PENGKELASAN CORAK PERISIAN HASAD MENGGUNAKAN RANGKAIAN NEURAL PERLINGKARAN BERULANG

MUDZFIRAH ABDUL HALIM

UNIVERSITI KEBANGSAAN MALAYSIA

PENGKELASAN CORAK PERISIAN HASAD MENGGUNAKAN RANGKAIAN NEURAL PERLINGKARAN BERULANG

MUDZFIRAH ABDUL HALIM

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEHI IJAZAH SARJANA SAINS KOMPUTER (TEKNOLOGI RANGKAIAN)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

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

14 Disember 2018

MUDZFIRAH ABDUL HALIM
P87569

PENGHARGAAN

Pertamanya, alhamdulillah syukur ke hadrat Allah S.W.T dengan izinNya dapat saya selesaikan kajian ini yang menjadi impian saya sejak memulakan perjalanan menimba ilmu di UKM. Terima kasih Allah S.W.T atas kasih sayang dan rahmat kerana memberi kefahaman, ilham, masa dan kesihatan yang baik untuk menyiapkan kajian ini. Engkau Tuhan yang memiliki ilmu pengetahuan yang mengajar kami sesuatu yang tidak diketahui. Sesungguhnya Engkau Tuhan Maha mengetahui lagi Maha bijaksana. Salam dan selawat ke atas junjungan Nabi Muhammad S.A.W.

Setinggi-tinggi penghargaan saya ucapkan kepada penyelia utama saya, Prof. Madya Dr. Azizi Abdullah atas bimbingan yang tidak berbelah bagi serta idea bernas sepanjang saya menuntut ilmu di UKM. Komitmen dan kerjasama Dr amatlah dihargai. Saya doakan semoga Allah S.W.T permudahkan segala urusan Dr sekeluarga sebagaimana Dr telah mudahkan urusan saya sepanjang proses penyeliaan berlangsung. Semoga Dr. sekeluarga sentiasa dimurahkan rezeki dan dirahmati Allah S.W.T. Saya juga ingin merakam ucapan terima kasih yang tidak terhingga kepada penyelia bersama saya, Dr. Khairul Akram Zainol Ariffin atas sokongan, dorongan serta motivasi yang diberikan. Segala jasa dan bimbingan amatlah saya hargai. Semoga Dr. sekeluarga sentiasa berada di bawah lindungan Allah S.W.T. Ucapan penghargaan terutama kepada Unit CAIT serta seluruh warga FTSM yang membantu menyediakan kemudahan pembelajaran semasa proses kajian ini.

Kepada kedua ibu bapa tercinta, Rohiah Taib dan Abdul Halim Razali, terima kasih atas kasih sayang, sokongan moral, kata semangat, titipan doa yang tidak pernah putus serta wang ringgit yang diberikan daripada mula kak cik mendaftar sehingga menyiapkan kajian ini. Tidak mampu diungkapkan betapa kak cik menghargai dan berterima kasih atas segala apa yang telah umi dan abah berikan. Boleh jadi umi dan abah bukan sesiapa bagi orang lain tapi ketahuilah bahawa umi dan abah adalah malaikat hidup kak cik. Semoga dengan cinta Allah S.W.T, kak cik doakan syurga sebagai ganjaran buat umi dan abah. Tidak lupa juga terima kasih setulusnya kepada kakak Mahfuzah, kak ngah Masturah, ateh Maftuhah dan adik Mumtazah kerana banyak membantu, mendoakan dan mendukung cita-cita kak cik.

Akhir sekali, saya ucapkan terima kasih kepada semua yang terlibat dalam menjayakan kajian ini secara langsung mahupun tidak langsung. Semoga ilmu yang diperoleh menjadi manfaat dan membawa barakah kepada semua.

"Kerana sesungguhnya sesudah kesulitan itu ada kemudahan, sesungguhnya sesudah kesulitan itu ada kemudahan". Q.S Al-Insyirah: 5-6)

ABSTRAK

Pembangunan teknologi rangkaian yang pesat telah memberi ilham kepada serangan siber seperti perisian hasad. Serangan ini menjadi salah satu ancaman besar kepada pengguna dan organisasi rangkaian. Ancaman ini menunjukkan perkembangan yang semakin aktif disebabkan oleh pertambahan serangan perisian hasad berubah-ubah. Oleh sebab itu, pelbagai kebaharuan pembangunan algoritma telah dicadangkan untuk mengesan serangan perisian hasad. Walau bagaimanapun, ia masih menghadapi masalah untuk membina model yang boleh dipercayai dan tepat yang berupaya menangani kuantiti data besar dengan corak sentiasa berubah-ubah. Perwakilan fitur menggunakan model Beg Perkataan (BP) adalah perwakilan yang biasa digunakan untuk mewakili kekerapan serangan perisian hasad. Namun begitu, menggunakan model BP memusnahkan aspek ruangan dan temporal corak serangan lalu mengakibatkan kehilangan maklumat dan pengindeksan yang kasar. Justeru itu, kajian ini menggunakan kaedah yang menggabungkan Rangkaian Neural Perlingkaran (RNP) dan Memori Jangka Masa Panjang dan Pendek (MJMPP) dicadangkan untuk mengkelas corak perisian hasad. Model RNP dan MJMPP digunakan dalam kajian ini untuk mengatasi masalah ruangan dan temporal untuk model BP. Hasil pengujian menunjukkan cadangan gabungan model RNP-MJMPP dan MJMPP-RNP mengatasi model tunggal Rangkaian Neural Berulang (RNB), RNP dan MJMPP dalam tugas mengkelas perisian hasad. Model cadangan RNP-MJMPP dan MJMPP-RNP masing-masing memperolehi tahap ketepatan pengkelasan sebanyak 96.76% dan 98.53% ke atas set data Drebin.

CLASSIFYING MALWARE PATTERNS USING RECURRENT CONVOLUTIONAL NEURAL NETWORK

ABSTRACT

Recently, the rapid development of netwok technology has brought inspiration to various new cyber attacks such as malware. The attacks have become one of the biggest threats to user and organization's network. The threat shows an active growth due to the increasing variation of malware attacks. Thus, a large number of novel algorithms have been proposed for attack detection. However, it still facing the problem of building realiable and accurate models that are capable of handling large quatities of data and with changing patterns. The most commonly used feature to represent attack patterns is bag-of-words (BOW) where the frequency of each word is used for attack description. However, using such approach destroys the spatial and temporal information aspects of the attack patterns, resulting in information loss and coarse indexing. In this research, a method that combines Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) is proposed to classify malware patterns. The combination of CNN and LSTM is used to overcome the problem of spatiality and temporal information for the BOW. The experiment shows that the proposed combination of CNN-LSTM and LSTM-CNN model outperforms the single Multi-Layer Perceptron (MLP), CNN and LSTM neural network models on malware classification tasks. The proposed model CNN-LSTM and LSTM-CNN obtain state-of-the-art performance accuracy by 96.76% and 98.53% on the Drebin dataset respectively.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI RAJAH

SENARAI SINGKATAN

CAIT	Centre of Artificial Intelligence Technology
FTSM	Fakulti Teknologi dan Sains Maklumat
UKM	Universiti Kebangsaan Malaysia
RNB	Rangkaian Neural Buatan
RNP	Rangkaian Neural Perlingkaran
RNN	Rangkaian Neural Berulang
MJMPP	Memori Jangka Masa Panjang Pendek
PM	Pembelajaran Mesin
PMTP	Pembelajaran Mesin Tanpa Penyelia
PMB	Pembelajaran Mesin Berpenyelia
BP	Beg Perkataan

GLOSARI

Algoritma Pembelajaran Mendalam	Deep Learning Algorithm
Perisian hasad	Malware
Algoritma Rambatan Balik	Back Propagation algoritm
Rangkaian Neural Buatan (RNB)	Artificial Neural Network (ANN)
Rangkaian Neural Perlingkaran (RNP)	Convolutional Neural Network (CNN)
Rangkaian Neural Berulang (RNN)	Recurrent Neural Network (RNN)
Memori Jangka Masa Panjang Pendek (MJMPP)	Long Short Term Memory (LSTM)
Pembelajaran Mesin Tanpa Penyelia (PMTP)	Unsupervised Machine Learning
Pembelajaran Mesin Berpenyelia (PMB)	Supervised Machine Learning

BAB I

PENGENALAN

1.1 PENGENALAN

Bab ini membincangkan tentang keseluruhan kajian yang ingin dilaksanakan dan dihurai dalam beberapa bahagian subtopik iaitu berkenaan latar belakang kajian, pernyataan masalah, persoalan kajian dan objektif kajian. Selain daripada itu, skop serta metodologi kajian juga turut dibincangkan. Latar belakang kajian merangkumi segala aspek yang terlibat secara langsung dan tidak langsung dalam kajian yang telah dilaksanakan bagi mengupas tentang kajian ini. Pernyataan masalah pula menerangkan faktor yang mendorong kajian ini dilaksanakan hasil daripada permasalahan yang timbul dalam kajian lepas. Seterusnya, bab ini membincangkan dan menyingkap persoalan serta objektif yang ingin dicapai melalui kajian ini. Selain itu, skop kajian menerangkan tentang apa yang dinilai dan ingin dicapai dalam kajian ini. Metodologi kajian pula membincangkan secara terperinci tentang aktiviti yang terlibat dalam kajian ini bagi mendapatkan jawapan kepada persoalan dan objektif kajian yang ingin dicapai.

1.2 LATAR BELAKANG KAJIAN

Seiring dengan kepesatan dan kemajuan arus dunia teknologi rangkaian hari ini, insiden keselamatan siber berkaitan dengan perisian hasad juga semakin meningkat. Hal ini mengancam ramai pihak terutama mereka yang menggunakan telefon pintar di mana maklumat peribadi pengguna boleh dicapai dari seluruh dunia (Weber & Studer 2016). Terma perisian hasad ini merujuk kepada satu perisian yang mempunyai niat jahat yang ditujukan ke atas sistem komputer tertentu. Ia direka khas untuk mengawal dan melakukan penipuan ke atas sistem. Perisian hasad pertama yang ditemui adalah

pada tahun 1986 dan diberi nama sebagai Brain (Joseph & Fics 1997). Berikutan daripada fenomena itu, motivasi membangunkan perisian hasad telah bertukar daripada hanya sebagai hiburan kepada satu tujuan yang mendatangkan manfaat kepada pengaturcaranya (Pascanu et al. 2015).

Penyelidikan yang luas telah dilakukan berhubung pengkelasan perisian hasad menggunakan pelbagai teknik dan salah satu daripadanya adalah melalui teknik Pembelajaran Mesin (PM) (Dai et al. 2015; Damodaran et al. 2015; Niyaz et al. 2015; Dong & Wang 2016). Pengkelasan perisian hasad menggunakan PM mempunyai dua kaedah iaitu Pembelajaran Mesin Tanpa Penyelia (PMTP) dan Pembelajaran Mesin Berpenyelia (PMB). Pembelajaran mendalam teknik PM dipercayai dapat meningkatkan tahap abstraksi dalam data dengan menggunakan seni bina yang kompleks beserta komposisi transformasi bukan linear (Kim et al. 2016). Algoritma Rangkaian Neural (RN) adalah satu model pengiraan automatik di bawah teknik PM yang popular (Nix & Zhang 2017). Algoritma ini terdiri daripada beberapa lapisan untuk mempelajari perwakilan data dan mentafsirkannya kepada suatu kumpulan kelas. Algoritma RN kebiasaannya membaca set data yang besar dengan menggunakan algoritma Rambatan Balik yang menunjukkan cara bagaimana mengubah parameter dalaman di setiap lapisan yang diambil daripada lapisan yang sebelumnya.

Terdapat pelbagai pendekatan yang telah diaplikasi dalam kajian lepas berkaitan pengkelasan perisian hasad menggunakan algoritma RN. Pemilihan fitur dan bilangan fitur untuk digunakan sebagai input fitur pengkelasan adalah berbeza mengikut kajian. Pemilihan fitur dan penggunaan bilangan fitur memberi kesan kepada prestasi model pengkelasan yang dibangunkan. Hal ini kerana pemilihan dan dan penggunaan fitur yang tidak perlu menyebabkan pengesanan negatif palsu yang tinggi (Zhou et al. 2012). Kajian Feizollah et al. (2017) menggunakan dua jenis fitur perisian untuk membuat pengkelasan dan mencapai ketepatan prestasi sebanyak 91% manakala kajian Drebin (Arp et al. 2014) yang menggunakan lapan fitur mencapai 94% ketepatan pengkelasan. Penggunaan bilangan fitur yang lebih banyak menambah maklumat pelengkap mengenai fitur yang ada pada sesebuah perisian. Selain daripada itu, isu perwakilan fitur perisian hasad juga menjadi topik kajian penting dalam bidang

ini kerana prestasi pengkelasan dipengaruhi oleh perwakilan fitur yang digunakan (Wu & Rehg 2011). Hal ini bermakna, perwakilan fitur adalah penting dalam proses membina model pengkelasan perisian hasad yang berkesan. Antara model yang terkenal dalam perwakilan fitur perisian adalah model Beg Perkataan (BP) yang menunjukkan prestasi pengkelasan yang memberangsangkan (Hughes et al. 2017; Winn et al. 2005). Model BP menerangkan fitur perisian dengan menggunakan kekerapan fitur dalam sesebuah perisian. Terdapat dua perwakilan fitur yang dapat dihasilkan menggunakan model BP iaitu perwakilan kewujudan fitur yang dikenali sebagai perwakilan perduaan dan perwakilan kekerapan fitur yang dikenali sebagai perwakilan vektor.

Kajian yang berkaitan penggabungan pengkelas untuk membuat pengkelasan telah menjadi topik kajian yang popular dengan kemunculan pelbagai pengkelas termasuklah Rangkaian Neural Perlingkaran (RNP) dan Memori Jangka Masa Panjang dan Pendek (MJMPP). Sekalipun begitu, susunan perlapisan antara pengkelas juga memainkan peranan penting dalam membuat gabungan model pengkelasan. Kajian Sainath et al. (2015) menggabungkan dua model pengkelas dengan meletakkan RNP di lapisan awal agar masalah ruangan fitur dapat diatasi terlebih dahulu. Sementara itu, kajian Xu, Li & Deng (2016) pula mendahulukan MJMPP di awal lapisan model pengkelasan agar ciri temporal data dapat dipelajari sebelum dihantar kepada RNP untuk meningkatkan signal ruangan data. Dalam kajian ini, susunan perlapisan model pengkelasan akan diukur keberkesanannya melalui prestasi ketepatan pengkelasan model.

1.3 PERNYATAAN MASALAH

Dalam mengkelas perisian hasad, terdapat tiga topik utama yang difokuskan dalam kajian ini. Isu yang pertama adalah isu pemilihan bilangan fitur perisian yang ingin digunakan sebagai input model pengkelasan. Pemilihan fitur dilaksanakan di awal pemprosesan data untuk mewakili ciri-ciri perisian. Secara umumnya, penggunaan fitur yang lebih banyak akan menghasilkan maklumat yang lebih jelas dan nyata bagi tujuan pengkelasan perisian hasad. Walau bagaimanapun, pemilihan bilangan fitur juga harus dilakukan berdasarkan kepentingan fungsi fitur itu sendiri dalam perisian

bagi membantu meningkatkan pengkelasan (Mazlan & Hamid 2018). Oleh yang demikian, bilangan fitur yang digunakan memainkan peranan penting dalam mempengaruhi ketepatan pengkelasan model. Bilangan fitur yang lebih banyak digunakan sebagai input menjadikan prestasi ketepatan pengkelasan lebih tinggi dan berkesan.

Isu seterusnya adalah isu perwakilan fitur perisian hasad menggunakan model Beg Perkataan (BP). Perwakilan fitur digunakan bertujuan untuk membolehkan set data diproses oleh algoritma PM. Meskipun begitu, terdapat cabaran dalam kedua-dua perwakilan perduaan dan perwakilan vektor model BP. Perbezaan antara perwakilan kehadiran dan kekerapan fitur memberi impak kepada keterhadan model pengkelas mempelajari fitur perisian dengan berkesan (Grosse et al. 2017). Perwakilan perduaan mempersembahkan nilai perduaan kewujudan fitur manakala perwakilan vektor mempersembahkan maklumat kekerapan fitur hadir dalam perisian. Penggunaan perwakilan perduaan hanya terhad kepada kewujudan fitur serta tertumpu kepada nilai 0 dan 1 sahaja. Oleh yang demikian, penggunaan perwakilan vektor adalah langkah untuk menjadikan pengkelasan lebih peka dan berkesan apabila maklumat kewujudan fitur dinyatakan dengan terperinci (Kim et al. 2017). Prestasi ketepatan pengkelasan lebih tinggi apabila menggunakan perwakilan vektor yang memiliki penerangan fitur aplikasi yang lebih deskriptif berbanding perwakilan perduaan.

Prestasi gabungan model pengkelasan perisian hasad adalah isu ketiga yang menjadi perhatian kajian ini. Perwakilan fitur yang digunakan dalam kajian ini menyebabkan pengkelasan perisian hasad tidak dapat dilaksanakan di tahap yang paling optimal. Hal ini demikian kerana model BP memusnahkan maklumat ruangan dan temporal fitur dan secara tidak langsung membataskan kuasa deskriptif fitur (Xu et al. 2015). Hubungan maklumat ruangan dan temporal adalah penting bagi tujuan pengkelasan kerana maklumat ini menjadi penghubung antara fitur. Maklumat ini juga membantu untuk lebih memahami bagaimana sesebuah perisian berkait antara satu sama lain dan membolehkan pengkelasan dibuat dengan membandingkan maklumat fitur (Zhang & Mayo 2010). Sehubungan dengan itu, RNP dan MJMPP adalah pengkelas yang paling sesuai dan menepati apa yang dijanjikan oleh algoritma RN dalam mengatasi masalah ruangan dan temporal perwakilan fitur model BP (Xu et al.

2016). Bagi masalah kemusnahan ruangan fitur, RNP mengenal pasti dan mempelajari maklumat antara fitur menggunakan penapis perlingkaran fitur tempatan perisian. MJMPP pula membuat pemerhatian pada satu masa di mana input adalah variasi dan menghasilkan output yang dikumpulkan daripada keseluruhan input. Bersesuaian dengan masalah temporal data perwakilan fitur model BP, MJMPP yang mempunyai satu ciri dalaman akan sentiasa berwaspada dengan kehadiran struktur temporal walaupun telah dimusnahkan. Namun begitu, prestasi ketepatan model gabungan yang dibina bergantung kepada susunan perlapisan yang dibuat.

1.4 PERSOALAN KAJIAN

Persoalan kajian berperanan penting untuk membentuk dan memberi fokus yang jelas kepada kajian ini. Persoalan kajian adalah seperti berikut:

1. Di antara penggunaan lapan fitur dan sepuluh fitur sebagai input, yang manakah menunjukkan prestasi pengkelasan yang lebih baik?
2. Apakah metodologi yang digunakan untuk menghasilkan perwakilan fitur untuk membina model pengkelasan perisian hasad?
3. Di antara perwakilan fitur menggunakan set data perduaan dan set data vektor yang dihasilkan oleh model BP, yang manakah menunjukkan prestasi pengkelasan yang lebih baik?
4. Adakah gabungan model RNP-MJMPP dan model MJMPP-RNP menunjukkan prestasi pengkelasan yang lebih baik dan boleh dipercaya untuk menangani masalah ruangan dan temporal?

1.5 OBJEKTIF KAJIAN

Objektif kajian menerangkan matlamat dan hala tuju kajian. Kajian ini membangunkan model pengkelasan yang menyelesaikan masalah ruangan dan temporal data dalam membuat pengkelasan perisian hasad. Objektif kajian adalah seperti berikut:

OBJ 1) Mengenal pasti keberkesanan bilangan fitur perisian yang digunakan sebagai input pembinaan model pengkelasan perisian hasad.

OBJ 2) Menganalisis keberkesanan penggunaan perwakilan fitur model BP bagi pembinaan model pengkelasan perisian hasad.

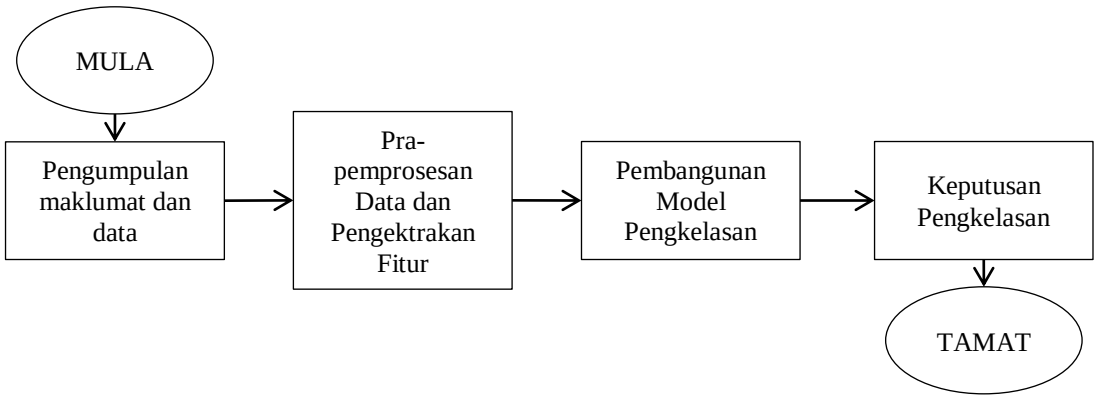
OBJ 3) Menguji prestasi pengkelasan perisian hasad menggunakan pendekatan gabungan model RNP-MJMPP dan model MJMPP-RNP dalam mengatasi masalah ruangan dan temporal data.

1.6 SKOP KAJIAN

Skop kajian ini memfokuskan kepada pengkelasan perisian hasad menggunakan Algoritma RN. Algoritma RNP-MJMM dan MJMPP-RNP dipilih untuk mengkelas set data fitur Drebin menggunakan teknik PMB. Set data Drebin melalui pra-pemprosesan data menggunakan model BP untuk menukar nilai perkataan fitur kepada nilai vektor. Model pengkelasan perisian hasad yang dibina mengkelas perisian samada ianya perisian hasad atau perisian normal dan membandingkannya dengan label sedia ada. Prestasi ketepatan model pengkelasan perisian hasad dibandingkan dengan model algoritma RN yang lain.

1.7 METODOLOGI KAJIAN

Dalam kajian ini, terdapat beberapa aktiviti utama yang dijalankan seperti yang telah diringkaskan dan diterjemah melalui Rajah 1.1.



Rajah 1.1 Metodologi kajian

a. Langkah 1: Pengumpulan Maklumat dan Data

Kajian kesusasteraan dilakukan di permulaan kajian untuk mengetahui dan mencari isu semasa berkenaan dengan pembangunan terkini pengkelasan perisian hasad. Set data yang dipilih untuk digunakan sebagai bahan ujikaji dalam kajian ini adalah set data Drebin. Set data ini adalah set data aplikasi perisian Android yang banyak digunakan dalam kajian lepas bagi tujuan pengkelasan perisian hasad. Bilangan asal set data ini mempunyai 129,013 sampel perisian yang mempunyai ciri-cirinya tersendiri. Menggunakan sumber perisian hasad set data Drebin, pengkelasan kelas perisian akan dilaksanakan keatas fitur perisian dan membandingkannya dengan label sedia ada.

b. Langkah 2: Pra-pemprosesan Data dan Pengekstrakan Fitur

Set data Drebin perlu melalui fasa pra-pemprosesan data di mana tapisan dijalankan bagi menyingkir lewahan data. Seterusnya, pentokenan dilaksanakan bagi mengenal pasti fitur yang terdapat di dalam set data fitur Drebin. Fitur yang dikenal pasti untuk digunakan di dalam kajian ini kemudiaannya diekstrak melalui model BP. Model ini melakukan transformasi atribut daripada perkataan kepada angka. Transformasi format fitur perlu dijalankan agar pengiraan aritmetik boleh dilaksanakan.

c. Pembangunan Model Pengkelasan

Sebelum pembangunan model pengkelasan dijalankan, terdapat beberapa ekperimen dijalankan bagi mendapatkan parameter yang paling optimum untuk hasil pengkelasan yang terbaik. Perpustakaan perisian dipilih bersesuaian dengan keperluan pembangunan model ini supaya tidak diganggu dengan pepijat yang sukar dibaiki. Bahasa pengaturcaraan python versi 3.6.3 digunakan untuk menulis skrip model pengkelasan dalam kajian ini.

d. Keputusan Pengujian

Keputusan setiap latihan dan pengujian yang dijalankan oleh model direkod dan perbandingan akan dilakukan untuk mengenal pasti kaedah yang paling tepat dalam membuat pengesanan keatas perisian hasad. Proses pengiraan dan perbandingan akan disertakan bagi kesemua model. Setelah itu, kesimpulan akan dibuat bagi merumuskan dapatan semasa kajian.

1.8 KEPENTINGAN KAJIAN

Kajian ini mampu memberi manfaat di dalam bidang akademik dan menyumbang kepada jalan penyelesaian bagi industri pengesanan perisian hasad di dunia sebenar. Kajian yang berterusan diperlukan bersesuaian dengan peredaran teknologi yang sentiasa berkembang dari masa ke semasa.

1.9 ORGANISASI PENULISAN

Secara keseluruhan, kajian ini mengandungi enam bab. Bab pertama ini membincangkan maklumat umum kajian seperti pengenalan terhadap kajian, pernyataan masalah, objektif kajian, persoalan kajian, skop kajian, dan kepentingan kajian. Perbincangan lanjut mengenai kajian ini dibincangkan secara terperinci dalam bab lain seperti berikut:

Bab II mengandungi kajian kesusasteraan tentang kajian lepas yang berkaitan dengan perisian hasad, dan metodologi yang digunakan untuk mengkelas perisian hasad. Kajian ini akan tertumpu kepada kaedah yang telah dilaksanakan serta cabaran dan kelemahan yang dijumpai. Dapatan maklumat daripada kajian kesusasteraan diguna sebagai panduan dan hala tuju kajian ini.

Bab III menerangkan metodologi yang digunakan sepanjang kajian ini secara terperinci. Dalam bab ini juga menghuraikan tentang seni bina kajian, kerangka kerja kajian, reka bentuk eksperimen, penilain prestasi serta alatan kajian bagi pembangunan model pengkelasan perisian hasad.

Bab IV menerangkan kaedah pemprosesan awal data dan pengekstrakan fitur yang diguna dalam kajian ini.

Bab V membincangkan secara keseluruhan proses pembinaan pengkelas serta analisis terhadap hasil keputusan yang diperolehi melalui fasa pengujian eksperimen. Ketepatan model pengkelas mengikut set data perduaan dan set data vektor akan dinilai dan perbandingan antara pengkelas akan dibincangkan. Perbandingan prestasi model gabungan membuat pengkelasan juga turut dihuraikan.

Bab VI menbincangkan sumbangan kajian dan merumuskan keseluruhan kajian yang dijalankan. Cadangan penambah baikan pada masa hadapan disenaraikan bagi tujuan perluasan kajian di masa hadapan.

1.10 RUMUSAN

Bab ini membincangkan latar belakang kajian, pernyataan masalah, persoalan kajian, objektif kajian dan skop kajian. Seterusnya, bab ini juga menceritakan tentang metodologi kajian secara ringkas serta kepentingan melaksanakan kajian ini. Kajian ini bermatlamat untuk mengkaji keberkesanan penggunaan bilangan fitur yang berbeza ke atas prestasi pengkelasan. Selain daripada itu, kajian ini juga bertujuan untuk mengkaji keberkesanan perwakilan fitur yang diproses menggunakan model BP. Di samping itu, tujuan kajian ini juga dilaksanakan adalah untuk mengkaji dan membandingkan prestasi pengkelasan perisian hasad menggunakan pendekatan gabungan model RNP-MJMPP dan model MJMPP-RNP dalam mengatasi masalah ruangan dan temporal data. Diharapkan kajian ini dapat memenuhi segala objektif yang dinyatakan dan menjawab semua persoalan kajian yang telah digariskan dalam bab ini. Dalam bab seterusnya, kajian kesusasteraan berkaitan pengkelasan perisian hasad, teknik PM, algoritma RN serta keterangan lanjut set data Drebin yang diguna untuk kajian ini diperincikan.

BAB II

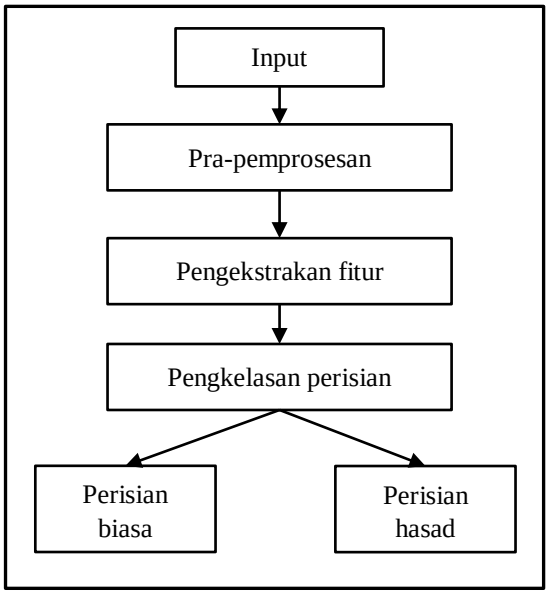
KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Secara amnya, bab ini akan memberi kefahaman yang lebih jelas tentang perisian hasad serta teknik pengkelasannya. Di samping itu, bab ini juga memberi ulasan daripada kajian kerja lepas bersesuaian dengan objektif yang telah dijelaskan dalam Bab 1. Kajian kesusasteraan penting untuk memperoleh maklumat latar belakang mengenai topik kajian dan pencapaian yang telah dilakukan melalui pembacaan dan analisis yang kritis. Perbincangan bab ini dimulakan dengan penerangan berkenaan pengkelasan perisian hasad di bahagian 2.2. Seterusnya, kajian lepas berkenaan pengkelasan perisian hasad dihuraikan.

2.2 PENGKELASAN PERISIAN HASAD

Kajian ini berkisar tentang pengkelasan perisian hasad menggunakan kaedah PM. Perisian hasad adalah satu terma untuk sebarang program perisian yang mempunyai hasad terhadap sistem kemudiannya bertindak melakukan tindakan yang tidak di ingini tanpa kebenaran pemilik dan menyumbang kepada kerosakan serius (Kramer & Bradfield 2010). Perisian hasad direka oleh penggodam topi hitam yang mempunyai pengetahuan mendalam tentang pengaturcaraan komputer yang pada kebiasaannya dicipta bagi kepentingan pihak ketiga. Perisian ini mendapat akses ke dalam peranti dengan mudah melalui rangkaian Internet, emel, aktiviti muat turun aplikasi dan sebagainya. Perisian hasad mampu menimbulkan ancaman kepada keselamatan rangkaian. Rajah 2.1 menunjukkan proses asas yang terlibat untuk membuat pengkelasan perisian hasad menggunakan teknik PM.



Rajah 2.1 Proses asas dalam sistem pengkelasan perisian hasad. Input perisian akan melalui pem-prosesan awal data dan pengekstrakan fitur. Kemudian, pengkelasan dilaksanakan untuk menentukan perisian adalah perisian normal atau perisian hasad.

Bidang kajian pengkelasan perisian hasad banyak dikaji dalam kajian lepas. Di peringkat pra-pemprosesan data, fitur input ditapis dan dianalisis. Pra-pemprosesan data mengesan informasi penting daripada input perisian untuk diekstrak dan digunakan dalam peringkat pengkelasan. Proses ini dilaksanakan bagi memastikan fitur terbaik dapat digunakan untuk menyumbang kepada keputusan pengkelasan. Terdapat pelbagai teknik dan kriteria bagi tujuan pemilihan fitur input yang digunakan dalam kajian lepas. Teknik berkenaan antaranya dengan menentukan fitur yang relevan dan menapis fitur yang tidak relevan dengan pembelajaran perisian hasad. Selain daripada itu, teknik menapis fitur berdasarkan kriteria tingkah laku spesifik perisian hasad juga turut digunakan. Menurut Ranveer, & Hiray (2015), fitur yang ada pada setiap perisian memberi penerangan deskriptif tentang aktiviti yang terlibat dengan perisian. Oleh yang demikian, pemilihan fitur yang digunakan sebagai input perlulah teguh dan dipercayai kerana ianya mampu mempengaruhi keberkesanan pengkelasan. Fitur yang tidak relevan mampu memesongkan pengesanan dan membuat pengkelasan yang salah.

Fitur input yang dipilih kemudian melalui proses pengekstrakan fitur. Proses ini mengekstrak fitur yang ingin digunakan kepada satu data baru yang boleh dibaca oleh algoritma PM (Lin et al. 2015). Terdapat pelbagai kaedah yang digunakan dalam

kajian lepas dan salah satu yang terkenal daripadanya adalah model Beg Perkataan (BP). Model BP seperti yang diketahui mencari perkataan dan menukar nilai perkataan kepada nilai angka (Jin et al. 2016). Model ini digunakan dalam kajian lepas untuk mencari perkataan fitur yang telah dipilih di peringkat pra-pemprosesan (Passalis & Tefas 2017; Rui et al. 2016; Zhang et al. 2017). Setelah perlaksanaan pengekstrakan fitur selesai, fitur digunakan di peringkat pengkelasan. Dengan sebab itu, pemilihan dan perwakilan fitur serta kaedah pengkelasan yang dipilih perlulah teguh dan dipercayai agar pengkelasan dapat dibuat dengan tepat dan berkesan. Sub topik seterusnya membincangkan kajian lepas tentang pengkelasan perisian hasad. Kaedah pengkelasan perisian hasad juga dikategorikan sebagai permasalahan dalam bidang pengkelasan. Pengkelasan dilaksanakan dengan mempelajari input data yang diberikan dan membuat pengkelasan ke atas data baru. Terdapat pelbagai teknik PM yang dapat digunakan untuk diimplikasikan dalam pengkelasan perisian hasad seperti Algoritma RN.

2.2.1 Kajian Lepas Berkaitan Kaedah Pengkelasan Perisian Hasad

Bahagian ini membincangkan beberapa kajian lepas yang berkaitan dengan pengkelasan perisian hasad. Bagi mendapatkan pengkelasan yang mempunyai kebergantungan sifar kepada manusia, teknik pengkelasan PM memberikan banyak kelebihan berbanding teknik pengkelasan perisian hasad yang lain. Oleh itu, kajian ini memfokuskan pengkelasan perisian hasad yang menggunakan kaedah pengkelasan PM. Kaedah PM bukanlah satu teknik yang baru diperkenalkan. Lebih daripada sedekad yang lalu, ramai pengkaji telah mula menggunakan kaedah PM (Louridas & Ebert 2016) untuk mengkelas perisian hasad dan membandingkan prestasi antara satu algoritma dengan algoritma lain (Gandotra et al. 2014; Ranveer & Hiray 2015). Ternyata, terdapat pelbagai teknik yang diaplikasikan untuk mengkelas perisian hasad dalam kajian lepas telah menyumbang kepada perkembangan kajian ini.

Kaedah PM secara umumnya menjalankan sesuatu tugasan dengan mempelajari satu set data latihan dan melaksanakannya ke atas satu set data yang belum pernah ditemui sebelumnya. Strategi kaedah PM terbahagi kepada dua iaitu Pembelajaran Mesin Tanpa Penyelia (PMTP) dan Pembelajaran Mesin Berpenyelia

(PMB). Set data latihan dalam PMB mempunyai label yang membolehkan sistem mempelajari bagaimana cara mengkelas set data pengujian. PMTP pula hanya mempunyai set data tanpa label yang memerlukan sistem mencarinya sendiri. Kaedah PM membuat pengkelasan perisian hasad dengan mempelajari corak fitur perisian terlebih dahulu sebelum membuat sesuatu keputusan pengkelasan. Antara algoritma yang sering digunakan dalam PM untuk pengkelasan adalah Regresi Logik, Pokok Keputusan, Mesin Sokongan Vektor (MSV), Rangkaian Neural (RN), Rangkaian Bayesian. Keterangan lanjut algotitma PM diberi dalam Jadual 2.1.

Jadual 2.1 Rumusan algoritma PM

Algoritma PM	**Penyelidik**	**Kaedah**	**Kebaikan**	**Kelemahan**
Regresi Logik	(Liu et al. 2016)	Algoritma ini sesuai bagi analisa data. Ia digunakan untuk menerangkan data mempelajarinya dan mengkelas data kepada satu kelas.	Membuat pengkelasan berbilang kelas dengan berkesan.	Berkemungkinan untuk membuat pengkelasan berdasarkan variabel yang berdikari. Jika data yang dimasukkan adalah variebel yang salah, maka pengkelasan yang dibuat tidak bermakna.
Pokok Keputusan	(Jia & Huang 2010)	Algoritma ini erkenal bagi masalah pengkelasan. Kaedah yang digunakan adalah dengan mempelajari set data latihan dan hasilkan model. Algoritma ini membolehkan model dibina dengan mudah dan berkesan.	Berkesan untuk pengkelasan data kecil mahupun besar.	Pembinaan pokok keputusan mengambil masa yang agak lama dan intensif.
Mesin Sokongan Vektor	(Zheng et al. 2013)	Kaedah ini adalah kaedah yang paling terkenal. Data bagi algoritma ini disertakan bersama label dari dua kelas. Pembinaan model dibangun menggunakan data latihan dan pengujian dilakukan ke atas data latihan bagi mendapatkan kelas sebenar data tersebut.	Kadar pembelajaran dan keputusan yang tinggi walaupun menggunakan data yang kecil.	Kaedah pengkelasan algoritma mengambil masa yang lama untuk melatih data latihan.
Rangkaian Neural	(Dong et al. 2010).	Rangkaian Neural mengadaptasi cara otak berfungsi menerima input dan menghasilkan hasil yang diingini.	Pembelajaran dihasilkan daripada data yang sedikit. Pembelajaran adalah lebih berkesan apabila data bertambah.	Masa yang lama diambil bagi memproses data latihan.

bersambung...

...sambungan

Rangkaian Bayesian	(Yue et al. 2015)	Algoritma ini menggunakan hubungan jangkaann di antara variabel yang dikaji. Pembelajaran data dibuat ke atas set data yang ditanda bersama label.	Mampu menghubungkan maklumat pembolehubah yang dikaji dengan diberi.	Sukar mengendali data yang mempunyai fitur berterusan.

Sumber: Singh, & Nene (2013)

Secara keseluruhannya, pengkelasan menggunakan kaedah PM lebih fokus kepada pembelajaran dan analisis data. Selain dari menggunakan teknik pengkelasan yang berbeza, kajian lepas turut menggunakan pelbagai set data yang popular. Beberapa kajian pengkelasan perisian hasad yang lepas ditunjukkan dalam Jadual 2.2.

Jadual 2.2 Kajian lepas pengkelasan perisian hasad menggunakan kaedah PM

Artikel	**Objektif**	**Metod**	**Set data/**	**Keputusan**
(Ahmed et al. 2009)	Mencadangkan model pengkelasan ruangan-temporal perisian hasad (panggilan API)	*Instance based learner* (*IBk*), pokok keputusan (J48), Bayesian Naif (BN), RIPPER, Mesin Sokongan Vektor (MSV)	-Set data perisian hasad: *VX Heavens Virus Collection*, -Set data perisian normal: pemasang aplikasi boleh laku, aplikasi boleh laku Windows XP.	Penggunaan input ruangan dan temporal menaikkan pengesanan perisian hasad pengkelas PM.
DroidMat (Wu et al. 2012)	Pengkelasan menggunakan bilangan fitur berbeza. (Hasrat, Keizinan dan API)	K-Means, EM, KNN, Bayesian Naif	Set data Contagio mobile, set data aplikasi Google Play	Keberkesanan penggunaan fitur Keizinan+ hasrat + API lebih tinggi berbanding fitur keizinan+API dan fitur hasrat+Inten dalam meningkatkan ketepatan pengkelasan.

bersambung…

…sambungan

(Saxe & Berlin 2015)	Meningkatkan prestasi pengkelasan.	Algoritma Rangkaian Neural	Set data persendirian Invincea: 400,000 set data perduaan	Keputusan prestasi pengkelasan tidak dapat ditingkatkan sepenuhnya kerana data input yang digunakan adalah data perduaan.
(Pascanu et al. 2015)	Mempelajari bahasa perisian hasad untuk mengkelas ancaman yang tak dikenali.	ESN, RNN, regresi logic, RNB	Set data persendirian: Set data vector	Penimbunan pengkelas menunjukkan kadar positif betul meningkat sehingga 98.3% dan kadar positif salah sebanyak 0.1%
(Huang & Stokes 2016)	Membuat pengkelasan perduaan ke atas urutuan API dan urutan objek nol	Beberapa lapisan rangkaian neural.	Set data Microsoft Corporation	Prestasi pengkelasan perisian hasad meningkat sebanyak 2.94%
(Kim et al. 2016)	Menguji keberkesanan algoritma pembelajaran mendalam.	Algoritma MJMPP –RNN.	KDD CUP-99	Perbandingan antara model pengeklasan menggunakan model MJMPP menunjukkan keputusan pengkelasan lebih tinggi berbanding menggunakan pengkelas PM yang lain.
(Tobiyama et al. 2016)	Meningkatkan pengkelasan hasad dengan menggabungkan Rangkaian Neural	Analisis dinamik untuk pengekstrakan fitur.Algoritma PM untuk pengkelasan.	Memproses tingkah laku fail log perisian dan merekodkan log tingkah laku perisian yang diekstrak sebagai set data fitur	Keputusan perbandingan AUC=0.96.

Kajian Ahmed et al. (2009) mengkelas perisian hasad menggunakan teknik PM ke atas fitur panggilan API yang diekstrak daripada set data yang dikumpulkan. Mereka menggunakan informasi ruangan dan temporal fitur yang diekstrak. Hasil pengkelasan menunjukkan ketepatan gabungan informasi ruangan dan temporal data adalah lebih tinggi berbanding penggunaan informasi ruangan atau temporal sahaja.

Seterusnya, kajian Wu et al. (2012) melaksanakan pengkelasan perisian hasad Android ke atas set data Contagio Mobile. Mengambil kira mekanisma keselamatan Android, pengkelasan dijalankan dengan mempelajari fitur keizinan, hasrat dan API yang diekstrak daripada set data perisian. Hasil pengkelasan menunjukkan pemilihan fitur yang ingin digunakan sebagai input memainkan peranan penting dalam meningkatkan ketepatan pengkelasan. Pengkelasan menggunakan fitur keizinan+hasrat+API lebih tinggi ketepatan berbanding pengkelasan yang menggunakan salah satu fitur sahaja.

Keputusan prestasi kaedah pembelajaran tradisional PM menunjukkan ketepatan pengkelasan yang memberangsangkan. Namun begitu, kaedah ini menghadapi masalah apabila jumlah data yang digunakan adalah besar (Qiu et al. 2016). Dengan itu, kaedah pembelajaran perwakilan input data perlulah teguh untuk menyelesaikan masalah ini (Bengio, Yoshua, Aaron Courville 2013). Perwakilan fitur adalah solusi bagi mempersembahkan data kepada satu bentuk informasi yang berguna dan bermakna apabila membina model pengkelasan. Namun begitu, keberkesanan model pengkelasan mempelajari perwakilan fitur adalah bergantung kepada jumlah penggunaan fitur yang digunakan dalam pengkelasan tersebut (Li et al. 2018). Matlamat pemilihan fitur adalah untuk mencapai bilangan fitur yang paling berkesan untuk membantu meningkatkan ketepatan pengkelasan.

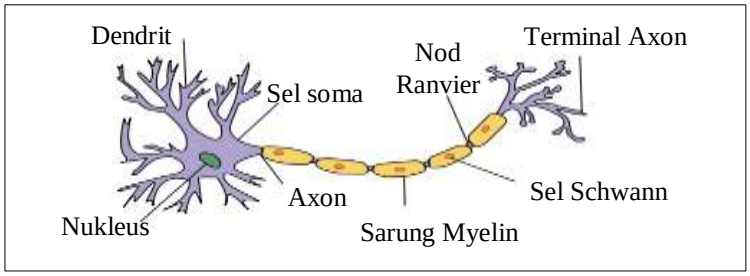
Yu et al. (2011) menyatakan bahawa algoritma pembelajaran mendalam PM mampu mempelajari corak perwakilan fitur yang rumit secara hierarki dan mencapai prestasi ketepatan mengatasi teknik PM yang cetek (*shallow*). Algoritma pembelajaran mendalam seperti algoritma RN lebih mudah sesuai dengan bidang baru (Humphrey et al. 2013). Kajian Saxe & Berlin (2015) menggunakan teknik suapan ke depan RN dalam kajian mereka bagi proses analisis fitur. Namun begitu, keputusan prestasi pengkelasan tidak dapat ditingkatkan sepenuhnya kerana perwakilan input yang digunakan adalah data perduaan. Selain daripada hanya dengan menggunakan satu lapisan algoritma RN sahaja untuk pengkelasan, penimbunan beberapa algoritma RN juga digunakan dan berkesan bagi pengkelasan perisian hasad seperi yang dilaksanakan oleh kajian Huang & Stokes (2016). Penimbunan algoritma pembelajaran mendalam ini mampu mempelajari corak perwakilan fitur yang

digunakan dengan lebih berkesan (Arel et al. 2010). Pascanu et al. pula menggabungkan algoritma RN iaitu Rangkaian Neural Berulang (RNN) dan algoritma Rangkaian Neural Buatan (RNB) sebagai model pengkelasan (Pascanu et al. 2015). Model pengkelasan yang dibina mempelajari kod perisian hasad dan perisian tidak hasad lalu menghasilkan satu perwakilan fitur dalam fail baru. Peningkatan pengkelasan berjaya direkodkan menggunakan model pengkelasan gabungan yang dicadangkan.

Menerusi model bahasa RNB yang dilaksana oleh Pascanu et al. (2015), Kolosnjaji et al. (2016) dan Athiwaratkun, & Stokes (2017) telah mengembangkan idea model bahasa pengkelasan. Mereka fokus kepada pembelajaran perisian hasad melalui model bahasa yang boleh mempelajari urutan perisian hasad. Model bahasa dibangunkan dengan menggunakan algoritma Memori Jangka Masa Pendek (MJMPP) dan Unit Pagar Berulang (UPB). Kajian Athiwaratkun & Strokes berjaya mencapai 31.30% peningkatan pengkelasan perisian hasad dengan gabungan model bahasa yang dibangunkan. Dalam kajian seterusnya, 95.37% keputusan keberkesanan RNN berjaya direkod menggunakan set data DARPA pada tahun 2015 oleh penyelidik Korea (Kim 2016). Kajian ini kemudiannya dikembangkan menggunakan algoritma MJMPP keatas set data KDD untuk membuat pengkelasan (Kim et al. 2016). Model berjaya mencapai ketepatan pengkelasan sebanyak 96.93%, mengatasi masalah temporal serta meramal daripada urutan yang memerlukan ketepatan masa dan pengiraan.

2.3 RANGKAIAN NEURAL

Matlamat utama pembinaan model pengkelasan perisian hasad adalah untuk mencapai tahap ketepatan pengkelasan yang paling optimal. Umum mengetahui, pelbagai kaedah pengkelasan yang boleh digunakan untuk mengkelas perisian hasad. Kajian ini menggunakan Algoritma Rangkaian Neural (RN) yang diadaptasi dari struktur neuron otak manusia seperti yang ditunjukkan dalam Rajah 2.2.



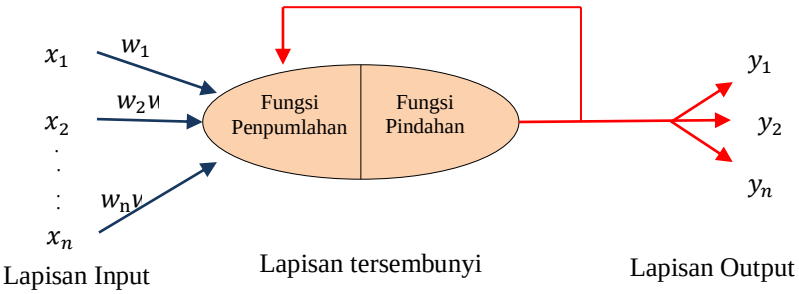
Rajah 2.2 Struktur neuron otak manusia

Sumber: (Grosse et al. 2017)

Komunikasi antara satu neuron kepada neuron yang lain berlaku melalui penghubung yang dipanggil sinapsis (Krogh 2008). Setiap neuron mengandungi sejumlah dendrit, soma dan juga axon. Dendrit menerima signal input berupa impuls elektrik lalu memodifikasikannya dan menghantarnya ke soma. Soma akan mengumpulkan semua signal yang dihantar oleh dendrit sehingga menepati satu ambang. Signal seterusnya dihantar kepada neuron lain melalui axon. Otak belajar secara semulajadi melalui pemerhatian dan deria yang ada.

2.3.1 Rangkaian Neural Buatan

Pada tahun 1950, satu model Rangkaian Neural Buatan (RNB) telah diperkenalkan oleh saintis bernama Frank Rosenblatt (Gupta 2013). Model ini meniru fungsi rangkaian neural otak manusia dalam menterjemah sesuatu input data. Senibina RNB dapat dilihat seperti dalam Rajah 2.3.



Rajah 2.3 Senibina Rangkaian Neural Buatan (RNB)

Sumber: Grosse et al. (2017)

RNB adalah model yang menghasilkan output berdasarkan jumlah berat daripada input yang telah diberikan kepada model. Model ini terbahagi kepada tiga lapisan asas iaitu lapisan input, lapisan tersembunyi dan lapisan output. RNB beroperasi dengan membaca dan mempelajari input x_n di lapisan input dan meletakkan pemberat w_n kepada setiap input. Pemberat ini merujuk kepada kepentingan input untuk menghasilkan output y_n. Lapisan tersembunyi yang terletak di antara lapisan input dan lapisan output akan menerima pemberat ditetapkan bagi setiap input dan menjumlahkannya sebelum dipindahkan kepada lapisan output. Fungsi pindahan memainkan peranan penting untuk mengemaskini berat input dan menentukan format output. Algoritma RNB seringkali digunakan di lapisan pengkelasan sebelum keputusan kelas perisian dihasilkan (Nauman et al. 2017; Nix & Zhang 2017). Fungsi pindahan yang boleh digunakan untuk menghasilkan output adalah seperti berikut:

Klasifikasi perduaan (dua kelas):Fungsi pengaktifan logik, dimana terdapat hanya satu neuron sahaja di lapisan output. Klasifikasi jenis ini boleh menggunakan dua fungsi pindahan samada fungsi sigmoid ataupun tanh:

- Sigmoid Fungsi sigmoid adalah seperti berikut:

$$\sigma(x) \equiv \frac{1}{1+e^{-x}} \tag{2.1}$$

Fungsi ini akan melakukan operasi ke atas nombor nombor sebenar dan jadikannya diantara nombor 0 hingga 1 sahaja. Ia akan memberi gambaran jelas tentang kadar pembuangan neuron.

$$y = \begin{cases} 0\ jika\ \sum_n w_n x_n \leq ambang \\ 1\ jika\ \sum_n w_n x_n > ambang \end{cases} \tag{2.2}$$

Penentuan nilai output y adalah berpandukan penjumlahan hasil input dan berat input $w_n x_n$. Jika jumlah hasil input dan berat kurang ataupun sama dengan nilai ambang, output adalah 0, dan jika jumlah hasil kurang daripada nilai ambang maka output adalah 1.

- Tanh Fungsi tanh seperti berikut:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.3}$$

Fungsi ini akan melakukan operasi ke atas nombor sebenar dan jadikan ia nombor diantara nombor 1 dan -1. Beza antara fungsi sigmoid dan tanh adalah pada perbezaan asimtot tanh yang terletak pada $y = -1$, dan bukannya $y = 0$. Dengan itu, kepelbagaian nilai untuk tanh adalah lebih besar berbanding sigmoid.

$$y = \begin{cases} -1\ jika \sum_n w_n x_n \leq ambang \\ 1\ jika \sum_n w_n x_n > ambang \end{cases} \tag{2.4}$$

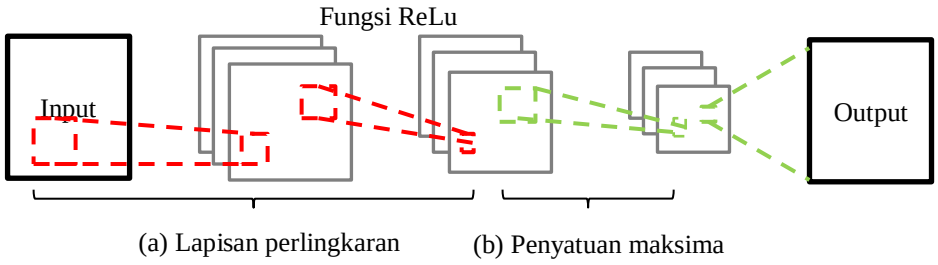
Nilai output y adalah -1 jika hasil penjumlahan input dan berat $w_n x_n$ kurang ataupun sama dengan nilai ambang yang ditetapkan manakala nilai output adalah 0 sekiranya hasil penjumlahan input dan berat melebihi nilai ambang.

Keputusan dalam fungsi pindahan seterusnya dihantar ke lapisan output. Lapisan output mempamerkan satu nilai output yang diambil terus daripada fungsi pindahan. Dalam kajian ini, fungsi sigmoid RNB digunakan bagi mengkelas input kepada satu output akhir daripada dua kelas samada kelas perisian hasad atau kelas perisian bukan hasad. RNB sangat sesuai digunakan untuk fungsi pemetaan daripada input kepada output.

Walau bagaimanapun, kekurangan algoritma ini adalah ia memerlukan input dibahagikan kepada beberapa input kecil untuk tujuan pengkelasan (Sutskever et al. 2014). Dengan itu, ia menyebabkan input bertindih antara satu sama lain semasa proses simulasi penghasilan output dijalankan. Seterusnya, RNB bersifat tanpa identiti yang hanya belajar fungsi tetap penghampiran dan tidak peka dengan struktur temporal input. Hal ini memberi impak yang sangat besar dalam proses memahami dan mengurus struktur temporal kerana rangkaian tidak dapat menyedari kehadiran struktur temporal dengan jelas. Selain daripada itu, RNB mempunyai kekangan dari segi bilangan input dan output yang tetap. Ini bermakna, hanya jumlah input dan output tertentu sahaja dapat dilaksanakan jika menggunakan kaedah RNB.

2.3.2 Rangkaian Neural Perlingkaran

Rangkaian Neural Perlingkaran (RNP) adalah model RN yang mampu mengesan dan menganalisa corak data (Kalchbrenner et al. 2014). Model ini mendapat inspirasi daripada fungsi korteks visual dalam otak manusia yang mampu mengesan corak objek secara berkala dan hierarki. Model RNP mengambil tiga kelebihan utama korteks visual manusia untuk diadaptasi. Yang pertama adalah hubungan tempatan antara neuron yang mampu menghasilkan set fitur baru daripada input (Li et al. 2015). Contohnya dengan melihat objek, manusia mampu menyenaraikan fitur yang terdapat pada objek tersebut dari segi bentuk, warna dan sebagainya. Seterusnya, kelebihan mempelajari hierarki fitur input. Ciri ini merujuk kepada kebolehan otak mengkaji fitur yang paling menonjol pada objek. Ciri kepelbagaian ruangan (*spatial*) dalam otak juga mempengaruhi pembangunan model RNP (Shin et al. 2016) yang mana merujuk kepada kebolehan otak manusia apabila melihat satu objek contohnya kasut yang terdiri daripada jenama dan bentuk yang berbeza dan pelbagai corak, namun manusia tetap boleh mengenal pasti bahawa objek itu adalah kasut. Senibina RNP digambarkan dalam Rajah 2.4 diperkenalkan oleh Yann LeCun dan Léon Bottou (LeCun et al. 1998; LeCun & Bengio 1995).



Rajah 2.4 Senibina Rangkaian Neural Perlingkaran

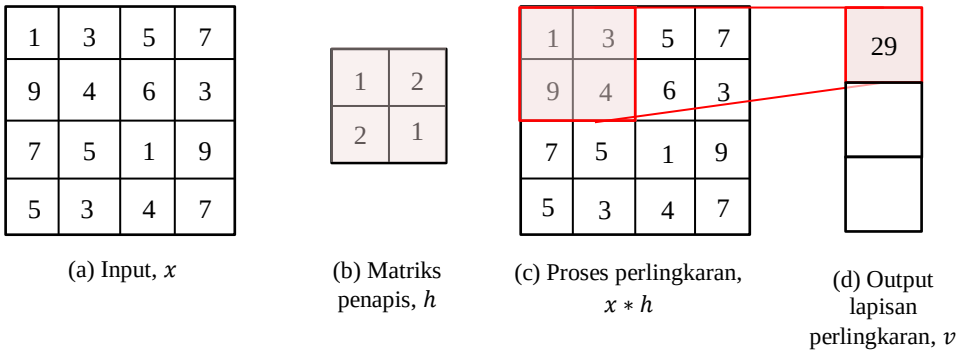
Sumber: Jaderberg et al. (2016)

Terdapat dua lapisan istimewa di dalam RNP iaitu lapisan perlingkaran bersama fungsi Relu seperti yang digambarkan dalam Rajah 2.4(a) dan lapisan penyatuan maksima yang ditunjukkan dalam Rajah 2.4(b) (Tobiyama et al. 2016). Setiap lapisan mempunyai tugasan tersendiri dan menyediakan hasil berguna untuk

satu sama lain. RNP dipercayai dapat meningkatkan prestasi RNB terutamanya dari membaca ruangan data (Graham 2014). Input RNP dibaca dan diproses oleh lapisan perlingkaran dan dihantar ke lapisan penyatuan maksima sebelum dihasilkan sebagai output. Format input yang paling terkenal untuk model RNP adalah data tiga dimensi yang digunakan dalam pengecaman imej. Namun begitu, kajian ini menggunakan input satu dimensi.

a) Lapisan perlingkaran

Rajah 2.5 menunjukkan proses pembelajaran input di lapisan perlingkaran (Schulz & Behnke 2012).



Rajah 2.5 Lapisan Perlingkaran

Input x dibaca dan diproses dengan menggunakan satu penapis h yang melingkari informasi dan menghasilkan data baru, v. Penapis h seperti yang ditunjukkan dalam Rajah 2.5(b) adalah matriks yang meniru fungsi otak yang memanfaatkan hubungan tempatan antara neuron sewaktu memproses input. Hubungan tempatan di lapisan perlingkaran berlaku dengan menghubungkan penapis di sekitar satu lingkungan tempatan input sahaja. Saiz tempatan input ditentukan oleh saiz penapis yang digunakan di mana senibina matriks penapis lebih kecil berbanding saiz input (He et al. 2015). Proses perlingkaran dalam lapisan ini diilustrasi seperti dalam Rajah 2.5(c) di mana nilai yang ada pada setiap penapis mendarab nilai input dan kesemua hasil darab dijumlah bagi membentuk satu kemasukan baru matriks output v sepeti yang ditunjukkan dalam Rajah 2.5(d). Proses perlingkaran berterusan sehingga seluruh nilai

input dilingkar. Penghasilan output v boleh difahami dengan lebih lanjut menggunakan formula seperti berikut:

$$v[n] = x[n] * h[n] = \sum_{k} x[k] * h[n-k]\,, k \in [-\infty, +\infty] \tag{2.5}$$

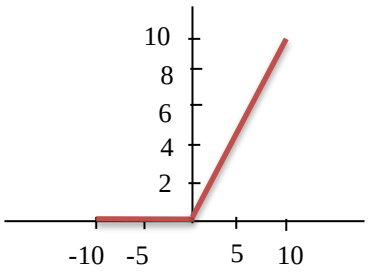
Pergerakan penapis dikawal oleh langkah yang dikenali sebagai *stride*. Sekiranya nilai *stride* adalah 1 maka penapis akan bergerak satu langkah daripada nilai input yang telah dilingkari. Jika *stride* ditetapkan kepada 2, maka penapis akan melompat dua langkah daripada input lama untuk melingkari nilai input baru. Nilai *stride* akan mempengaruhi saiz output yang dihasilkan. Penapis bergerak bermula dari kiri ke kanan matriks input dan akan ke bawah apabila penapis telah sampai ke penjuru paling kanan matriks input. Selain daripada stride, parameter yang mempengaruhi saiz output adalah *zero-padding*. Parameter ini meletak nilai 0 di hujung sempadan input. Saiz input boleh dikira menggunakan persamaan berikut:

$$saiz\ v = \frac{W - F + 2P}{S + 1} \tag{2.6}$$

W adalah saiz matriks input, F adalah saiz matriks penapis yang digunakan, S adalah nilai *stride* dan P adalah jumlah *zero-padding* di sempadan input. Output v daripada pendaraban di lapisan perlingkaran kemudiannya melalui satu fungsi pengaktifan yang dinamakan fungsi ReLU (Nair & Hinton 2010). Fungsi ini boleh ditakrifkan seperti dalam persamaan (2.7).

$$f[v] = max(0, v) \tag{2.7}$$

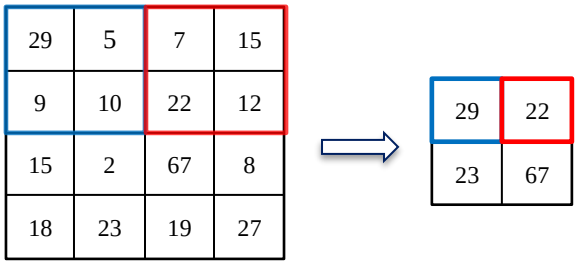
Fungsi ini bertujuan mengeluarkan semua nilai negatif dalam matrik v dengan menukar semua nombor negatif kepada nilai sifar dan kekalkan nombor bernilai positif. Graf fungsi ReLu dapat dilihat dalam Rajah 2.6.



Rajah 2.6 Fungsi ReLu

b) Penyatuan maksima

Lapisan seterusnya adalah lapisan penyatuan maksima seperti yang dapat dilihat dalam Rajah 2.7 (Yoshioka et al. 2015). Lapisan ini menyaring semua nilai output v yang melalui fungsi relu menggunakan satu pintu tingkap. Pintu tingkap ini adalah matriks yang mengurangkan dimensi input bagi lapisan seterusnya. Melalui tingkap penyatuan ini, nilai input akan dibandingkan antara satu sama lain dan satu nilai tertinggi akan dipilih. Langkah ini akan diulang bagi nilai yang seterusnya. Setelah semua perbandingan nilai dilakukan, maka terhasil satu output y.



Rajah 2.7 Operasi penyatuan maksima

Operasi dalam Rajah 2.7 menggunakan pintu tingkap yang bersaiz 2x2 dan pergerakannya menggunakan langkah *stride* [2,2]. Operasi ini menghasilkan output matriks 2x2. Saiz output daripada proses penyatuan maksima boleh dikira menggunakan persamaan (2.9) bagi lebar output dan persamaan (2.10) bagi mengira ketinggian output.

$$W = \frac{W_w - P_w}{S_w} + 1 \tag{2.9}$$

$$H = \frac{H_H - P_H}{S_H} + 1 \tag{2.10}$$

W_W adalah saiz kelebaran input, maka lebar input dalam Rajah 2.7 adalah 4. P_w adalah saiz kelebaran pintu tingkap iaitu 2. S_w adalah bilangan langkah *stride* iaitu 2 langkah untuk setiap perbandingan kiri ke kanan input. Untuk dimensi ketinggian output H seperti dalam persamaan (2.10), H_H mewakili nilai ketinggian ataupun sampel yang digunakan iaitu 4. P_H adalah ketinggian pintu tingkap yang digunakan manakala S_H adalah langkah *stride* yang digunakan untuk menggerakkan pintu tingkap dari atas ke bawah matriks input, masing-masing bernilai 2.

Algoritma RNP adalah satu algoritma yang cukup berkesan dalam menyelesaikan masalah ruangan dan mengekstrak data. Namun begitu, RNP mempunyai masalah dalam mengingati informasi input (Severyn & Moschitti 2015). RNP biasanya melupakan tugas yang telah diproses sebelumnya sekaligus mengabaikan ciri temporal pada input urutan. Dengan itu, RNP memproses elemen yang hadir pada data tanpa mempertimbangkan hubungan informasi antara input data.

2.3.3 Rangkaian Neural Berulang

Rangkaian Neural Berulang (RNN) adalah satu solusi yang menyelesaikan isu yang dihadapi RNB dan RNP. Kekuatan utama RNN terletak pada kebolehannya mengingati informasi. RNN telah dikenalkan sejak tahun 1980 namun dilupakan kerana masalah kecerunan (Hochreiter & Schmidhuber 1997). Kebanyakan pengkaji menggunakan algoritma pembelajaran mesin seperti Ripper, k-Jiran terdekat, Bayes naïf, dan Mesin Sokongan Vektor (Idika & Mathur 2007) namun bagi permasalahan untuk melatih urutan dan memproses temporal data, RNN adalah yang terbaik untuk digunakan (Athiwaratkun & Stokes 2017; Brownlee.; Hochreiter & Schmidhuber 1997; Kim et al. 2016; Pascanu et al. 2015). Struktur temporal data adalah maklumat yang sentiasa berubah-ubah mengikut masa. Oleh yang demikian, data lebih sesuai diproses oleh RNN kerana pembelajaran setiap urutan input itu dilakukan secara