UMPUKAN LEMBUT SETEMPAT UNTUK SISTEM PENGESANAN PENCEROBOHAN BERASASKAN HOS: SUATU PERBANDINGAN

NURUL SHAZIRA SAIFUZZAMAN
AZIZI ABDULLAH

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pada masa kini, Sistem Pengesanan Pencerobohan (SPP) memainkan peranan yang penting dalam menangani serangan daripada pengguna hasad. Sistem ini merupakan satu alat teknologi yang digunakan untuk mengesan pelbagai bentuk pencerobohan. Berbanding dengan kaedah traditional iaitu tembok api, SPP berasaskan teknik pengesanan anomali berkebolehan untuk mengesan serangan baru dengan membuat perbandingan kelakuan antara normal dan anomali. Kaedah Pembelajaran Mesin (PM) seperti Pembelajaran Mesin Berpenyelia (PMB) dan Pembelajaran Mesin Tanpa Berpenyelia (PMTB) dilihat mampu menyelesaikan masalah pengelasan dalam SPP berasaskan anomali. Bagi membuktikan penyataan tersebut, set data KDD Cup 99' dipilih sebagai bahan ujikaji di mana ia mempunyai set data sekitar 5 juta rekod dengan 5 kategori iaitu Normal, Dos, U2R, R2L dan Probe. Berikutan set data yang sedia ada terlalu besar dan memerlukan masa yang lama untuk diproses, maka kajian ini dipisah dua kepada SPP berasaskan hos dan rangkaian di mana kajian ini menumpukan pada bahagian hos. Kategori yang terlibat pada bahagian hos ialah Normal, R2L dan U2R. Kaedah PMTB yang popular iaitu teknik kluster digunakan untuk mengecilkan data yang besar kepada beberapa kumpulan dengan menyatukan kemiripan fitur bagi setiap kumpulan yang dikenali sebagai kluster. Tujuan utama kajian ini adalah untuk mendapatkan keputusan pengesanan bagi setiap serangan yang berlaku pada hos dan dibandingkan dengan keputusan pengesanan tanpa melakukan teknik kluster. Bagi mencapai objektif tersebut, pendekatan umpukan lembut digunakan dan SVM untuk membantu membuat pengelasan kategori. Hasil kajian mendapati bahawa pengelasan terhadap serangan menggunakan teknik kluster adalah sebanyak 28.76% manakala pengelasan tanpa menggunakan teknik kluster mendapat keputusan sebanyak 92.85%.

PENGENALAN

Internet kian berkembang ke serata dunia dan telah diguna pakai secara meluas sehingga dunia dikenali sebagai dunia tanpa sempadan. Perkembangan teknologi yang pesat dengan munculnya rangkaian komputer tanpa wayar atau lebih mudah dikenali sebagai 'wifi' menjadikan pengguna lebih terdedah kepada pelbagai serangan dan ancaman terhadap sistem komputer mereka. Laporan tahunan daripada Computer Emergency Response Team (CERT) menerangkan pertambahan bilangan insiden keselamatan komputer setiap tahun (CERT 2002). Jika penggodam atau "hacker" berjaya menyelinap masuk ke dalam sistem sesebuah organisasi maka virus tersebut mampu memberi kesan kepada syarikat atau perniagaan dan mengakibatkan kerugian yang besar. Sistem Pengesanan Pencerobohan (SPP) merupakan satu mekanisme pertahanan sistem komputer daripada diserang pengguna hasad. Ia dipasang di dalam satu sistem rangkaian dan ianya merupakan satu sistem dalam komputer di mana ia boleh mengesan perisian berbahaya seperti virus, cecacing(worm), trojan horse dan rootkit.

SPP dibahagikan kepada dua kategori iaitu *Host-based Intrusion Detection System* (HIDS) dan *Network-based Intrusion Detection System* (NIDS).

Jenis-jenis sistem ini direka bentuk untuk mengumpul dan menguruskan maklumat mengenai aktiviti pada sistem tertentu. Pelayan yang mengandungi maklumat penting adalah sasaran utama untuk percubaan pencerobohan. Sistem Pengesanan Pencerobohan ini terbahagi kepada dua jenis teknik pengesanan iaitu pengesanan berasaskan anomali dan tandatangan. Kedua-dua ini mempunyai kelebihan dan kekurangan masing-masing. Ketepatan pengesanan tandatangan atau nama lain teknik pengesanan penyalahgunaan adalah tinggi tetapi ia tidak boleh mengesan aktiviti serangan baru yang kelakuannya tidak diketahui. Manakala untuk teknik pengesanan anomali, ia mempunyai kadar penggera palsu yang tinggi tetapi ia dapat mengesan serangan baru yang belum dikenali sebelumnya sekiranya terdapat perbezaan dengan kelakuan aktiviti normal.

Projek ini akan menumpukan pada Sistem Pengesanan Pencerobohan berasaskan Hos (HIDS). HIDS ini merujuk kepada pencerobohan yang berlaku pada sistem hos. Kebiasaannya HIDS ini dipasang pada *server* atau pelayan dan akan lebih tertumpu kepada menganalisis sistem pengendalian dan aplikasi, penggunaan sumber dan aktiviti sistem lain yang menetap di hos.

PENYATAAN MASALAH

Sistem Pengesanan Pencerobohon berasaskan Hos (HIDS) telah banyak dibangunkan. Tujuan utamanya bagi mengesan pencerobohan oleh pengguna hasad yang berlaku di hos. Namun begitu, penyelidikan ke atas data laporan yang telah dihasilkan oleh sesuatu SPP masih lagi kekurangan dari sudut kadar ketepatan dan kesilapan ketika membuat keputusan. Antara punca permasalahan sistem tersebut adalah disebabkan oleh pemilihan teknik pengesanan yang tidak sesuai serta menggunkan fitur yang kurang tepat untuk mendeskripsi kegiatan normal dan anomali.

Sesetengah sistem yang menggunakan teknik pengesanan tandatangan tidak memberi keputusan yang memberangsangkan pada keputusan akhir. Hal ini kerana teknik tersebut tidak dapat mengesan sebarang pencerobohan baru atau yang tidak mempunyai corak dipadankan dalam rekod sistem. Tidak seperti teknik pengesanan anomali, ia mampu mengesan sebarang serangan baru tanpa perlu sentiasa mengemaskini rekod pencerobohan dalam sistem yang dibangunkan. Perbezaan penggunaan teknik pengesanan memberi impak yang berbeza pada keputusan pengelasan.

Berdasarkan bahan ujikaji set data KDD Cup 99', terdapat sebanyak 42 attribut termasuk label bagi setiap set data latihan dan pengujian. Atribut-attribut tersebut digunakan untuk menerangkan ciri-ciri serangan, namun tidak semua attribut itu digunakan untuk menerang sesuatu serangan. Berdasarkan kajian literasi, kebanyakan pengkaji menggunakan attribut yang tidak sama bagi setiap kajian dan memberikan setiap keputusan yang berlainan. Jadi, pemilihan fitur yang bersesuaian mendorong kepada peningkatan purata ketepatan.

OBJEKTIF KAJIAN

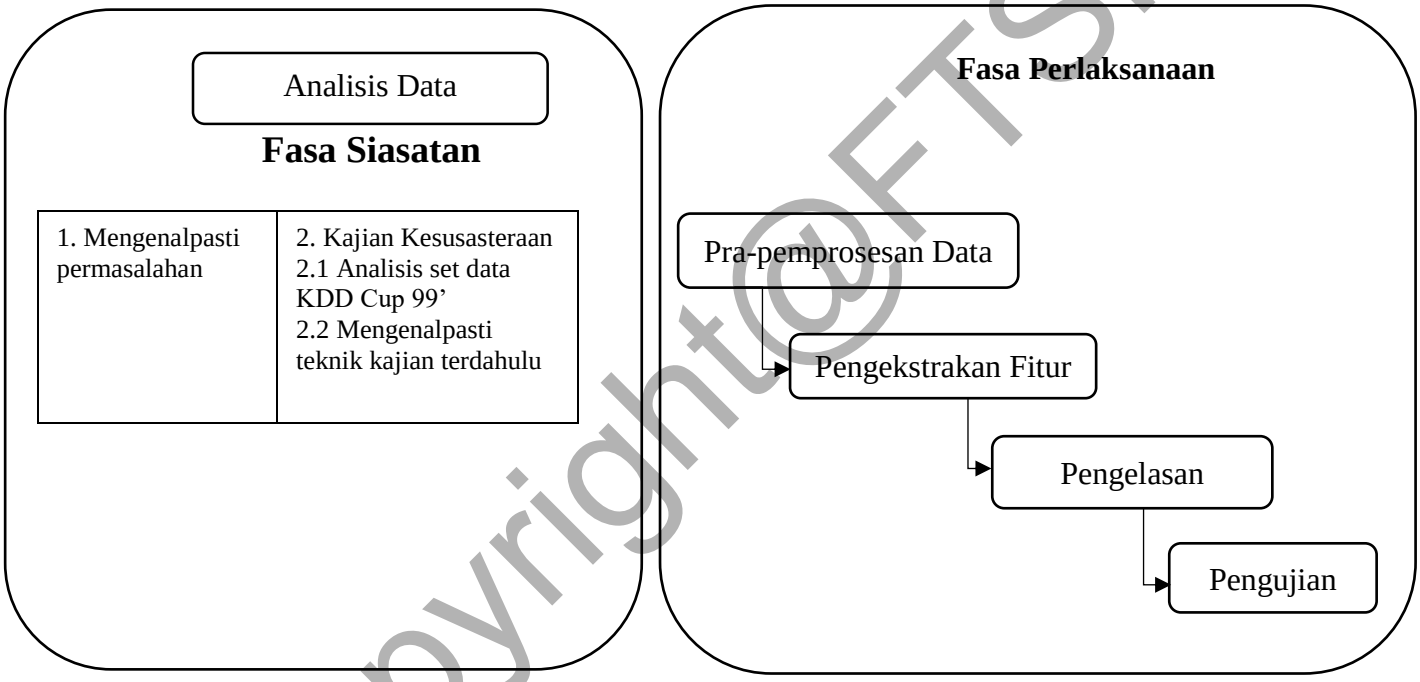
Projek ini bertujuan memperkenalkan satu model ramalan untuk membuat pengesanan serangan dalam sistem komputer kepada orang ramai. Secara umum objektif kajian adalah untuk menghasilkan sistem pengesanan pencerobohan berasaskan hos. Konsep ini dapat meningkatkan kualiti sistem keselamatan komputer yang sedia ada.

Kertas ini membincangkan tentang pembangunan sistem dan menjelaskan bagaimana ia berfungsi. Kajian ini juga bertujuan untuk mengetahui dan mengasingkan atribut atau ciri-ciri untuk serangan yang berlaku di hos. Objektif seterusnya adalah untuk mengkaji

keberkasanan pemilihan teknik pengesanan anomali berbanding teknik lain seperti teknik tandatangan. Model ramalan yang dibangunakan ini juga dikaji terhadap keberkesanan keputusan pengesanan menggunakan kaedah pembelajaran mesin iaitu mesin sokongan vektor (SVM).

METOD KAJIAN

Kaedah yang digunakan semasa penghasilan projek adalah penting bagi memastikan projek berjalan dengan lancar serta sebarang kelemahan dapat dikenalpasti pada peringkat awal perlaksanaan projek. Dalam membangunkan sistem ini, terdapat lima fasa atau peringkat yang terlibat dalam kaedah ini iaitu:



Rajah 1 Model Pembangunan Sistem Pengesanan Pencerobohan berasaskan Hos

I. Analisis Data

Langkah pertama merupakan fasa siasatan di mana permasalahan yang berlaku perlu dikenalpasti telebih dahulu. Setelah mengenalpasti permasalahan, kajian kesusasteraan terhadap masalah tersebut perlu dilakukan dengan menentukan beberapa perkara yang penting. Dalam kajian ini, pemilihan set data KDD Cup sebagai bahan ujikaji di mana analisis terhadap set data dilaksanakan terhadap jenis-jenis serangan yang terkandung di dalamnya. Ini termasuklah menentukan pemilihan set data bagi tujuan latihan dan ujian.

II. Pra-pemprosesan Data

Pemprosesan awal data adalah proses transformasi daripada data mentah kepada bentuk yang bersesuaian supaya analisis secara aritmetik dapat dilakukan. Transformasi data perlu dilakukan disebabkan set data mentah latihan dan pengujian mengandungi sampel yang banyak, bertindan, dan mengandungi nilai aksara. Kaedah Pembelajaran Mesin seperti SVM sukar dilaksanakan ke atas set data mentah tersebut kerana format data tidak bersesuaian dengan format yang ditetapkan. Sehubungan dengan itu, dalam fasa ini melibatkan pengumpulan data dan menformat data untuk dianalisis oleh algoritma pengesanan.

Tapisan data bertindan atau tapisan kelewahan ini diperlukan supaya rekod yang berlebihan atau berulang dapat ditapis. Hal ini kerana, set data latihan terdiri daripada rekod yang sama dan berulang. Bagi menapis kelewahan ini penggunaan kerangka kerja java seperti LinkedHashSet diperlukan. Cara ini juga dikenali sebagai *hash* atau kaedah nilai cincangan. Bagi mengatasi masalah ini, set data latihan baru akan dihasilkan di mana hanya rekod yang unik sahaja dipilih. Kaedah ini digunakan dengan setiap baris rekod diberikan nilai cincangan yang tertentu dan seterusnya dibandingkan dengan baris rekod seterusnya. Jika terdapat baris rekod yang unik nilai cincangannya dengan baris sebelumnya, maka baris rekod tersebut akan dimasukkan ke dalam senarai set data yang baru. Proses ini dilakukan sehingga kesemua baris rekod diproses.

Langkah seterusnya adalah dengan mengasingkan aytribut yang bernilai aksara jenis abjad. Nilai ini perlu diwakili dalam bentuk angka seperti mana yang diperlukan dalam SVM bagi membolehkan pengiraan aritmetik dilaksankan. Dalam kedua-dua set data, didapati attribut *protocol, services* dan *flag* sahaja mengandungi nilai berbentuk aksara abjad**.**

III. Pengekstrakan Fitur

Selepas melaksanakan fasa pemprosesan awal, pengekstrakan fitur perlu dilakukan. Terdapat sebanyak 42 attribut termasuk label dalam set data KDD bagi setiap set data latihan dan pengujian. Pemilihan attribut yang digunakan dan membuang attribut yang tidak digunakan berdasarkan kajian terdahulu sedikit sebanyak membantu dalam kajian ini. Pengurangan attribut tersebut dapat

membantu memudahkan proses dan menjimatkan masa. Seterusnya, membuat analisis keperluan data dengan terperinci.

Dalam fasa ini juga berlaku konsep pengkuantitian vektor di mana ia digunakan untuk menghasilkan model beg fitur. Kaedah Kluster Purata-K merupakan sebahagian daripada algoritma pengkuantitian vektor. Kaedah ini digunakan untuk memecahkan set data yang besar kepada beberapa kluster atau kumpulan yang kecil di mana setiap sampel yang berada dengan niali purata dikumpulkan bersama. Nilai purata bagi setiap kumpulan itu dikenali sebagai sentroid.

Beg fitur diperlukan supaya rajah histogram dapat dibentuk bagi menggambarkan kekerapan taburan. Konsep beg fitur ini adalah untuk memetakan setiap fitur kepada sentroid yang diperolehi semasa pengkuantitian vektor. Terdapat dua jenis kaedah pemetaan antaranya ialah umpukan lembut dan umpukan kasar. Dalam kajian ini akan menggunakan kaedah umpukan lembut di mana umpukan lembut adalah nilai jarak minimum yang diukur akan dipilih bukan sentroid walaupun setiap vektor fitur dipetakan kepada sentroid kluster berdasarkan jarak minimum.

Kegunaan umpukan lembut ini adalah untuk membina dan membentuk fitur yang baru. Penggunaan Jarak Euclidean dipilih supaya nilai jarak yang minimum diambil untuk membentuk set data latihan yang baru untuk kluster setempat yang akan dihasilkan. Nilai jarak yang diperolehi bagi setiap sub-kluster sperti kluster terhadap U2R, R2L dan Normal disusun mengikut tertib menaik bagi memudahkan pemilihan jarak paling minimum.

IV. Pengelasan

Pengelasan adalah salah satu kaedah Pembelajaran Mesin Berpenyelia (PMB) dengan menggunakan kaedah popular iaitu kaedah kepintaran buatan seperti mesin sokongan vektor (SVM). Pengelasan adalah proses mengklasifikasikan data mengikut kriteria yang hampir sama yang telah dilatih.

V. Pengujian

Pada fasa ini, latihan dan pengujian akan dilaksanakan terhadap kaedah sistem pengesanan yang digunakan agar dapat menentukan sama ada memenuhi skop dan objektif yang telah ditetapkan. Fasa penilaian ini juga merupakan peringkat akhir bagi projek ini. Dokumentasi projek akan disiapkan termasuklah penyusunan dan penyediaan maklumat, penulisan ilmiah tentang projek ini.

HASIL KAJIAN

Bahagian ini membincangkan hasil daripada proses pembangunan sistem pengesanan pencerobohan berasaskan hos. Penerangan yang mendalam tentang reka bentuk diperihal. Fasa reka bentuk adalah fasa yang penting dalam pembangunan projek.

Kajian ini tidak menghasilkan reka bentuk yang banyak kerana ia merupakan satu model ramalan sistem pengesanan pencerobohan di mana skop ujikaji hanya tertumpu untuk set data KDD Cup 99'. Set data tersebut mempunyai set data latihan dan set data ujian. Untuk simulasi kajian ini, input yang digunakan adalah set data ujian di mana data ini tidak dilatih seperti set data latihan. Set data latihan dilatih oleh SVM untuk membuat pengelasan bagi setiap kategori serangan atau bukan serangan mengikut kemiripan fitur yang telah digunakan. Jadi, set data ujian hanya digunakan sebagai input untuk membuat pengesanan samada model tersebut memberi keputusan yang betul atau tidak.

Rajah 2 menunjukkan antara muka bagi model ramalan sistem pengesanan pencerobohan di mana pada bahagian kiri merupakan input set data ujian yang digunakan. Apabila satu set data dipilih untuk dikesan kategorinya, maka butang klasifikasi ditekan. Selepas butang klasifikasi ditekan keputusan akan dipaparkan di bahagian kanan atas samada set data tersebut di bawah kategori Normal, R2L atau U2R. SVM akan membuat keputusan samada ia merupakan kategori yang betul atau salah. Jika kategori tersebut adalah salah SVM akan memaparkan kategori yang betul pada bahagian kanan bawah.

Kerja-kerja untuk melatih set data latihan mengikut kategori yang betul berdasarkan fitur-fitur adalah sangat rumit dan mengambil masa yang panjang. Teknik kajian ini memerlukan tumpuan yang tinggi dan teliti kerana untuk melatih set data yang besar adalah sangat sensitif yang perlu penelitian kerja yang tinggi.

Rajah 2 Antara Muka Model Ramalan Sistem Pengesanan

Objektif utama pengujian kajian ini adalah untuk memerhatikan corak dan parameter SVM yang mampu memberikan ketepatan purata yang tinggi bagi kedua-dua set data yang terdiri daripada 15 atribut dan 41 atribut. Pemilihan set data yang mengandungi 15 atribut tersebut adalah berdasarkan kepada pengkaji terdahulu yang menggunakan algoritma Pokok Keputusan Prune berbanding Mesin Sokongan Vektor. Dalam kajian ini, pemilihan atribut tersebut adalah berdasarkan kepada penggunaan algoritma pengelasan yang mempunyai peratusan yang tinggi serta yang menggunakan beberapa atribut untuk menjalankan uji kaji terhadap pengelasan bagi serangan.

Atribut-atribut yang telah dipilih seperti *duration, service, src_bytes, dst_bytes, hot, num_file_creations, srv_count, host_count, host_srv_count, host_diff_srv_rate, host_same_src_port_rate, host_same_src_diff_host_rate, dst_srv_serror_rate, dst_host_srv_serror_rate* dan *host_rerror_rate*. Pemilihan sebanyak 15 atribut ini adalah berdasarkan kepada gabungan antara serangan R2L dan U2R. Kesemua atribut tersebut adalah berdasarkan kepada hasil daripada penggunaan algoritma Pokok Keputusan Prune. Berdasarkan kepada pengkaji terdahulu, keputusan bagi setiap serangan telah didapati sebanyak 80.62% untuk serangan R2L manakala bagi serangan U2R sebanyak 99.92%. Bagi membuat perbandingan tersebut, 15 atribut dan 41 atribut telah digunakan dalam ujian peringkat pertama. Hasil yang telah didapati ialah seperti Jadual 4.3.1 dan 4.3.2.

Jadual 1: Hasil keputusan tanpa menggunakan kluster

Bilangan Atribut	Parameter c	Parameter g	Ketepatan
15	32.0	2.0	92.8548
41	128.0	8.0	94.3981

Berdasarkan kepada keputusan dalam Jadual 4.3.1 didapati bahawa peratusan yang mempunyai ketepatan yang tinggi adalah bilangan atribut yang mempunyai 41 atribut iaitu sebanyak 94.3981% berbanding ketepatan bagi 15 atribut. Pada Jadual 4.3.2 ketepatan peratusan bagi setiap serangan telah di analisa bagi mendapatkan keputusan yang sebenar bagi setiap kategori seperti Normal, R2L dan U2R.

Jadual 2: Nilai parameter berserta peratusan ketepatan yang diperolehi

Bilangan Atribut	Kategori Serangan		
	Normal	R2L	U2R
15	91%	92%	93%
41	93.8%	94.74%	95.6%

Hasil yang diperolehi berdasarkan Jadual 4.3.1 dan 4.3.2, bilangan atribut 41 telah dijadikan rujukan dan di analisa untuk diuji di dalam penggunaan kluster dan umpukan lembut atau ditafsirkan sebagai mencari jarak minimum yang diperolehi untuk sentoid setiap kategori bagi mencari beg fitur sebagai set data untuk diproses di dalam algoritma pengelasan iaitu SVM.

Seterusnya, dapat menghasilkan model ramalan berdasarkan kepada 41 atribut. Pada peringkat kedua ini, 41 atribut telah dipilih kerana mempunyai ketepatan yang lebih tinggi berbanding 15 atribut. Oleh hal demikian, kajian pada peringkat ini menumpukan kepada 41 atribut sahaja untuk diuji ketepatan peratusan yang dimiliki untuk beberapa kluster berdasarkan kepada Jadual 4.3.3 dan 4.3.4

Jadual 3: Ketepatan untuk penggunaan teknik kluster

Bil Attribut	Sentroid	Parameter c	Parameter g	Ketepatan
41	10	0.5	0.0078125	35.8391
41	50	0.125	0.0078125	35.8391
41	100	0.125	0.0078125	35.8391

Pada Jadual 4.3.3 mendapati hasil ketepatan bagi setiap kluster tidak berubah iaitu sebanyak 35.8391% bagi setiap sentroid kategori. Meskipun hasil ketepatan tidak berubah tetapi parameter c berubah dan menghasilkan peningkatan kadar bagi pengesahan berbilang kelas semakin meningkat. Semakin tinggi parameter c dan parameter g, semakin tinggi kadar bagi pengesahan berbilang dan ketepatan peratusan untuk serangan tersebut. Keputusan bagi setiap sentroid memiliki ketepatan yang sama berkemungkinan disebabkan perbezaaan sentroid untuk kluster serangan sangat hampir menyebabkan tidak dapat meningkatkan peratusan ketepatan bagi serangan.

Hasil daripada keputusan peratusan dalam Jadual 4.3.3, didapati bahawa setiap kategori serangan untuk ketiga-tiga kluster tersebut. Jadual 4.3.4 menunjukkan hasil peratusan untuk setiap serangan.

Jadual 4: Nilai parameter berserta peratusan ketepatan yang diperolehi

Bilangan Atribut	**Kategori Serangan**		
	Normal	**DoS**	**Probe**
41	35%	35%	35%
41	35%	35%	35%
41	35%	35%	35%

Peratusan bagi setiap kategori serangan adalah hampir sama selepas membuat kluster. Hal ini terjadi berkemungkinan terdapat sedikit ralat semasa menjalankan kluster. Selain itu, mungkin juga disebabkan penggunaan teknik umpukan lembut untuk mencari jarak minimum menggunakan jarak Euclidean terdapat sedikit masalah semasa mecari nilai jarak tersebut.

KESIMPULAN

SPP adalah satu mekanisme pertahanan sistem komputer yang digunakan untuk mengesan serangan dan ia memberi amaran sekiranya terdapat penceroboh melakukan aktiviti luar biasa. Analisis kelakuan serangan adalah sangat penting dalam dunia sekarang. Oleh itu, dengan masih tiadanya SPP tersebut di dalam sistem komputer, kemungkinan sistem komputer pengguna berada dalam keadaan tidak selamat walaupun sudah mempunyai anti-virus mahupun tembok api. Tembok api bertindak sebagai penghalang antara komputer dalam

rangkaian. Jadi, sistem pengesanan pencerobohan ini perlu dipasang di dalam setiap sistem komputer pengguna agar serangan-serangan dalam sistem komputer dapat dikesan dari semasa ke semasa.

Secara keseluruhannya, projek ini akan memberi penekanan ke atas penyelesaian masalah bagi mengenalpasti dan mengesan penceroboh rangkaian sebenar yang berlaku pada hos serta memberi amaran kepada pengguna sekiranya terdapat pencerobohan. Kajian ini juga diharapkan dapat mencapai objektif serta menjawab persoalan kajian yang telah digariskan di atas.

RUJUKAN

Molina, J., & Cukier, M. (2009). Evaluating Attack Resiliency for Host Intrusion Detection Systems. *Work*, *4*, 1–9.

MyCERT. (2016). Reported Incidents bas ed on General Incident C las s ification Statis tics 2 016, 5802.

Sains, F., Teknologi, D. A. N., Bagi, K., Tumbuhan, J., & Centella, I. (2014). Universiti kebangsaan malaysia, 15328.

Umpukan Lembut Kluster Sejagat Dan Setempat Untuk Sistem Pengesanan Pencerobohan : Satu Kajian Perbandingan. (n.d.).

Bergmark, S. (2015). Development of an Intrusion Detection System. De Boer, P., & Pels, M. (n.d.). Host-based Intrusion Detection Systems.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Aggarwal, P. & Sharma, S. K. 2015. Analysis of Kdd Dataset Attributes - Class Wise for Intrusion Detection. *Procedia Computer Science* 57(842-851.

Darpa. 1999. Kdd Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, https://kdd.ics.uci.edu/databases/kddcup99/task.html

Warusia Yassin, Nur Izura Udzir, Zaiton Muda & Md. Nasir Sulaiman. 2013. Anomaly-Based Intrusion Detection through K-Means Clustering and Naives Bayes Classification.

"What Is Network Traffic? - Definition from Techopedia." n.d. https://www.techopedia.com/definition/29917/network-traffic.