yang berpotensi untuk menyebabkan kecurian data dan maklumat daripada setiap pemegang akaun bank dan pembeli atas talian perlulah di analisis dan dikawal bagi mengelakkan sebarang serangan siber yang tidak di ingini berlaku. Kegagalan sesebuah syarikat dalam menangani masalah serangan siber, terutamanya risiko pancingan data dan kebocoran maklumat boleh mengakibatkan pelbagai kerugian terhadap organisasi dan juga merosakkan nama syarikat.

Salah satu akibat dari kebocoran data adalah kerugian kewangan. Banyak percubaan pelanggaran telah berjaya dan menyebabkan kerugian sejumlah besar wang yang dimiliki oleh organisasi perniagaan. Selain terpaksa mengeluarkan kos yang besar untuk memulihkan kesemua kerosakan yang dihadapi, organisasi yang terjejas juga perlu memaklumkan dan memberi pampasan kepada mereka yang terjejas - sama ada pelanggan atau pelabur. Tidak berakhir di situ, syarikat-syarikat ini juga perlu melabur dalam banyak sistem pertahanan keselamatan siber. Mereka perlu mengeluarkan kos untuk membayar kemas kini perisian, anti-virus, perkhidmatan penyelenggaraan

tembok api dan sistem keselamatan pangkalan data yang boleh melindungi pelayan mereka.

Bukan itu sahaja, malah reputasi dan nama syarikat juga akan turut terjejas. Sebuah syarikat akan membelanjakan wang yang besar untuk membina jenamanya bagi mengekalkan kedudukan organisasi mereka dalam pasaran tertentu pada tahap yang memuaskan. Walau bagaimanapun, usaha ini boleh dijatuhkan oleh penjenayah siber, kerana percubaan jahat mereka akan membawa kepada merosakkan jenama dan reputasi syarikat (Hossein Abrishan et. al, 2021). Apabila perkara ini berlaku, pesaing-pesaing dalam sektor yang sama akan mengambil peluang daripada situasi ini dan mengukuhkan strategi pemasaran dan promosi mereka untuk memenangi pelanggan dan pelabur syarikat mangsa.

1.2.4 Ciri-ciri laman web pancingan data

Sesebuah laman web pancingan data lazimnya akan menyerupai reka bentuk laman web asal, contohnya seperti laman-laman perbankan dalam talian, dimana setiap dari imej grafik, teks malah logo dari laman web bank yang sah telah dicuri dan digunakan dalam laman web palsu bagi mengelirukan pengguna (Reem Mohamed Ibrahim dan Yahia A. Fadlalla, 2018).

Hal ini menyebabkan kekeliruan dikalangan pengguna internet, jika mereka tidak berhati-hati apabila melayari laman web yang dikehendaki. Hal ini sering berlaku dikalangan pengguna yang lebih berumur, dimana kebanyakan pengguna dari golongan ini tidak begitu mahir dalam menggunakan internet. Mereka tidak dapat membezakan ciri-ciri yang terdapat pada sesebuah laman web palsu dengan laman yang sebenar, serta tidak dilengkapi dengan pengetahuan dan kesedaran tentang serangan-serangan siber, sekaligus menyebabkan sejumlah besar dari mereka menjadi sasaran pihak penggodam.

Walaupun begitu, terdapat satu ciri laman web pancingan data yang dapat dilihat dan dibezakan dengan mata kasar, iaitu URL. Salah satu langkah pertama yang perlu diambil untuk mengenal laman web pancingan data ialah melihat URL. Banyak ciri-ciri yang terdapat dalam URL yang boleh dilihat dan dikenalpasti sebagai URL yang meragukan, yang berkemungkinan akan membawa kepada laman web palsu. Panjang

URL, alamat IP yang digunakan, penggunaan token HTTPS dan bilangan sub domain dalam sesebuah URL dapat memberitahu kita tentang kesahihan laman web tersebut.

Ciri-ciri URL inilah yang banyak membantu para pengkaji dalam membangunkan sistem dan sistem bagi mengesan laman web pancingan data. Seperti yang diterangkan oleh Dragoș Glăvan et. al (2020), dapat disimpulkan bahawa terdapat tiga ciri utama yang mudah yang boleh dilihat oleh para pengguna, iaitu :

1. Protokol – Penggunaan HTTP atau HTTPS
2. Nama domain
3. Penggunaan dan bilangan simbol sengkang (-) dalam URL

1.3 OBJEKTIF KAJIAN

Kajian ini membentangkan kaedah pendekatan yang boleh digunakan untuk mengesan dan menganalisis serangan pancingan data terhadap pengguna laman perbankan internet. Secara khususnya, matlamat kajian ini adalah seperti berikut :

1. Objektif 1 : Membangunkan sistem pengesahan berasaskan web yang boleh mengesan serangan pancingan data dengan tepat berdasarkan ciri-ciri dari URL dan kaedah Pokok Keputusan.

2. Objektif 2 : Menilai keberkesanan dan ketepatan sistem pengesahan berasaskan web dalam mengesan serangan pancingan data.

1.4 SOALAN KAJIAN

Kajian ini akan menjawab soalan-soalan berikut :

1. Sistem apakah yang sesuai digunakan bagi mengesan laman web palsu?
2. Apakah tahap keberkesanan dan ketepatan sistem pengesahan laman web palsu?

1.5 SKOP KAJIAN

Kajian ini memfokuskan kepada pembangunan sistem pengesahan berasaskan laman web yang boleh mengesan serangan pancingan data dengan tepat berdasarkan ciri-ciri dari URL dan kaedah Pokok Keputusan. Semua ciri pada laman web palsu yang dibangunkan oleh penggodam akan dikenalpasti dan digunakan sebagai sumber data utama dalam membangunkan sistem *anti-phishing* ini. Kajian ini juga akan memberi tumpuan kepada penilaian terhadap keberkesanan dan ketepatan sistem pengesahan ini dalam mengesan serangan pancingan data. Hasil keputusan dari eksperimen ini akan turut dibincangkan dalam kertas ini.

1.6 KEPENTINGAN KAJIAN

Hasil kajian ini akan menyumbang secara tidak langsung dan memberi manfaat kepada institusi yang terlibat dalam pasaran dalam talian dan sekaligus menyambung kerja-kerja sedia ada untuk menyelesaikan masalah pancingan data di laman web dengan cara berikut :

1. Membangunkan sistem pengesahan berasaskan web yang boleh mengesan serangan pancingan data yang boleh diakses oleh semua pengguna internet tanpa sebarang perlu memuat turun sebarang perisian ke dalam peranti.
2. Mengenalpasti kelemahan sistem pengesanan dan memberi cadangan penambahbaikan terhadap fungsi dan ketepatan sistem tersebut.

1.7 ORGANISASI PENULISAN

Terdapat lima bab yang akan dibincangkan di dalam penulisan ini. Bab 1 merungkai definisi pancingan data, pernyataan masalah, objektif dan soalan kajian serta kepentingan kajian yang dilakukan ini. Bab 2 membincangkan tentang kajian dan pembangunan sistem pancingan data yang sedia ada, termasuk kaedah yang telah digunakan oleh para pengkaji sebelum ini. Pelbagai data dan statistik pancingan data turut dibincangkan bagi memastikan proses dan pemahaman penulis terhadap tajuk ini lebih baik. Semua kajian yang pernah dijalankan akan di semak dan difahami sebagai

bahan rujukan dalam melaksanakan projek ini. Seterusnya, Bab 3 membincangkan metodologi yang terlibat dalam proses rangka kerja dan pembangunan sistem anti-*phishing*. Segala maklumat yang dirujuk pada Bab 2 akan digunakan dan data yang diperoleh akan digunakan sebagai sampel atau set data bagi membangunkan sistem dan menguji ketepatan sistem. Bab 4 pula merujuk kepada ujian dan eksperimen yang akan dijalankan terhadap sistem *anti-phishing* tersebut. Kesemua pembolehubah dan parameter yang terlibat akan dijelaskan, dan susun atur ujian akan diterangkan secara terperinci. Selepas ujian dijalankan, keputusan yang terhasil akan di analisis dan dibincangkan. Sebarang kelemahan yang terdapat pada sistem *anti-phishing* tersebut juga akan dikawal dan cadangan penambahbaikan akan dimasukkan supaya pengkaji akan datang boleh membuat penilaian dan membangunkan sistem yang lebih baik dan bebas ralat. Bab yang terakhir, iaitu Bab 5, dimana rumusan dan kesimpulan akan dibuat berdasarkan hasil ujian dan cadangan kajian seterusnya akan diberikan.

BAB II

TINJAUAN KESUSASTERAAN

2.1 ULASAN BAB

Dalam bahagian ini, kita akan membincangkan faktor-faktor yang boleh menyebabkan seseorang pengguna dalam talian jatuh ke dalam perangkap penggodam. Semua penjelasan dan bukti-bukti kes pancingan data akan diulaskan. Setiap sampel dari kajian yang sedia ada akan disemak dan dihuraikan bagi mendapatkan ulasan yang lebih kukuh mengenai serangan siber ini. Sebarang berita atau kemas kini terkini mengenai topik pancingan data akan ditambah terus sepanjang tempoh kajian. Melalui cara ini kita akan dapat maklumat yang lebih terperinci dalam menentukan teknik penyelesaian yang paling berkesan bagi melawan serangan pancingan data.

2.2 FAKTOR SERANGAN

Perkembangan pesat teknologi pasaran dan pengiklanan atas talian telah memberi banyak faedah dah kesenangan kepada pelbagai pihak. Pada masa yang sama, pihak syarikat yang terlibat juga sedar bahawa mereka perlu melindungi platform laman web mereka dari sebarang ancaman dari pihak-pihak yang tidak bertanggungjawab. Setiap institusi perlu mengenal pasti kepelbagaian dalam faktor serangan dan seterusnya mengambil langkah sewajarnya untuk membentung risiko ancaman ini. Dalam kertas ini, kita mengklasifikasikan dua faktor yang menyumbang kepada isu serangan siber ini, iaitu tabiat pengguna dan kelemahan sistem perkhidmatan dalam talian.

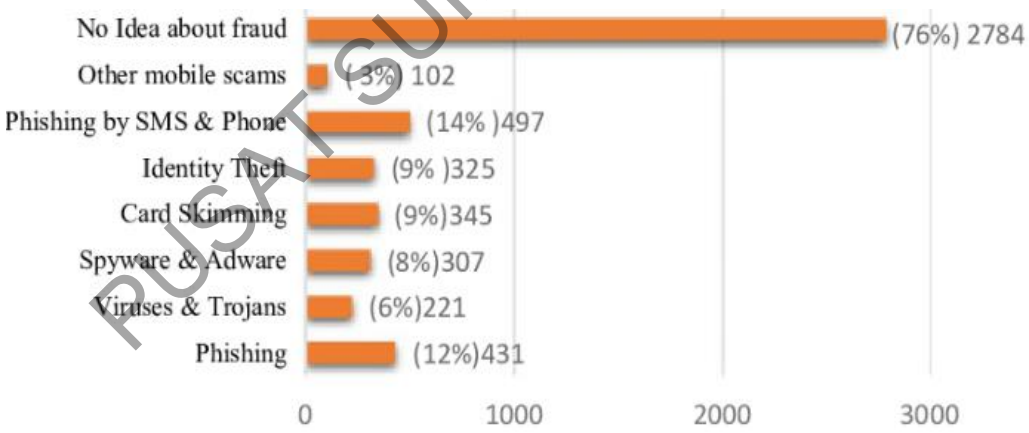
2.2.1 Tabiat pengguna

Seperti yang kita sedia maklum, manusia adalah faktor paling lemah yang boleh membawa kepada sebarang bentuk serangan siber. Penggodam sering mengambil

kesempatan untuk menggunakan kejuruteraan sosial atau “social engineering” untuk memanipulasi sasaran mereka. Penggodam akan menyamar sebagai seorang yang mempunyai kedudukan tinggi dalam organisasi tertentu atau berpura-pura menjadi mana-mana entiti yang dikenali, contohnya seperti institusi perbankan.

Dengan menggunakan kejuruteraan sosial seperti ini, ia terbukti berjaya apabila penyerang menggunakan teknik penyamaran bagi mencuri data mangsa mereka. Dakwaan di atas dibuktikan melalui statistik dan laporan mengenai insiden kebocoran maklumat dan pancingan data yang berlaku kepada di seluruh dunia, khususnya dalam sektor perniagaan dalam talian dan e-banking.

Selain itu, paras kesedaran pengguna juga menyumbang kepada ancaman ini. Rajah 2.1 di bawah adalah keputusan dari hasil kaji selidik yang dijalankan oleh penyelidik terdahulu, dan ia membuktikan bahawa kebanyakan pengguna laman e-perbankan tidak mempunyai kesedaran tentang penipuan dalam talian. (Musfika Nuha et. al, 2021).



Rajah 2.1 Kesedaran pengguna terhadap penipuan e-perbankan

Berdasarkan kajian yang dilakukan oleh Musfika Nuha et al (2020), keputusan menunjukkan bahawa 76% pengguna laman e-perbankan tiada kesedaran terhadap penipuan perbankan mudah alih. Tambahan pula, penemuan penyelidik menunjukkan bahawa 86.3% daripada mangsa penipuan e-perbankan atau perbankan mudah alih tidak mempunyai pengetahuan awal tentang jenis penipuan ini. Pada masa yang sama, 13.7% mangsa dalam sektor tersebut mempunyai pengetahuan yang jelas tentang penipuan dan ancaman pancingan data dalam talian. Jelas sekali, di sebalik serangan ini, kekurangan

pengetahuan dan kesedaran boleh menjadi faktor utama yang menyumbang kepada peningkatan kes jenayah siber.

2.2.2 Kelemahan sistem perkhidmatan dalam talian

Kebiasaannya, para penggodam mempunyai banyak kaedah yang canggih untuk memanipulasi mekanisma keselamatan sistem-sistem dalam talian, dan mereka juga boleh mengeksploitasi kelemahan sistem ini untuk memulakan serangan. (Reem Ibrahim dan Yahia A. Fadlalla, 2018). Pada hari ini, dapat kita lihat bahawa perlindungan dua faktor telah diperkenalkan dalam aplikasi perbankan dan pembayaran mudah alih, dan kebanyakan laman media sosial pada hari ini telah didatangkan dengan ciri teknologi pengesahan biometrik, seperti pengecaman cap jari dan sebagainya.

Walau bagaimanapun, sistem biometrik boleh menyumbang kepada dua jenis kegagalan pengesahan; yang pertama ialah pemadanan palsu. Ia berlaku apabila dua sampel daripada individu yang sama mempunyai tahap persamaan yang rendah dan sistem tidak dapat memadankan kedua-dua sampel dengan betul. Manakala kegagalan yang kedua pula ialah padanan palsu yang boleh mengakibatkan pencerobohan oleh penggodam dan pemancing data. (Reem Mohamed Ibrahim dan Yahia A. Fadlalla, 2018).

Cabaran utama dalam mengesan serangan pancingan data adalah untuk mencari teknik yang digunakan. Kebiasaannya, penggodam akan meningkatkan strategi mereka secara berterusan dan mencipta laman web yang mampu melindungi diri mereka daripada pelbagai bentuk pengesanan. Sehubungan itu, kaedah pengesanan pancingan data dengan teknik yang sesuai amat diperlukan untuk menentang para penjenayah siber ini. (AlMaha Abu Zuraiq, Mouhammd Alkasassbeh. 2019).

Mengambil contoh serangan pada laman web perbankan internet, menurut Nilay Yildirim dan Asaf Varol (2019), IDC meramalkan bahawa apabila bank-bank cuba untuk mencipta kepercayaan digital dengan pelanggan mereka, kos perbelanjaan bagi menaiktaraf ciri-ciri keselamatan ke atas perbankan dalam talian meningkat sebanyak

20%. Kos ini dapat dihuraikan dengan melihat kepada jadual 2.1 yang dibentangkan oleh penulis seperti di bawah:

Jadual 2.1 Ancaman perbankan dalam talian

Serangan keatas peranti	**Serangan keatas rangkaian**	**Serangan keatas pusat data**
Pelayar web: Pancingan data *Man-In-The-Middle*	Wi-Fi: Lemah / Tiada penyulitan Sijil SSL palsu / tamat tempoh	Pelayan web: Salah konfigurasi *Cross-Site Scripting (XSS)*
Sistem e-perbankan : Kata laluan yang lemah Tiada penyulitan data		Pangkalan data: *SQL injection*

Dengan melihat kepada jadual ini, serangan pancingan data dikategorikan sebagai serangan yang paling biasa digunakan keatas pelayar web, kerana ia tidak memerlukan sebarang kemahiran teknikal. Kelemahan pihak mangsa dalam menetapkan ciri kata laluan yang perlu digunakan oleh para pengguna juga memberi ruang kepada pihak penggodam untuk melakukan pencerobohan terhadap laman web ini dan seterusnya bertindak mencuri data pengguna.

Selain itu, Reem Mohamed Ibrahim dan Yahia A. Fadlalla (2018) turut membincangkan tentang kelemahan penggunaan Teks SMS kata laluan satu kali (OTP). Paradigma utama SMS OTP ialah untuk menghantar kod Transaksi kepada nombor telefon yang berdaftar dengan pihak bank atau syarikat yang menawarkan khidmat pembayaran atas talian. Pengguna memasukkan kod OTP ini untuk memberi kebenaran bagi meneruskan transaksi untuk sistem dalam talian. Dalam mekanisme ini, kod OTP dihantar sebagai mesej teks ke telefon pintar pengguna. Walau bagaimanapun, keselamatan SMS OTP bergantung pada kerahsiaan mesej SMS yang bergantung kepada keselamatan rangkaian mudah alih. Namun, pelbagai bentuk serangan siber yang berlaku kebelakangan ini telah membuktikan bahawa kerahsiaan mesej SMS juga boleh ditembusi oleh para penggodam. Tambahan pula, penggodam telah mencipta virus Trojan bagi tujuan mengekploitasi kos OTP ini.

2.3 KESAN-KESAN TERHADAP PENGGUNA

Jumlah serangan pancingan data yang telah dilaporkan setakat 2017 adalah sangat besar. Anggaran bilangan serangan yang dilaporkan mencecah 156 juta sehari (Ibrahim Mohammed Alseadoon1 et. al., 2017). Rui Chen et. al (2020), kesan utama yang kebiasaannya dilaporkan oleh mangsa serangan pancingan data adalah kehilangan wang. Membuat kesilapan terutamanya dalam perbankan dalam talian akan mengakibatkan kerugian tyerhadap pengguna, lebih-lebih lagi apabila mereka menggunakan data-data peribadi bagi melakukan transaksi atas talian (Ibrahim Mohammed Alseadoon1 et. al (2017). Bukan sahaja mereka akan kehilangan wang, malah maklumat sensitif mereka juga dicuri dan diexploitasi oleh para penggodam. Hal ini sering terjadi kerana laman web pancingan data sering terlihat seperti laman web yang sah, bagi mengelirukan dan menarik pengguna memberikan maklumat sensitif tanpa mereka sedar (Ibrahim Mohammed Alseadoon1 et. al., 2017). Menurut Hossein Abrishan et. al (2021), kebiasaannya penggodam menggunakan teknikal, sosial, dan kelemahan psikologi orang ramai untuk memperoleh maklumat sensitif mangsa dan gunakan maklumat tersebut untuk mencuri aset kewangan mereka atau melancarkan serangan lain ke atas mangsa.

Malah, APWG (Kumpulan Kerja Anti-Phishing) mentakrifkan pancingan data sebagai mekanisma jenayah yang menggunakan kedua-dua kejuruteraan sosial dan penipuan teknikal untuk mencuri data identiti peribadi dan kelayakan akaun kewangan pengguna. Pancingan data ialah kaedah yang sangat popular digunakan dalam serangan rangkaian dan membawa kepada kebocoran privasi, kecurian identiti dan kerosakan harta benda (Peng Yang et. al., 2019). Apabila seseorang pengguna menjadi mangsa, ia akan sekaligus menjejaskan emosi dan mental mangsa, yang seterusnya akan menyebabkan mangsa berasa sedih, marah, serta hilang kepercayaan.

Kesan yang sama dapat dilihat jika serangan ini berlaku keatas organisasi yang lebih besar. Dari sudut pandangan organisasi, kesan serangan pancingan data dinilai dari segi tanggungan kos kerosakan, kerugian kewangan, dan kejatuhan nama dan prestasi sesebuah organisasi (Hamidreza Shahbaznezhad et. al., 2020). Gregor Petrič dan Kai Roer (2022) melaporkan bahawa serangan pancingan data merupakan punca utama pelanggaran keselamatan sesebuah syarikat. Ini kerana, serangan pancingan data

boleh menyebabkan kerugian kewangan yang besar bukan sahaja untuk pengguna, tetapi untuk keseluruhan organisasi tersebut. Jika pengguna gagal dalam proses pengesanan, dan serangan ini membawa kepada seseorang pekerja untuk muat turun fail yang telah dilengkapi dengan perisian hasad oleh penggodam, serangan itu boleh memberi impak kepada nilai sebuah organisasi dan mampu menjatuhkannya (Rui Chen et. al (2020). Data yang bocor dari sesebuah syarikat amatlah bernilai kepada penggodam, mereka mampu menjual malah mengoksploitasi data tersebut untuk tujuan kepentingan peribadi. (Hamidreza Shahbaznezhad et. al., 2020). Jika serangan ini dilaporkan dan diketahui oleh masyarakat umum, khususnya mereka yang menjadi pelanggan kepada syarikat tersebut, akan tercemar dan runtuhlah reputasi syarikat ini. Syarikat juga akan berdepan risiko kehilangan kepercayaan dari pelanggan, dan lebih teruk, apabila pelanggan menarik diri dari syarikat tersebut (Gregor Petrič dan Kai Roer, 2022).

Meskipun telah banyak kempen kesedaran yang dianjurkan oleh hampir kesemua organisasi, terdapat juga segelintir pekerja yang mengambil mudah dan tidak endah akan bahayanya serangan pancingan data ini. Kecenderungan pekerja untuk berinteraksi dengan e-mel pancingan data, mahupun melawat laman-laman sesawang palsu ketika berada dipejabat meletakkan aset organisasi pada risiko yang amat tinggi (Peng Yang et. al., 2019). Para pekerja yang mengakses laman web palsu akan mengklik pada pautan dan mendedahkan beberapa maklumat sensitif, termasuklah data dan informasi sulit syarikat, yang kemudiannya disalahgunakan oleh penyerang (Gregor Petrič dan Kai Roer, 2022).

Secara keseluruhannya, impak dan kesan yang ditinggalkan oleh penyerang setelah melakukan serangan pancingan data terhadap seseorang individua atau sesebuah organisasi mampu memberi kerosakan bukan sahaja pada harta, malah juga keatas emosi dan tekanan dari masyarakat luar. (Ibrahim Mohammed Alseadoon1 et. al., 2017).

2.4 CIRI-CIRI LAMAN WEB PANCINGAN DATA

Kebanyakan laman web pancingan data direka bentuk dengan baik bagi meniru dan menyerupai laman web yang asal. Ini adalah salah satu punca yang mengakibatkan

pengguna yang menghabiskan lebih sedikit masa dalam talian mungkin tidak dapat membezakan antara laman web palsu dan laman web yang sah (Ibrahim Mohammed Alseadoon1 et. al., 2017). Walau bagaimanapun, menurut Ibrahim Mohammed Alseadoon1 et. al (2017), terdapat petunjuk khas yang boleh membawa kepada pengesanan penipuan dalam laman web pancingan data. Contohnya, laman web pancingan data mempunyai nama domain (URL) yang hampir serupa dengan domain tapak web yang asal, dengan hanya satu atau dua perbezaan huruf atau nombor.

Kebanyakan laman web palsu juga mempunyai maklumat yang boleh mendedahkan identiti mereka seperti sijil yang tidak sah (Hamidreza Shahbaznezhad et. al., 2020). Selain dari URL, terdapat pelbagai lagi ciri yang dapat diperhatikan bagi membezakan sesebuah laman web palsu dengan laman web asli, seperti reka letak, CSS, teks, imej dan juga kod HTML (Peng Yang et. al., 2019). Hossein Abrishan et. al (2021) menjelaskan bahawa sebahagian besar laman web palsu yang dibangunkan oleh penggodam mengandungi imej yang kelihatan tulen, contohnya seperti penggunaan logo tulen daripada syarikat atau bank-bank terkenal. Malah, laman web ini juga sering menggunakan nama yang sama dengan laman web asal.

Satu sumber dari laman web Repositori Pembelajaran Mesin UCI atau dikenali sebagai The UCI Machine Learning Repository telah mengektrak dan menyenaraikan ciri-ciri unik yang lazimnya didapati pada laman web pancingan data. Ciri-ciri yang diekstrak termasuklah alamat IP, jangka hayat domain, status SSL, Kedudukan laman web, rekod DNS, nama domain yang dimuat naik dan beberapa lagi ciri yang telah dikenalpasti sebagai palsu. Ciri-ciri ini telah dipilih dan digunakan sebagai sumber utama dalam membangunkan sistem *anti-phishing*, dan penerangan untuk setiap ciri akan diterangkan dengan lebih lanjut dalam Bab 3.

2.5 KAEDAH MENGESAN SERANGAN PANCINGAN DATA

Merujuk kepada kajian yang dilakukan oleh Alhuseen O. Alsayed dan Anwar L. Bilgrami (2017), dapat kita simpulkan bahawa gambaran keseluruhan pendekatan pengesanan terhadap serangan pancingan data boleh diklasifikasikan secara meluas kepada dua jenis, iaitu teknik berasaskan kesedaran kepada pengguna dan teknik berasaskan teknikal dan pembangunan sistem dan perisian komputer.

2.5.1 Teknik kesedaran pengguna

Walaupun sistem pembayaran dalam internet dan media sosial telah digunakan secara meluas di banyak negara maju, kesedaran pengguna di negara membangun adalah lebih perlahan daripada yang dijangkakan (Fadare Olusolade Aribake and Zahurin Mat Aji, 2021). Ini adalah disebabkan oleh tahap kepercayaan yang rendah dalam platform transaksi dalam talian, seperti rasa tidak selamat atau kurang keyakinan (Viswanadham, 2017).

Menurut Alhuseen O. Alsayed dan Anwar L. Bilgrami (2017), dalam usaha membendung isu serangan pancingan data terhadap pengguna internet, semua institusi yan terlibat haruslah mendidik pelanggan mereka tentang risiko serangan pancingan data serta menyediakan langkah-langkah yang boleh diambil untuk melindungi maklumat peribadi dan kewangan mereka. Pengguna juga perlu berhati-hati terhadap serangan pancingan data dengan mencuba untuk mengenal pasti ciri-ciri yang terdapat di dalam laman web pancingan data. Mereka boleh dilatih mengenai kesedaran keselamatan melalui akses kepada kandungan dan bahan bacaan yang lengkap. Kaedah ini boleh dicapai melalui pelbagai bentuk usaha seperti menambah kandungan ke laman-laman web yang terlibat, pemberitahuan melalui e-mel atau meletakkan poster di cawangan-cawangan organisasi (Agilandeeswari et. Al, 2019).

Memberi pengguna lebih banyak maklumat tentang ancaman pancingan data boleh meningkatkan pengetahuan mereka tentang kemungkinan kelemahan yang terdapat dalam sesebuah laman web e-perbankan. Bagi memastikan keberkesanan penyampaian maklumat ini, pendidikan dan kesedaran terhadap setiap individu yang menggunakan perkhidmatan dalam talian perlu dilaksanakan dan dijadikan amalan secara berterusan (Fadare Olusolade Aribake dan Zahurin Mat Aji, 2021).

Mengambil contoh dari sektor kewangan, satu kajian ke atas bank tempatan telah dijalankan oleh Zeti Suhana Zainudin dan Nurul Nuha Abdul Molok (2018) untuk mengkaji faktor-faktor yang mempengaruhi peningkatan ancaman jenayah siber dan kesedaran dalam kalangan pengamal keselamatan siber di institusi perbankan. Kajian penyelidik ini bertujuan untuk menyiasat strategi keselamatan yang digunakan oleh

pihak bagi melindungi mereka daripada serangan siber. Penemuan kajian ini mendedahkan bahawa kesedaran ancaman serangan siber adalah lebih berkesan melalui pembelajaran tidak formal berbanding pembelajaran formal. Oleh itu, institusi perbankan perlu membuat pertimbangan sewajarnya dalam mempertahankan maklumat pengguna mereka daripada jatuh ke tangan individu yang berniat jahat dan melakukan jenayah siber yang secara dasarnya, menyalahgunakan kecanggihan teknologi masa kini untuk membuat jenayah yang menyalahi undang-undang (Fadare Olusolade Aribake dan Zahurin Mat Aji, 2021).

Selain bank, para peniaga dan syarikat yang menjalankan pelbagai perkhidmatan atas talian juga boleh menggunakan kemudahan sistem dalam talian masing-masing bagi mengukuhkan kempen kesedaran keselamatan siber melalui paparan poster dan pengedaran risalah di setiap rangkaian cawangan. Usaha ini dapat mewujudkan impak yang lebih besar dalam usaha memerangi ancaman pancingan data di laman-laman mereka supaya pengguna lebih dilindungi dan seterusnya dapat melaksanakan transaksi dengan lebih yakin.

Inovasi teknologi yang dibangunkan oleh penjenayah siber sedemikian telah membawa kepada pengukuhan Pengurusan Pengetahuan (KM) dalam meningkatkan pengetahuan dan kesedaran para pengguna internet (Fadare Olusolade Aribake dan Zahurin Mat Aji, 2021). KM menandakan pendekatan yang bernas dan sistematik untuk menjamin operasi penuh pangkalan pengetahuan organisasi, ditambah pula dengan kemahiran asas, pengetahuan individu, inovasi dan idea untuk mewujudkan organisasi yang lebih cekap dan berkesan. Inisiatif membangunkan KM oleh penyelidik ini dilihat sebagai wajar dalam mengukuhkan penggunaannya dalam penggunaan internet bagi memerangi serangan pancingan data.

2.5.2 Teknik pembangunan sistem keselamatan

Melihat kepada pendekatan yang diambil oleh penyelidik kajian-kajian terkini, pelbagai idea dan cadangan merekabentuk sistem *anti-phishing* dibincangkan. Terdapat juga cadangan-cadangan penambahbaikan terhadap sistem perkhidmatan dalam talian yang sedia ada. Sebagai contoh, Waleed A. Hammood et al (2020) menyemak akses

pengguna perbankan dalam talian pada telefon bimbit dan mengkaji fungsi kad SIM. Pengklonan, ancaman kepada kad SIM juga diterangkan. Hasil kajian menunjukkan bahawa semua sistem keselamatan untuk akses pengguna dalam talian mengandungi kata laluan dalam bentuk biometrik atau PIN. Oleh itu, penyelidik mencadangkan idea akses pengguna berdasarkan Nombor Identiti Peralatan Mudah Alih Antarabangsa (IMEI) untuk mengukuhkan keselamatan pengguna.

Satu lagi sistem dibangunkan oleh Tej Narayan Thakur dan Noriaki Yoshiura (2021), dimana sistem yang dinamakan sebagai sistem AntiPhiMBS ini berfungsi sebagai medium *anti-phishing* untuk sistem pembayaran dalam talian. Sistem ini menggunakan ID pengguna untuk pengesahan dan kemudiannya dipautkan kepada ID aplikasi yang hanya diketahui oleh pengguna, dan sistem e-perbankan sahaja. Hasil ujian pengesahan yang dijalankan oleh penyelidik melalui AntiPhiMBS telah berjaya membuktikan bahawa AntiPhiMBS-Auth yang dicadangkan adalah bebas ralat, dan institusi yang terlibat boleh melaksanakan sistem yang disahkan seperti AntiPhiMBS untuk mengurangkan serangan pancingan data terhadap laman web masing-masing.

Selain dari sistem yang dicadangkan di atas, terdapat banyak lagi kajian dan rangka kerja yang dibangunkan oleh para penyelidik, dan kebanyakan mereka penggunaan pendekatan berasaskan algoritma pembelajaran mesin atau lebih dikenali sebagai machine learning untuk pengesanan pancingan data. Pelbagai teknik diuji, dan pemilihan teknik yang sesuai dipilih berdasarkan keputusan dari kepelbagaian set data, ujian ketepatan, kecekapan dan kemudahan yang ditawarkan oleh setiap algoritma. Dalam kajian ini, kita akan membincangkan tentang teknik-teknik machine learning yang pernah diuji oleh para penyelidik terdahulu, dan seterusnya membangunkan sistem *anti-phishing* dengan menggunakan algoritma yang bersesuaian dengan kehendak keselamatan laman web masa kini.

2.6 KAJIAN-KAJIAN YANG BERKAITAN

Terdapat banyak kajian yang dilakukan oleh para penyelidik terhadap penggunaan teknik pembelajaran mesin dalam usaha mengekang isu pancingan data. Algoritma pembelajaran mesin telah menjadi salah satu teknik yang ampuh dalam mengesan laman web pancingan data (R. Kiruthiga dan D. Akila, 2019).

Ammara Zamir et. al (2019) membincangkan tentang kesesuaian sesuatu algoritma pembelajaran mesin dalam membina sistem yang boleh mengesan kehadiran pancingan data dengan tepat. Ciri-ciri set data untuk mengesan pancingan data ini dianalisis dengan menggunakan pelbagai teknik termasuk perolehan maklumat, nisbah keuntungan, Relief-F dan penyingkiran ciri rekursif (RFE). Antara algoritma utama yang digunakan adalah Pokok Keputusan, Hutan Rawak, rangkaian neural atau lebih dikenali sebagai neural network, mesin vektor sokongan dan Naïve Bayes. Selepas itu, dua sistem digunakan dengan menggabungkan pengelas pemarkahan tertinggi untuk meningkatkan ketepatan klasifikasi laman web yang mengandungi ciri pancingan data. Dalam ulasan mereka, dapat dihuraikan bahawa penggabungan tiga algoritma, iaitu Random Forest, rangkaian neural dan bagging mengatasi semua teknik lain dengan mencapai ketepatan sebanyak 97.4%.

Teknik penggabungan yang sama juga telah digunakan oleh Lizhen Tang dan Qusay H. Mahmoud (2021), di mana perbandingan antara pelbagai algoritma dalam pembelajaran mesin dihuraikan dengan lebih terperinci, seperti di dalam jadual 2.2 di bawah.

Berdasarkan jadual ini, penyelidik merumuskan bahawa teknik Random Forest telah mengesan dan memberi ketepatan yang paling tinggi berbanding teknik-teknik lain, termasuk kesemua penggabungan algoritma yang lain. Walau bagaimanapun, keputusan ketepatan ini bergantung kepada sampel set data yang digunapakai, dan juga tetapan yang ditetapkan dalam setiap fungsi algoritma yang dipilih.

Jadual 2.2 Perbandingan hasil ujian ketepatan algoritma pembelajaran mesin (Lizhen Tang dan Qusay H. Mahmoud, 2021)

Sistem / Algoritma	**Jenis**	**Set Data**	**Cabaran**	**Kekangan**	**Ketepatan**
Hutan Rawak	Satu	ISCXURL-2016	Mencapai kualiti yang tinggi dan tindak balas masa yang kurang tanpa kebergantungan pada perkhidmatan dari pihak ketiga.	Tidak menggunakan data set yang berbeza bagi melatih sistem, membandingkan keputusan ataupun menilai kekukuhan sistem.	99.57%

bersambung…

...sambungan			Juga menggunakan ciri-ciri yang terhad daripada URL		
Hutan Rawak	Satu	Laman Sesawang (*phishTank, OpenPhish, Alexa*, Pembayaran Dalam Talian) 5223 contoh; 2500 URL pancingan data; 2723 URL yang sah; 20 ciri	Kesemua data diperolehi daripada laman sesawang yang sah, 20 ciri juga diekstrak secara manual. Beberapa ciri mesti diperolehi daripada khidmat pihak ketiga dah sesetengah ciri perlu dihuraikan daripada kod HTML laman sesawang.	Tidak menggunakan data set yang berbeza bagi melatih sistem, membandingkan keputusan ataupun menilai kekukuhan sistem. Set data ujian adalah kecil.	99.50%
PSL [1] + PART	Gabungan	Laman Sesawang (*phishTank, OpenPhish, RelBank*) 30,500 contoh; 20,500 URL pancingan data; 10,000 URL yang sah; 3000 sampel data yang sah 18 ciri	Diekstrak daripada 3000 ciri yang komprehesif dan menggunakan set parameter yang berbeza pada sistem pembelajaran mesin bagi membandingkan keputusan ujian.	URLs yang sah dalam data set kesemuanya mempunyai kaitan dengan bank dah sesetengah cir-ciri hanya terhad sistem e - perbankan.	99.30%
ISHO + SVM	Gabungan	UCI	Algoritma ISHO digunakan untuk memilih ciri - ciri yang lebih cekap	Data set UCI adalah sumber terbuka dan mengandungi 11,500 contoh dengan ciri - ciri yang telah dinormalkan tetapi tidak mengandungi URL yang asal dan pendekatan yang dicadangkan juga tidak mempunyai prosedur pengambilan.	99.64%

bersambung...

...sambungan

Adaboost	Satu	Laman Sesawang (*phishTank, MillerSmiles*, Carian *Google*); Saiz data tidak dinyatakan; Setiap contoh mempunya 30 ciri	Sistem yang dicadangkan menggunakan WEKA 3.6, Python dan MATLAB 2	Tidak menggunakan data set yang berbeza bagi melatih sistem, membandingkan keputusan ataupun menilai kekukuhan sistem.	99.30%
LBET (*Logistic Regression + Extra Tree*)	Gabungan	UCI	Menggabungkan algoritma pembelajaran *Meta* dan *Extra Trees* bagi mencapai tahap ketepatan yang tinggi serta kadar positif yang rendah.	Sumber data yang tidak mencukupi dan kekurangan proses pengekstrakan ciri	99.57%
Bootstrap Aggregating + Logistic Sistem Tree	Gabungan	UCI	Pengelas telah dilatih dan diuji berdasarkan 10 pengesahan bersilang untuk mengurangkan bias dan varian.	Sumber data yang tidak mencukupi dan kekurangan proses pengekstrakan ciri	99.42%

Terdapat juga kajian yang membuktikan teknik Rangkaian Neural juga mampu menghasilkan sistem yang memberi keputusan yang tepat dan kadar pengesanan laman web pancingan data yang tinggi. Y. Yerima dan Mohammed K. Alzaylaee (2020) menjalankan eksperimen dengan pendekatan yang sama, namun mereka telah memilih teknik yang lebih khusus, iaitu Konvolusi Rangkaian Neural (CNN). Mereka membentangkan reka bentuk sistem berasaskan CNN sebagai medium utama pengesanan laman web pancingan data dan menilai sistem tersebut dengan menggunakan set data yang diperolehi daripada 4898 laman web pancingan data dan 6157 laman web tulen. Analisis perbandingan mereka menunjukkan bahawa sistem berasaskan CNN berjaya mencapai prestasi pengesanan pancingan data yang terbaik sebanyak 98.2%, berbanding dengan algoritma-algoritma lain.

Pada tahun 2020, Aljofey et al. juga telah mencadangkan penggunaan sistem CNN bagi mengesan laman web pancingan data dengan hanya berdasarkan pautan URL. Mereka mengekstrak ciri-ciri yang terdapat pada pautan URL asal, dan kemudian menjalankan ujian dengan menggunakan sistem CNN. Keputusan eksperimen mereka telah menunjukkan bahawa sistem ini memperoleh ketepatan sebanyak 95.02% pada 318642 set data yang mereka gunakan.

Satu lagi kajian yang dijalankan oleh Ying-fang Li et. al (2017) yang menjalankan kajian dengan menggunakan sistem Pokok Keputusan pula merumuskan bahawa algoritma ini mempunyai keupayaan pengelasan yang sangat baik dan kestabilan yang tinggi jika dibandingkan dengan kaedah pembelajaran mesin yang lain.

Begitu juga dengan Erzhou Zhu et. al (2020), di mana mereka mencadangkan satu sistem yang dinamakan DTOF-ANN yang dibangunkan dengan menggunakan kaedah gabungan antara Pokok Keputusan dan Rangkaian Neural Tiruan bagi mengesan pancingan data. Hasil ujikaji terhadap sistem ini didapati bahawa bahawa DTOF-ANN mempamerkan prestasi yang lebih tinggi daripada kebanyakan kaedah dan sistem yang sedia ada oleh penyelidik terdahulu (Erzhou Zhu et. al, 2020).

Namun begitu, setiap hasil dari ujian penyelidikan bergantung kepada kualiti dan kebolehpercayaan ciri-ciri laman web yang diekstrak dan jumlah data set yang digunakan. (Ammara Zamir et. al, 2019). Dengan ini, kajian yang lebih terperinci terhadap kesesuaian sesuatu algoritma perlu dijalankan bagi mendapatkan keputusan yang boleh menjadi tanda aras dalam mengenalpasti teknik yang paling sesuai dalam mengesan setiap laman web palsu.

Cara paling mudah bagi pemancing data untuk menipu pengguna internet adalah dengan cara menjadikan halaman web pancingan data serupa dengan sasaran mereka. Walau bagaimanapun, banyak ciri dan ciri tersendiri boleh membezakan laman web asal yang sah daripada laman web pancingan data klon seperti ralat ejaan, pengubahan imej, alamat URL yang panjang dan tidak normal (M.A. Adebowale et. al, 2019).

Selaras dengan itu, para penyelidik telah mengambil banyak ciri yang terdapat di laman web dan menjadikannya sebagai set data dan contoh bagi setiap eksperimen

yang dijalankan. Ankit Kumar Jain dan B. B. Gupta (2018) mengesan serangan pancingan data dengan mengekstrak dan menganalisis setiap pautan yang terdapat dalam kod sumber HTML laman web. Kemudian, data ini dimasukkan kedalam *machine learning* dan diuji. Keputusan ujian mereka menunjukkan bahawa kaedah ini sangat cekap dalam mengklasifikasi laman web pancingan data kerana ia menunjukkan kadar 98.42% bagi ketepatan pengesanan pautan pancingan data. Selain dari pautan dan URL, pelbagai ciri laman web boleh digunapakai sebagai data contoh bagi ujian pengesanan serangan siber ini. Dragoș Glăvan et. al (2020) menghuraikan setiap perincian ciri laman web yang kerap dijadikan sasaran penjenayah siber dalam memalsukan sesebuah laman web mengikut kekerapan, seperti yang ditunjukkan di dalam Jadual 2.3 di bawah.

Jadual 2.3 Ciri laman web yang boleh dipalsukan oleh pemancing data

Ciri-ciri	**Penerangan**
Penggunaan HTTPS	Sambungan rangkaian yang selamat menggunakan HTTPS
Sub-domain dalam URL	Beberapa *sub domain* dalam URL dianggap tidak selamat dan membawa kepada laman web palsu
Simbol sengkang (-)	digunakan untuk membuktikan bahawa URL tersebut adalah sah
Menggunakan alamat IP dalam nama domain	Penggidam cuba menyembunyikan nama dengan menggunakan nombor
URL yang panjang	Digunakan untuk menyembunyikan kata kunci yang mencurigakan
URL yang dimasukkan	Kesemua teks / gambar haruslah dimuat naik melalui URL domain
Penggunaan tetingkap pop-up	Menggunakan tetingkap pop-up untuk kemasukan kata laluan adalah tidak beretika
Rekod DNS	URL dianggap mencurigakan jika tiada rekod DNS
Trafik laman web	Laman web dianggap palsu jika tiada trafik
Jangka hayat domain	Laman web dianggap palsu jika tempah penggunaan dan pendaftaran kurang dari setahun

Contoh-contoh lain yang dapat diambilkira di sesebuah laman web adalah seperti kadar trafik laman web, penggunaan teks dan imej klon, saiz rangka antaramuka, kod sumber Javascript dan banyak lagi (M.A. Hossain et. al, 2019).

2.7 KAJIAN SAMBUNGAN

Kajian ini adalah kerja penambahbaikan daripada penyelidikan yang telah dilakukan oleh Akila D. (2019) yang bertajuk *Phishing Websites Detection Using Machine*

Learning. Dalam kajian tersebut, penyelidik telah menganalisis serangan siber dan mencadangkan pembangunan sistem sebagai medium pengesanan pancingan data dengan berkesan dengan menggunakan ciri-ciri semantik dalam laman sesawang yang menggunakan Bahasa Cina. Sebanyak 11 ciri telah diekstrak dan dikategorikan ke dalam lima kelas untuk memperoleh ciri statistik laman web tersebut. Set data yang mengandungi URL yang sah diperoleh daripada laman web DirectIndustry dan sampel pancingan data data diperoleh daripada laman *Anti-Phishing Alliance of China*. Sistem yang dicadangkan adalah unik kepada halaman web berbahasa Cina dan ia mempunyai pergantungan terhadap bahasa yang tertentu.

Berdasarkan ciri-ciri di atas, penyelidik telah mencadangkan cara yang cekap untuk mengesan pancingan data pada laman web URL dengan menggunakan pendekatan Pokok Keputusan. Teknik ini mengekstrak ciri-ciri yang terdapat pada laman web dan mengira nilai heuristik. Nilai ini seterusnya akan dimasukkan kedalam algoritma Pokok Keputusan untuk menentukan sama ada laman web tersebut adalah itu laman pancingan data atau bukan. Set data yang digunakan untuk tujuan ini diambil dari laman PhishTank dan Google. Proses ini merangkumi dua fasa, iaitu fasa pra-pemprosesan dan fasa pengesanan, di mana ciri diekstrak berdasarkan peraturan yang telah ditetapkan dalam fasa pra-pemprosesan dan nilai yang terhasil dimasukkan ke algoritma Pokok Keputusan. Hasil dari ujikaji ini, keputuan menunjukkan nilai ketepatan sebanyak 89.40%.

Berdasarkan keputusan yang diperoleh ini, satu sistem akan dicadangkan dalam kertas ini, dimana sistem yang akan dibangunkan bukan sahaja hanya dapat mengektrak kandungan sesetengah laman web bahasa Cina, malah sistem tersebut akan mengesan kewujudan laman web palsu berdasarkan pada ciri-ciri yang terdapat dalam URL sesebuah laman web tersebut. dengan ini, pengesanan akan menjadi lebih meluas kepada seluruh laman web yang ada di internet dan tidak terhad kepada laman web yang berbahasa Cina sahaja.

2.8 RUMUSAN

Setiap penyelidikan yang dilakukan adalah bertujuan untuk mengkaji kaedah yang sesuai bagi mengesan laman web pancingan data. Ini adalah kerana pancingan data

menjejaskan banyak pihak yang menjalankan perniagaan dari pelbagai bidang yang berbeza-beza, seperti e-dagang, perniagaan dalam talian, perbankan dan pemasaran digital. Seperti yang telah dibincangkan dalam bab ini, lazimnya, serangan pancingan data dijalankan dengan mencipta laman web palsu yang menyerupai laman web sebenar. Oleh itu, pelbagai cara dan kaedah telah dikenal pasti oleh para pengkaji terdahulu bagi mengekang serangan ini, antaranya adalah dengan menggunakan algoritma pembelajaran mesin dan juga pembangunan sistem keselamatan.

BAB III

METODOLOGI

3.1 ULASAN BAB

Sepertimana yang telah dibincangkan dalam Bab 1, pancingan data telah menjadi ancaman keselamatan utama yang mengakibatkan kerugian besar terhadap bank dan juga pelanggan. Serangan pancingan data ini semakin meningkat dari hari ke hari kerana kekurangan teknik pengesanan yang cekap dan langkah pencegahan yang berkesan. Teknik pengesanan yang komprehensif harus dibangunkan untuk mengesan dan memaklumkan pengguna web tentang serangan pancingan data bagi memastikan bahawa data sensitif mereka tidak akan bocor semasa serangan ini. Dalam bahagian ini, unit analisis, kaedah dan sebarang reka bentuk instrumen yang digunakan akan diterangkan dengan jelas. Setiap kaedah eksperimen dan sampel data yang digunakan dipilih berdasarkan nasihat daripada penyelia agar sesuai dengan objektif kajian, di mana ia akan mempengaruhi keputusan di akhir kaji selidik ini. Sampel data, algoritma dan teknik pembangunan sistem pengesanan pancingan data akan dihuraikan secara terperinci bagi memastikan keputusan yang tepat dapat diperoleh.

3.2 REKABENTUK PENYELIDIKAN

Kaedah yang digunakan dalam metodologi kajian melibatkan empat fasa iaitu kajian teoretikal, kajian empirikal, pembangunan rangka kerja dan juga penilaian. Setiap fasa mempunyai prosedur yang direkabentuk bagi mencapai ketiga-tiga objektif yang telah dinyatakan dalam Bab 1. Rajah 3.2 dibawah menunjukkan pecahan keempat-empat fasa tersebut, dan setiap satunya akan dihuraikan dengan lebih lanjut dalam seksyen yang seterusnya.

Rajah 3.1 Fasa kaedah penyelidikan

3.3 FASA 1 : KAJIAN TEORETIKAL

Fasa ini menggunakan pendekatan menjalankan kajian kesusasteraan bagi mengkaji kesan-kesan serangan pancingan data terhadap pengguna internet mengenalpasti ciri-ciri laman web palsu yang dibangunkan oleh penggodam dalam usaha memancing data. Pendekatan ini digunakan bagi memastikan kesemua maklumat dan data yang diperlukan tepat dan boleh dipercayai. Ciri-ciri yang terdapat pada laman web palsu yang kerap digunakan oleh para penggodam juga dapat diperincikan melalui kaedah ini, dah kesemua informasi ini adalah dari sumber yang tepat dan terkini. Ciri-ciri yang telah dikenalpasti ini kemudian digunapakai dalam fasa ketiga sebagai tetapan bagi membangunkan sistem pengesan pancingan data.

3.4 FASA 2 : KAJIAN EMPIRIKAL

Perancangan untuk fasa ini dimulakan dengan mengadaptasi dan menambahbaik cadangan sedia ada yang telah cetuskan oleh pengkaji terdahulu. Kajian ini adalah sambungan daripada penyelidikan yang dilakukan oleh Akila D., (2019) yang bertajuk *Phishing Websites Detection Using Machine Learning*, bagi menambah baik cadangan beliau untuk menggunakan kaedah Pokok Keputusan, menerusi platform pembelajaran mesin bagi mengesan pancingan data. Fasa ini juga digunakan bagi mencapai objektif yang ketiga iaitu :

Objektif 1 : Membangunkan sistem untuk mengesan laman web palsu yang boleh digunapakai oleh semua pihak, termasuk pengguna dalam usaha membanteras isu pancingan data terhadap laman pembayaran dalam talian.

3.5 FASA 3 : PERLAKSANAAN RANGKA KERJA

Pelbagai teknik telah digunakan oleh para penyelidik terdahulu bagi mengesan pancingan data, namun serangan ini masih berleluasa dan sukar ditangani. Maka, satu

eksperimen akan dijalankan dengan pendekatan pembelajaran mesin, secara khususnya menggunakan teknik Pokok Keputusan. Fasa ini juga dapat mencapai Objektif 3, dan bagi fasa ini, empat langkah telah dilaksanakan seperti Rajah 3.5 di bawah :

Rajah 3.2 Langkah rekabentuk penyelidikan

3.5.1 Pengumpulan data

Seperti yang telah dinyatakan dalam Bab 1, ciri-ciri dari laman web palsu yang telah dikenalpasti akan diterangkan dengan lebih terperinci di dalam seksyen ini. Kesemua ciri yang disenaraikan diekstrak dari URL dari laman-laman web pancingan data. Set data primer diambil dari laman web Repositori Pembelajaran Mesin UCI atau dikenali sebagai The UCI Machine Learning Repository, mengandungi senarai URL dari laman web *phishing*. Laman ini mengandungi koleksi pangkalan data, teori domain dan penjana data yang digunakan oleh komuniti pembelajaran mesin untuk analisis empirikal algoritma pembelajaran mesin. Set data ini memberi penerangan tentang ciri penting yang telah terbukti kukuh dan berkesan dalam meramalkan tapak web pancingan data. Untuk projek ini, data yang digunakan adalan sampel URL yang telah diklasifikasikan antara jenis sah (1), meragukan (0) dan pancingan data (-1). Data ini mengandungi 11055 contoh URL merangkumi 30 atribut, seperti Jadual 3.5.1.1 di bawah :

Jadual 3.1 Atribut sampel URL set data

Ciri-ciri pada URL	
Alamat IP (*having_IP_Address*)	Menghantar maklumat ke emel (*Submitting_to_email*)
Panjang URL (*URL_Length*)	URL tidak normal (*Abnormal_URL*)
Pemendekkan URL (*Shortining_Service*)	Ubah hala (*Redirect*)
Simbol @ (*having_At_Symbol*)	*Mouse over* (*on_mouseover*)

bersambung...

...sambungan

Pengalihan slash // (*double_slash_redirecting*)	Right Click (*RightClick*)
Nama awal & akhir (*Prefix_Suffix*)	Tetingkap (*popUpWindow*)
Sub domain (*having_Sub_Domain*)	*IFrame Redirection* (*Iframe*)
SSL (*SSLfinal_State*)	Jangka hayat domain (*age_of_domain*)
Jangka hayat domain (*Domain_registeration_length*)	Rekod DNS (*DNSRecord*)
Ikon (*Favicon port*)	Trafik laman web (*web_traffic*)
Token HTTPS (*HTTPS_token*)	Kedudukan laman web (*Page_Rank*)
Domain yang dimuat naik (*Request_URL*)	Indeks Google (*Google_Index*)
Pautan (*URL_of_Anchor*)	Putan ke laman web (*Links_pointing_to_page*)
Pautan dalam tag (*Links_in_tags*)	Laporan statistic (*Statistical_report*)
Server Form Handler (*SFH*)	Keputusan (*Result*)

Bagi menjalankan ujian pembelajaran mesin, ciri-ciri laman web palsu dikenalpasti seperti yang tertera dalam jadual 3.2 di atas. Berikut adalah penerangan tentang ciri tersebut, beserta tetapan 'if else' (rule) bagi mengenal pasti laman web palsu, berdasarkan maklumat penyelidik dari The UCI Machine Learning Repository :

1. Penggunaan Alamat IP (*having_IP_Address*)

   Jika alamat IP digunakan sebagai alternatif nama domain dalam URL, seperti "http://194.9.46.12/index.html", pengguna boleh yakin bahawa seseorang cuba mencuri maklumat peribadi mereka.

   Rule: If

   $$\begin{cases} \text{Jika terdapat Alamat IP dalam URL} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

2. Panjang URL (*URL_Length*)

   Penggodam boleh menggunakan URL panjang untuk menyembunyikan bahagian yang diragui dalam bar alamat.

   Rule: If

$$\begin{cases} \text{Panjang URL} < 54 \rightarrow \text{Laman web sah} \\ else\ if\ \text{Panjang URL} \geq 54 \text{ dan } \leq 75 \rightarrow \text{Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

3. Pemendekkan URL (*Shortining_Service*)

Jika URL terlalu pendek, ia mungkin bertujuan untuk menyembunyikan alamat URL sebenar (*phishing*).

Rule: If

$$\begin{cases} \text{URL terlalu pendek} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{Laman web sah} \end{cases}$$

4. Simbol @ *(having_At_Symbol)*

Menggunakan simbol "@" dalam URL menyebabkan penyemak imbas mengabaikan semua yang mendahului simbol "@".

Rule: If

$$\begin{cases} \text{URL mengandungi simbol @} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{Laman web sah} \end{cases}$$

5. Pengalihan // slash *(double_slash_redirecting)*

Kewujudan "//" dalam URL bermakna pengguna akan diubah hala ke laman web lain. Contoh URL sedemikian ialah: "http://www.lamansah.com//http://www.phishing.com". Maka, jika URL bermula dengan "HTTP", ini bermakna "//" sepatutnya muncul di kedudukan keenam. Walau bagaimanapun, jika URL menggunakan "HTTPS" maka "//" sepatutnya muncul di kedudukan ketujuh.

Rule: If

$$\begin{cases} \text{Kedudukan } // \text{ dalam URL} > 7 \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{Laman web sah} \end{cases}$$

6. Tanda sempang antara nama awal & akhir *(Prefix_Suffix)*

Simbol sempang (-) jarang digunakan dalam URL yang sah. Pemancing data cenderung untuk menambah nama awalan atau akhiran yang dipisahkan oleh (-) pada domain supaya pengguna merasakan bahawa mereka berurusan dengan halaman web yang sah. Contohnya http://www.fake-facebook.com/.

Rule: If

$$\begin{cases} \text{Tanda sempang } (-) \text{ dalam domain} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

7. SSL *(SSLfinal_State)*

Menyemak sijil SSL (SSL certificate) yang mengandungi HTTPS termasuk kesahihan pengeluar sijil dan jangka hayat sijil, kerana jangka hayat minimum sijil adalah dua tahun.

Rule: If

$$\begin{cases} \text{Mempunyai HTTP, sijil sah dan jangka hayat} \geq 1 \rightarrow \text{ Laman web sah} \\ else\ if \text{ Mempunyai HTTP, pengaluar sijil tidak sah} \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

8. Sub domain *(having_Sub_Domain)*

Nama domain mungkin termasuk kod negara, atau lebih dikenali sebagai country-code top-level domains (ccTLD), dan terdapat jugak banyak situasi dimana domain peringkat kedua (SLD) wujud di dalam URL menyebabkan banyak tanda titik (noktah) di dalam satu URL. Untuk mengenalpasti ciri laman web *phishing*, tanda noktah selepas www dalam URL diabaikan dan noktah yang tinggal akan dikira. Jika bilangan tanda noktah lebih daripada satu, maka URL diklasifikasikan sebagai "Meragukan" kerana ia mempunyai satu sub domain. Walau bagaimanapun, jika tanda noktah lebih daripada dua, ia diklasifikasikan sebagai "Pancingan data" kerana ia akan mempunyai banyak bilangan sub domain. Jika tidak, jika URL tidak mempunyai sub domain, ia ditetapkan sebagai "Sah".

Rule: If

$$\begin{cases} \text{Tanda noktah pada domain} = 1 \rightarrow \text{ Laman web sah} \\ else\ if \text{ Tanda noktah pada domain} = 2 \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

9. Jangka hayat domain *(Domain_registeration_length)*

Kebanyakan laman web pancingan data hanya wujud untuk jangka masa yang singkat, maka domain palsu digunakan kebiasaannya hanya bertahan selama satu tahun sahaja.

Rule: If

$$\begin{cases} \text{URL mengandungi simbol @} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

10. Token HTTPS *(HTTPS_token)*

Penggodam boleh menambah token "HTTPS" pada bahagian domain URL untuk menipu pengguna. Sebagai contoh, http://https-facebook.com/.

Rule: If

$$\begin{cases} \text{Terdapat token HTTP dalam domain URL} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

11. Ikon *(Favicon port)*

Favicon ialah imej grafik (ikon) yang dikaitkan dengan halaman web tertentu. Banyak pengguna menunjukkan favicon sebagai peringatan visual identiti sesebuah laman web dalam bar alamat. Jika favicon dimuatkan daripada domain selain daripada yang ditunjukkan dalam bar alamat, maka halaman web itu mungkin dianggap sebagai percubaan *phishing*.

Rule: If

$$\begin{cases} \text{Favicon dimuat turun dari domain lain} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

12. Domain yang dimuat naik *(Request_URL)*

Memeriksa sama ada objek yang terkandung dalam laman web seperti imej, video dan bunyi dimuatkan daripada domain lain. Dalam laman web yang sah, kebanyakan objek yang muat naik dalam domain yang sama.

Rule: If

$$\begin{cases} \%\text{ domain yang dimuat naik} < 22\% \rightarrow \text{ Laman web sah} \\ else\ if\ \ \%\text{ domain yang dimuat naik} \leq 22\% \text{ dan } 61\% \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

13. Pautan *(URL_of_Anchor)*

Jika teg <a> dan laman web mempunyai nama domain yang berbeza, ianya adalah serupa dengan ciri Request_URL diatas.

Rule: If

$$\begin{cases} \% \text{ Pautan ke domain lain} < 31\% \rightarrow \text{ Laman web sah} \\ else\ if\ \ \% \text{ Pautan ke domain lain} \geq 22\% \text{ dan} \leq 61\% \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

14. Pautan dalam teg *(Links_in_tags)*

Adalah perkara biasa bagi laman web yang sah menggunakan teg <Meta>, <Script> dan <Link> dipautkan ke domain halaman web yang sama.

Rule: If

$$\begin{cases} \% \text{ Pautan dalam teg} < 17\% \rightarrow \text{ Laman web sah} \\ else\ if\ \ \% \text{ Pautan dalam teg} \geq 17\% \text{ dan} \leq 81\% \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

15. Server Form Handler *(SFH)*

SFH yang mengandungi pautan kosong atau "about:blank" dianggap meragukan kerana tindakan harus diambil ke atas maklumat yang dimasukkan. Selain itu, jika nama domain dalam SFH berbeza daripada nama domain laman web, ini menunjukkan bahawa laman web itu mencurigakan kerana maklumat yang dimasukkan jarang dikendalikan oleh domain luar.

Rule: If

$$\begin{cases} \text{SFH menunjukkan "}about\text{: }blank\text{"} \rightarrow \text{ Laman web palsu } (Phishing) \\ else\ if \text{ SFH menunjukkan domain luar} \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web sah} \end{cases}$$

16. Menghantar maklumat ke emel *(Submitting_to_email)*

Skrip yang digunakan untuk semua borang atau form dalam laman web diubah ke fungsi "mail()" dalam PHP. Satu lagi fungsi yang mungkin digunakan oleh penggodam untuk tujuan ini ialah fungsi "mailto:".

Rule: If

$$\begin{cases} \text{Terdapat skrip mail() atau mailto:} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

17. URL tidak normal *(Abnormal_URL)*

Untuk laman web yang tulen, identiti biasanya dapat dilihat dari sebahagian daripada URLnya.

Rule: If

$$\begin{cases} \text{Nama hos tiada dalam URL} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

18. Ubah hala *(Redirect)*

Laman web yang sah kebiasaannya diubah hala hanya sekali. Laman web pancingan data diubah hala sekurang-kurangnya 4 kali.

Rule: If

$$\begin{cases} \text{Ubah hala} < 1 \rightarrow \text{Laman web sah} \\ else\ if\ \ \text{Ubah hala} \geq 2 \text{ dan} \leq 4 \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

19. Mouse over *(on_mouseover)*

Penggodam boleh menggunakan JavaScript untuk menunjukkan URL palsu dalam bar status kepada pengguna. "onMouseOver" boleh membantu untuk menyemak sama ada terdapat sebarang perubahan pada bar status.

Rule: If

$$\begin{cases} onMouseOver \text{ menukar bar status} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

20. Right Click *(RightClick)*

Penggodam menggunakan JavaScript untuk melumpuhkan fungsi right click, supaya pengguna tidak boleh melihat kod sumber laman web.

Rule: If

$$\begin{cases} \text{Fungsi } right\ click \text{ lumpuh} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

21. Tetingkap *(popUpWindow)*

Kebiasaannya ciri ini telah digunakan dalam beberapa laman web yang sah dan matlamat utamanya adalah untuk memberi amaran kepada pengguna tentang aktiviti penipuan atau menyiarkan pengumuman alu-aluan, dan tiada maklumat peribadi diminta diisi melalui popup window ini.

Rule: If

$$\begin{cases} Popup\ window \text{ meminta input teks} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

22. IFrame Redirection *(Iframe)*

Penggodam boleh menggunakan teg "iframe" dan menjadikan halaman web tambahan tidak kelihatan.

Rule: If

$$\begin{cases} \text{Mengandungi } iFrame \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

23. Jangka hayat domain *(age_of_domain)*

Kebanyakan laman web pancingan data hidup untuk tempoh yang singkat. Jangka hayat minimum domain yang sah ialah 6 bulan.

Rule: If

$$\begin{cases} \text{Jangka hayat domain} \geq 6 \text{ bulan} \rightarrow \text{Laman web sah} \\ Else \rightarrow \text{ Laman web palsu } (Phishing) \end{cases}$$

24. Rekod DNS *(DNSRecord)*

Jika rekod DNS kosong / tidak dijumpai maka laman web tersebut diklasifikasikan sebagai *phishing*, jika tidak, ia diklasifikasikan sah.

Rule: If

$$\begin{cases} \text{Domain tiada rekod DNS} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

25. Trafik laman web *(web_traffic)*

Memandangkan laman web pancingan data hidup untuk jangka masa yang singkat, ia mungkin tidak dikenali oleh pangkalan data Alexa (Alexa the Web Information Company., 1996). Laman web yang sah berada dalam kedudukan 100,000 teratas. Jika domain tidak mempunyai trafik atau tidak dikenali oleh pangkalan data Alexa, ia diklasifikasikan sebagai *phishing*.

Rule: If

$$\begin{cases} \text{Kedudukan laman web} < 100{,}000 \rightarrow \text{Laman web sah} \\ else\ if\ \text{Kedudukan laman web} > 100{,}000 \rightarrow \text{Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

26. Kedudukan laman web *(Page_Rank)*

PageRank ialah nilai antara "0" hingga "1". Ia bertujuan untuk mengukur betapa pentingnya sesebuah laman web. Semakin tinggi nilai PageRank, semakin penting laman web tersebut. Kira-kira 95% laman web pancingan data tidak mempunyai PageRank. Baki 5% laman web pancingan data mungkin mencapai nilai PageRank sehingga 0.2.

Rule: If

$$\begin{cases} \text{PageRank} < 0.2 \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{Laman web sah} \end{cases}$$

27. Indeks Google *(Google_Index)*

Kebiasaannya, laman web pancingan data hanya boleh diakses untuk tempoh yang singkat, oleh itu ia mungkin tidak berada dalam indeks Google.

Rule: If

$$\begin{cases} \text{Laman web berada dalam indeks Google} \rightarrow \text{Laman web sah} \\ Else \rightarrow \text{Laman web palsu } (Phishing) \end{cases}$$

28. Pautan ke laman web *(Links_pointing_to_page)*

Bilangan pautan yang menghala ke sesebuah laman web menunjukkan tahap kesahihannya. Disebabkan jangka hayatnya yang singkat, 98% laman web pancingan data tidak mempunyai sebarang pautan.

Rule: If

$$\begin{cases} \text{Tiada sebarang pautan} \rightarrow \text{ Laman web palsu } (Phishing) \\ else\ if \text{ Jumlah pautan} > 0 \text{ dan} \leq 2 \rightarrow \text{ Mencurigakan} \\ \text{Else} \rightarrow \text{Laman web sah} \end{cases}$$

29. Laporan statistik *(Statistical_report)*

Beberapa pihak seperti PhishTank dan StopBadware merumuskan banyak laporan statistik mengenai laman web pancingan data pada setiap tempoh masa tertentu.

Rule: If

$$\begin{cases} \text{Hos berada dalam laporan statistik} \rightarrow \text{Laman web palsu } (Phishing) \\ Else \rightarrow \text{ Laman web sah} \end{cases}$$

3.5.2 Pembahagian data

Melalui Pokok Keputusan, set data primer yang diambil dari laman web Repositori Pembelajaran Mesin UCI, dibahagikan kepada dua bahagian, iaitu set latihan dan set ujian (data yang gunakan untuk menguji untuk melihat sejauh mana sistem ini boleh digeneralisasikan). Pengagihan data dibuat dalam nisbah 70:30 :

```
from sklearn.model_selection import train_test_split

Xtrain, Xtest, ytrain, ytest = train_test_split(X, y, test_size=.3)
```

```
Xtrain.shape, Xtest.shape
```

```
((7738, 30), (3317, 30))
```

Rajah 3.3 Pengagihan set data primer

1. 70% data (7738 sampel) digunakan untuk Latihan
2. 30% data (3317 sampel) digunakan untuk Ujian

3.5.3 Membina sistem *Anti-Phishing*

Aplikasi *anti-phishing* dibangunkan daripada sistem yang di import dari platform machine learning, Jupyter Lab 6.4.8. Sistem ini dibangunkan melalui Jupyter Lab, yang kemudiannya diimport ke dalam Python untuk menukarnya menjadi aplikasi yang berfungsi. Pelbagai algoritma pengaturcara digunakan seperti Jadual 3.2 :

Jadual 3.2 Kadar peratusan algoritma yang digunakan

Algoritma	Peratusan (%)
Phyton	44.2
HTML	18.7
CSS	9

Algoritma ini digabung bagi membentuk satu laman web yang membolehkan pengguna memasukkan satu set URL untuk menyemak sama ada ia membawa ke laman pancingan data atau laman web yang sah.

3.5.4 Menguji sistem

Sistem ini diuji dengan memasukkan 100 sampel URL yang diambil dari laman Github (50% URL laman web palsu dan 50% laman web yang sah) dan dinilai peratusan ketepatannya.

3.6 FASA 4 : PENILAIAN

Setelah sistem *anti-phishing* berjaya dibangunkan, satu ujian akan dilaksanakan bagi menguji tahap ketepatan peratusan laman web pancingan data yang berjaya dikesan. Satu ujian simulasi akan dijalankan bagi menguji tahap ketepatan sistem pancingan data ini, dimana sebanyak 100 URL yang diambil dari laman Github akan dimasukkan ke dalam sistem untuk diuji ketepatannya. Seterusnya, setiap keputusan yang diperoleh akan dinilai dan dibincangkan selanjutnya dengan lebih terperinci dalam Bab 4.

3.7 JENIS KAEDAH DAN ALGORITMA

Pokok Keputusan adalah sistem yang digunakan secara meluas untuk tugas klasifikasi dan regresi. Pada asasnya, algoritma ini mempelajari hierarki soalan if/else, yang membawa kepada sesuatu keputusan. Set data yang mempunyai 11055 sampel URL di import ke platform Jupyter, dan data ini dibahagi kepada 70:30 (70% digunakan untuk Latihan dan 30% digunakan untuk Ujian). Hasilnya, sebanyak 7738 URL sampel dikumpulkan untuk diuji manakala 3317 pula dilatih untuk sistem ini.

Hasil ujian dan latihan mendapati bahawa ketepatan keatas set data yang dilatih mencapai 0.991, manakala set yang diuji mencapai sebanyak 0.958 seperti dalam rajah 3.3 dibawah :

```
from sklearn.tree import DecisionTreeClassifier

# instantiate the model
tree = DecisionTreeClassifier()

# fit the model
tree.fit(Xtrain, ytrain)
```

```
from sklearn.metrics import accuracy_score

#computing the accuracy of the model performance
acc_train_tree = accuracy_score(ytrain, y_train_tree)
acc_test_tree = accuracy_score(ytest, y_test_tree)

print("Decision Tree: Accuracy on training Data: {:.3f}".format(acc_train_tree))
print("Decision Tree: Accuracy on test Data: {:.3f}".format(acc_test_tree))
```

```
Decision Tree: Accuracy on training Data: 0.991
Decision Tree: Accuracy on test Data: 0.958
```


Rajah 3.3 Python skrip untuk mengira ketepatan prestasi sistem

Seterusnya, Rajah 3.4 memperlihatkan sistem yang terhasil ini di import ke dalam algoritma Python di mana sistem akan diadaptasikan menjadi sebuah sistem yang berfungsi.

```
# save Decision Tree model to file

import pickle
pickle.dump(tree, open("model.pkl", "wb"))
```

Rajah 3.4 Sistem di import ke dalam file .pkl