MENGATASI MASALAH KETIDAKSEIMBANGAN NSL-KDD MENGGUNAKAN TEKNIK PENSAMPELAN DATA : SEBUAH PERBANDINGAN

AMEERAH SAEEDATUS SYAHEERAH BINTI ABD HAMID

UNIVERSITI KEBANGSAAN MALAYSIA

MENGATASI MASALAH KETIDAKSEIMBANGAN NSL-KDD MENGGUNAKAN TEKNIK PENSAMPELAN DATA : SEBUAH PERBANDINGAN

AMEERAH SAEEDATUS SYAHEERAH BINTI ABD HAMID

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

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

20 Februari 2024

AMEERAH SAEEDATUS SYAHEERAH BINTI ABD HAMID
P98563

PENGHARGAAN

Bismillahirrahmanirrahim

Terlebih dahulu, saya ingin memanjatkan kesyukuran yang tidak terhingga kepada Allah SWT, atas kekuatan yang dikurniakan kepada saya sepanjang usaha ini. Dengan keyakinan yang tulus dalam anugerah-Nya, saya dapat mengharungi cabaran sepanjang laluan akademik ini.

Saya ingin merakamkan setinggi-tinggi penghargaan kepada penyelia projek saya, Dr. Wan Fariza Binti Paizi @ Fauzi. Kepakaran dan bimbingannya yang bernas dan tidak ternilai sepanjang proses pembangunan dan penulisan projek ini. Bimbingan beliau bukan sekadar akademik tetapi juga sokongan moral yang mendorong saya untuk berusaha menyiapkan projek ini. Kesabaran, dorongan, dan pengetahuan yang mendalam telah menjadi sumber inspirasi yang hebat untuk saya. Ucapan terima kasih yang tidak terhingga juga saya tujukan kepada pengajar kursus Sarjana Keselamatan Siber di FTSM. Ajaran mereka telah membentuk pertumbuhan intelek dan pandangan profesional saya secara mendalam. Kesediaan mereka untuk berkongsi ilmu dan menawarkan nasihat telah memainkan peranan penting dalam membentuk perjalanan akademik saya.

Kepada ibu bapa saya, tiada kata-kata yang dapat menyatakan rasa terima kasih saya dengan secukupnya. Kasih sayang anda yang tidak bersyarat, sokongan yang tidak berbelah bahagi, kepercayaan dan pengorbanan anda yang tidak berkesudahan telah meletakkan asas di mana saya berdiri hari ini. Dengan itu, saya sentiasa bersyukur. Kepada adik-beradik saya, terima kasih atas pemahaman, dorongan, dan kata-kata semangat yang tidak terkira banyaknya yang anda berikan semasa saya mengalami kesukaran. Sokongan dan kasih sayang anda menjadikan perjalanan ini lebih mudah ditempuhi.

Akhir sekali, saya amat berterima kasih kepada semua yang terlibat secara langsung dan tidak langsung sepanjang fasa menyiapkan projek ini.

ABSTRAK

Set data yang tidak seimbang menimbulkan cabaran yang ketara untuk model pembelajaran mesin, terutamanya dalam bidang sistem pengesanan pencerobohan di mana kelas minoriti sering memegang maklumat penting. Kajian ini menyiasat keberkesanan teknik pensampelan semula dalam menangani ketidakseimbangan kelas dalam set data NSL-KDD, set data penanda aras untuk pengesanan pencerobohan. Melalui analisis perbandingan, tiga kaedah pensampelan semula—Pengurangan Pensampelan Secara Rawak (RUS), Teknik Pensampelan Terlebih Minoriti (SMOTE), dan pendekatan hibrid SMOTE-RUS—dinilai dari segi keupayaan mereka untuk meningkatkan pengesanan kelas minoriti. Hasil kajian mendedahkan bahawa RUS mengatasi kedua-dua SMOTE dan SMOTE-RUS dalam mengesan kelas minoriti, dengan SMOTE mengikuti tempat kedua dan SMOTE-RUS menyediakan prestasi pertengahan antara SMOTE dan RUS. Penemuan ini memberikan pandangan berharga tentang pemilihan strategi pensampelan semula yang sesuai untuk mengendalikan ketidakseimbangan kelas dalam tugas pengesanan pencerobohan dan menunjukkan bahawa keberkesanan kaedah pensampelan semula sangat bergantung pada sifat sebuah set data.

OVERCOMING IMBALANCE NSL-KDD DATASET USING RESAMPLING TECHNIQUES : A COMPARISON

ABSTRACT

Imbalanced datasets pose a significant challenge for machine learning models, particularly in the realm of intrusion detection systems where minority classes often hold crucial information. This study investigates the efficacy of resampling techniques in addressing class imbalance within the NSL-KDD dataset, a benchmark dataset for intrusion detection. Through a comparative analysis, three resampling methods—Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), and a hybrid approach SMOTE-RUS—are evaluated in terms of their ability to enhance the detection of minority classes. Experimental results reveal that RUS outperforms both SMOTE and SMOTE-RUS in detecting minority classes, with SMOTE following closely behind, and SMOTE-RUS providing an intermediate performance. These findings provide valuable insights into the selection of appropriate resampling strategies for handling class imbalance in intrusion detection tasks showing that the effectiveness of resampling methods is highly contingent on the nature of the datasets.

KANDUNGAN

Halaman

PENGAKUAN		**ii**
PENGHARGAAN		**iii**
ABSTRAK		**iv**
ABSTRACT		**v**
KANDUNGAN		**vi**
SENARAI JADUAL		**ix**
SENARAI ILUSTRASI		**xi**
SENARAI SINGKATAN		**xiii**
BAB I	**PENGENALAN**	
1.1	Pendahuluan	1
1.2	Latar Belakang Kajian	1
	1.2.1 Cabaran Set Data NSL-KDD	4
1.3	Permasalahan Kajian	6
	1.3.1 Teknik Pensampelan Semula	7
1.4	Persoalan Kajian	9
1.5	Objektif Kajian	9
1.6	Kepentingan Kajian	9
1.7	Skop Dan Batasan Kajian	10
1.8	Kesimpulan	11
BAB II	**KAJIAN LITERASI**	
2.1	Pengenalan	13
2.2	Keselamatan Siber dan Kepentingannya	13
	2.2.1 Jenis Ancaman dan Serangan Siber	14
	2.2.2 Sistem Pengesanan Pencerobohan (IDS)	14
	2.2.3 Peranan IDS dalam Keselamatan Siber	15
2.3	Set Data NSL-KDD	15
	2.3.2 Kajian Terdahulu Menggunakan Set Data NSL-KDD	21
2.4	Ketidakseimbangan Kelas	23

	2.4.1 Cabaran Ketidakseimbangan Kelas Terhadap Pembelajaran Mesin	23
	2.4.2 Teknik – Teknik Menangani Ketidakseimbangan Kelas	25
2.5	Teknik Pensampelan Semula	30
2.6	Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE)	31
	2.6.1 Kelebihan dan Kelemahan Teknik SMOTE	31
2.7	Pensampelan Rawak Kurang (RUS)	32
	2.7.1 Kelebihan dan Kelemahan Teknik RUS	32
2.8	Teknik Pensampelan Sintetik Minoriti Lebih – Pensampelan Rawak Kurang (SMOTE-RUS)	33
	2.8.1 Kelebihan dan Kelemahan Teknik SMOTE-RUS	33
	2.8.2 Kajian Terdahulu Dengan Teknik Pensampelan SMOTE, RUS dan SMOTE-RUS	34
2.9	Mesin Sokongan Vektor (SVM)	36
	2.9.1 Kajian Terdahulu Dengan Mesin Sokongan Vektor (SVM)	38
2.10	Kesimpulan	40
BAB III	**KAEDAH KAJIAN**	
3.1	Pengenalan	41
3.2	Pengumpulan Data NSL-KDD	42
3.3	Pra Pemprosesan Data	42
	3.3.1 Pengkategorian Data	43
	3.3.2 Pengekodan Nilai Nominal	43
	3.3.3 Penskalaan Data	43
3.4	Pensampelan data dengan Teknik SMOTE	43
3.5	Pensampelan Data Dengan Teknik RUS	44
3.6	Pensampelan Data Dengan Teknik SMOTE-RUS	44
3.7	Model Ramalan Support Vector Machine (SVM)	45
	3.7.1 Pembangunan Model Ramalan	45
3.8	Penilaian Metrik	46
3.9	Kesimpulan	48
BAB IV	**PELAKSANAAN EKSPERIMEN**	
4.1	Pengenalan	49
4.2	Persediaan Eksperimen	49

4.3	Pra-Pemprosesan data	50
	4.3.1 Pengkategorian Pelbagai Serangan Dalam Kelas Serangans	51
	4.3.2 Pengekodan *One-Hot*	51
	4.3.3 Penskalaan Data	52
	4.3.4 Pembahagian Set Data NSL-KDD (KDDTrain+)	53
4.4	Pembangunan Model Ke Atas Set data NSL-KDD	54
	4.4.1 Parameter Support Vector Machine (SVM)	55
	4.4.2 Pembangunan Model Asas	56
	4.4.3 Pembangunan Model SMOTE	57
	4.4.4 Pembangunan Model RUS	60
	4.4.5 Pembangunan Model SMOTE-RUS	63
4.5	Kesimpulan	68
BAB V	**DAPATAN KAJIAN**	
5.1	Pengenalan	69
5.2	Keputusan Pembangunan Dan Pengujian Model ASAS	70
5.3	Keputusan Pembangunan Dan Pengujian Model SMOTE	75
5.4	Keputusan Pembangunan Dan Pengujian Model RUS	80
5.5	Keputusan Pembangunan Dan Pengujian Model SMOTE-RUS	85
5.6	Keputusan Perbandingan Model	87
5.7	Kesimpulan	91
BAB VI	**RUMUSAN DAN CADANGAN**	
6.1	Pengenalan	92
6.2	Rumusan Penemuan Dan Cadangan	92
RUJUKAN		**94**
LAMPIRAN		
Lampiran A	Output KNIME – Keputusan Pembangunan Model Asas, SMOTE, RUS dan SMOTE-RUS	102
Lampiran B	Output KNIME – Keputusan Pengujian Model Asas, SMOTE, RUS dan SMOTE-RUS	104

SENARAI JADUAL

No. Jadual		Halaman
Jadual 2.1	Saiz Sampel Kelas NSL-KDD	17
Jadual 2.2	41 Kategori Fitur NSL-KDD	19
Jadual 2.3	Jumlah Rekod NSL-KDD	20
Jadual 2.4	Perbandingan Teknik Data, Algoritma dan Hibrid	28
Jadual 2.5	Perbandingan Kajian Teknik Data, Algoritma dan Hibrid	29
Jadual 2.6	Kelebihan dan Kelemahan SMOTE	31
Jadual 2.7	Kelebihan dan Kelemahan Teknik RUS	32
Jadual 2.8	Kelebihan dan Kelemahan Teknik SMOTE-RUS	33
Jadual 4.1	Pembahagian Set Data	54
Jadual 4.2	Data SMOTE	59
Jadual 4.3	Data RUS	62
Jadual 4.4	Rekod Data Minoriti Selepas SMOTE	64
Jadual 4.5	Rekod Data Majoriti Selepas RUS	66
Jadual 4.6	Jumlah Set Data Keseluruhan	68
Jadual 5.1	Keputusan Pembangunan Model Asas	70
Jadual 5.2	Keputusan Pengujian Model Asas Menggunakan Set Data KDDTest+	72
Jadual 5.3	Keputusan Pembangunan Model SMOTE	75
Jadual 5.4	Keputusan Pengujian Model SMOTE Menggunakan Set Data KDDTest+	78
Jadual 5.5	Keputusan Pembangunan Model RUS	80
Jadual 5.6	Keputusan Pengujian Model RUS Menggunakan Set Data KDDTest+	82
Jadual 5.7	Keputusan Pembangunan Model SMOTE-RUS	85
Jadual 5.8	Keputusan Pengujian Model SMOTE-RUS Menggunakan Set Data KDDTest+	86

Jadual 5.9 Keputusan Perbandingan Teknik Pensampelan Semula 87

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 1.1	Carta Alir Pelaksanaan Teknik Pensampelan Data	12
Rajah 3.1	Carta Alir Metodologi Kajian	41
Rajah 3.2	Pembangunan Model Ramalan	45
Rajah 3.3	Metrik Kekeliruan	47
Rajah 4.1	Pengkategorian Serangan	51
Rajah 4.2	Nilai Nominal ke Numerikal *protocol_type*	52
Rajah 4.3	Penskalaan Min-Max	53
Rajah 4.4	Pembahagian Set Data	54
Rajah 4.5	Tetapan SVM	55
Rajah 4.6	Data Model Asas	56
Rajah 4.7	Pembangunan Model Asas	57
Rajah 4.8	Tetapan SMOTE	58
Rajah 4.9	Pembangunan Model SMOTE	58
Rajah 4.10	Data Model SMOTE	59
Rajah 4.11	Pembahagian Data RUS	60
Rajah 4.12	Tetapan RUS	61
Rajah 4.13	Data Model RUS	61
Rajah 4.14	Pembangunan Model RUS	62
Rajah 4.15	Rekod Set Latihan Asal	63
Rajah 4.16	Pensampelan SMOTE Untuk Kelas Minoriti	63
Rajah 4.17	Jumlah Kelas Minoriti Selepas SMOTE	64
Rajah 4.18	Tetapan Parameter SMOTE-RUS	65
Rajah 4.19	Jumlah Rekod RUS Kelas Majoriti	65
Rajah 4.20	Gabungan Data Majoriti dan Minoriti	66

Rajah 4.21	Jumlah data SMOTE-RUS	67
Rajah 4.22	Pembangunan Model SMOTE-RUS	67
Rajah 5.1	Pengujian Model Asas	72
Rajah 5.2	Pengujian Model SMOTE	77
Rajah 5.3	Pengujian Model RUS	81
Rajah 5.4	Pengujian Model SMOTE-RUS	85

SENARAI SINGKATAN

ADASYN	Pendekatan Pensampelan Sintetik Adaptif (Adaptive Synthetic Sampling Approach)
AE-DQN	Pembelajaran Pengukuhan Adversarial/Ejen Pelbagai dengan Pembelajaran Dalam-Q (Adversarial/Multi-Agent Reinforcement Learning – Deep-Q Learning
ANN	Rangkaian Neural Buatan (Artificial Neural Network)
BSMOTE	Sempadan SMOTE (Borderline SMOTE)
DARPA	Defense Advanced Research Projects Agency
DDOS	Penafian Perkhidmatan Teragih (DDOS)
DL	Pembelajaran Dalam (Deep Learning)
DNN	Rangkaian Neural Dalam (Deep Neural Network)
DOS	Penafian Perkhidmatan (Denial of Service)
DRL	Pembelajaran Pengukuhan Dalam (Deep Reinforcement Learning)
DT	Pokok Keputusan (Decision Tree)
ENN	Disunting Jiran Terdekat (Edited Nearest Neighbour)
ERT	Pokok Rawak Ekstrim (Extremely Randomized Trees)
GB	Peningkatan Kecerunan (Gradient Boosting)
IDS	Sistem Pengesanan Pencerobohan (Intrusion Detection System)
KDD99	Knowledge Discovery in Databases (1999)
KNN	Jiran Terdekat K (K-Nearest Neighbour)
LR	Regresi Logistik (Logistic Regression)
MLP	Perseptron Berbilang Lapisan (Multilayer Perceptron)
NB	Naives Bayes
NIDS	Sistem Pengesanan Pencerobohan Rangkaian (Network Intrusion Detection System)

NN	Rangkaian Neural (Neural Network)
NSL-KDD	Network Security Laboratory-Knowledge Discovery in Databases
PSO	Pengoptimuman Partikel Swarm (Particle Swarm Optimization)
R2L	Remot ke Lokal (Remote to Local)
RF	Hutan Rawak (Random Forest)
RUS	Pensampelan Rawak Kurang (Random Undersampling)
SMOTE	Teknik Pensampelan Sintetik Minoriti Lebih (Synthetic Minority Oversampling Technique)
SMOTE-ENN	SMOTE-Edited Nearest Neighbour
SMOTE-RUS	Synthetic Minority Oversampling Technique – Random Undersampling
SMOTE-Tomek	SMOTE-Tomek Links (Pautan Tomek – SMOTE)
SVM	Mesin Sokongan Vektor (Support Vector Machine)
U2R	Pengguna ke Pangkal (User to Root)
XGBoost	Peningkatan Kecerunan Ekstrim (eXtreme Gradient Boosting)

BAB I

PENGENALAN

1.1 PENDAHULUAN

Era digital telah membawa tahap kesalinghubungan dan pertukaran data yang tidak pernah berlaku sebelum ini, membolehkan organisasi dan individu berkomunikasi dan berurus niaga pada skala global. Kesalinghubungan ini, walaupun sangat bermanfaat, telah membuka pintu kepada pelbagai ancaman dan serangan keselamatan siber. Perlindungan aset digital dan maklumat sensitif telah menjadi kebimbangan utama bagi perniagaan, kerajaan dan individu. Akibatnya, bidang keselamatan siber telah muncul sebagai disiplin kritikal yang tertumpu pada melindungi rangkaian, sistem dan data daripada akses tanpa kebenaran, pelanggaran data dan aktiviti berniat jahat (Abu-Alhaija 2020).

Dalam konteks keselamatan siber, pengesanan pencerobohan rangkaian memainkan peranan penting dalam mengenal pasti dan mengurangkan ancaman kepada infrastruktur rangkaian. Sistem Pengesanan Pencerobohan Rangkaian (NIDS) direka bentuk untuk memantau trafik rangkaian dan mengenal pasti corak atau anomali yang mungkin menunjukkan kemungkinan percubaan pencerobohan. Mengesan dan bertindak balas terhadap ancaman sedemikian adalah penting untuk mencegah kebocoran data, gangguan perkhidmatan dan pelanggaran keselamatan siber yang lain (Ahmad et al. 2021).

1.2 LATAR BELAKANG KAJIAN

Set data NSL-KDD (*Network Security Laboratory-Knowledge Discovery in Databases*), yang diperoleh daripada set data asal KDD Cup 99 (*Knowledge Discovery*

in Databases Cup 1999), telah menjadi penanda aras standard untuk menilai prestasi teknik pengesanan pencerobohan.

Dalam kajian Al-Khassawneh (2023), beliau membincangkan pelaksanaan Sistem Pengesanan Pencerobohan (IDS) terhadap Penyedia Perkhidmatan Terurus (*Managed Service Providers)* untuk meningkatkan keselamatan dengan mengesan potensi ancaman. Kajian ini menekankan nilai set data NSL-KDD dalam menangani isu dalam set data KDD Cup 99 dengan meneroka dan membandingkan teknik pengesanan pencerobohan, antaranya Jiran Terdekat K (KNN), Mesin Sokongan Vektor (SVM) dan Hutan Rawak (RF). Hasil kajian menggunakan set data NSL-KDD, model yang dicadangkan iaitu gabungan algoritma RF dan pemilihan fitur berjaya meningkatkan ketepatan IDS dan membuka peluang kepada beliau untuk meneroka model pembelajaran mesin yang lain.

Kajian Dinesh dan Kalaivani (2023) menggunakan set data NSL-KDD sebagai penanda aras untuk menilai keberkesanan Sistem Pengesanan Pencerobohan (IDS). Algoritma meta-heuristik adalah elemen kecerdasan pengiraan, terutamanya digunakan untuk penyelesaian masalah pengoptimuman yang kompleks (Abdel-Basset et al. 2018). Algoritma meta-heuristik ini berkesan dalam mengoptimumkan model sistem pengesanan pencerobohan (IDS) dengan mencapai kadar ketepatan sebanyak 96% dan menekankan peranan pembelajaran mesin dalam menangani cabaran keselamatan siber.

Kajian oleh Shukla dan Sharma (2023) memfokuskan pada meningkatkan prestasi IDS untuk analisis trafik rangkaian, dengan menyedari bahawa kerumitan trafik rangkaian memerlukan analisis yang teliti. Untuk meningkatkan keberkesanan dan kecekapan IDS, kajian ini menggunakan gabungan teknik Hutan Rawak (RF) dan pemilihan ciri Pengoptimuman Partikel *Swarm* (PSO). Pelbagai pengelas seperti Jiran Terdekat K (KNN), Mesin Sokongan Vektor (SVM), Regresi Logistik (LR), Pokok Keputusan (DT) dan *Naives Bayes* (NB) digunakan untuk mengukur berbilang metrik IDS. Apabila digunakan pada set data NSL-KDD, kaedah gabungan RF dan PSO secara berkesan mengurangkan hingar dalam kadar penggera, meningkatkan kadar pengelasan dan meningkatkan prestasi IDS. Parameter prestasi menunjukkan peningkatan yang ketara dalam ketepatan (99.26%) dan pengurangan dimensi data melalui PSO.

Dalam kajian oleh Malik dan Saini (2023) pula, ia menilai keberkesanan Pembelajaran Pengukuhan Dalam (DRL) dalam menangani cabaran yang dihadapi oleh sistem pengesanan pencerobohan rangkaian (NIDS). Penyelesaian yang dicadangkan melibatkan penyepaduan Pembelajaran Pengukuhan Adversarial/Ejen Pelbagai dengan Pembelajaran Dalam-Q (AE-DQN) untuk mengesan anomali rangkaian. Penilaian pendekatan ini dijalankan menggunakan set data NSL-KDD dan melaporkan skor F1 sebanyak 79% dan ketepatan 80% untuk kaedah yang dicadangkan. Kajian yang dibentangkan oleh Pandey et al. (2023) dalam menangani isu serangan penafian perkhidmatan (DOS), beliau mencadangkan sistem berasaskan pembelajaran dalam (DL) untuk mengesan serangan penafian perkhidmatan teragih (DDOS). Mereka telah menggunakan algoritma Regresi Logistik (LR), Jiran Terdekat K (KNN) dan Hutan Rawak (RF) serta menilai model mereka menggunakan set data NSL-KDD. Keputusan menunjukkan bahawa model cadangan mereka adalah sangat tepat dalam mengesan serangan DDOS. Antara model-model yang diuji, KNN (98.79%) dan RF (99.4%) menunjukkan ketepatan tertinggi.

Tianyao, Huadong et al. (2023) membincangkan kepentingan pengesanan trafik yang tidak normal dalam keselamatan *internet* kerana ancaman serangan berniat jahat. Mereka mencadangkan kaedah untuk pengelasan pencerobohan menggunakan pemilihan ciri dan pengelas Mesin Sokongan Vektor (SVM) dengan set data NSL-KDD. Kaedah ini bertujuan untuk meningkatkan ketepatan pengelasan dengan mengurangkan ciri input melalui pemilihan fitur dalam NSL-KDD. Keputusan eksperimen menunjukkan teknik yang dicadangkan mencapai ketepatan klasifikasi 88.25% dengan KDDTest+ dan 72.42% dengan KDDTest-21.

Set data NSL-KDD mengandungi pelbagai jenis data trafik rangkaian, termasuk data normal dan serangan, yang merangkumi pelbagai kategori serangan. Walaupun set data NSL-KDD berfungsi sebagai sumber yang berharga untuk menilai model pengesanan pencerobohan, ia memberikan beberapa cabaran yang perlu ditangani (Sapre et al. 2019).

1.2.1 Cabaran Set Data NSL-KDD

Antara cabaran set data NSL-KDD (*Network Laboratory Lab – Knowledge Discovery in Databases)* yang memberi kesan kepada pembangunan dan prestasi model pengesanan pencerobohan adalah:

1. Ketidakseimbangan Kelas: Taburan data normal melebihi bilangan data serangan. Ketidakseimbangan kelas ini boleh membawa kepada prestasi model bias kerana algoritma akan condong ke arah kelas majoriti, menjadikannya sukar untuk membangunkan model yang tepat untuk mengesan serangan jarang (minoriti) dan mengakibatkan pengesanan serangan yang lemah.

2. Data Bising (*Noise*): Set data ini berkemungkinan mengandungi ciri bising yang boleh menimbulkan kekeliruan dan menghalang keupayaan model untuk membezakan corak serangan yang bermakna. Ia mungkin masih mengandungi hingar disebabkan ralat dalam pelabelan, kerumitan yang wujud dan kekaburan corak trafik rangkaian atau kehadiran ciri yang tidak berkaitan.

3. Evolusi Serangan: Dengan landskap serangan keselamatan siber yang kerap berubah dan penyerang siber yang sentiasa mengembangkan taktik mereka untuk mengelak mekanisme pertahanan sedia ada, set data ini tidak merangkumi semua kemungkinan senario serangan.

Cabaran yang wujud pada set data NSL-KDD menekankan keperluan untuk teknik yang efektif yang boleh meningkatkan prestasi model pengesanan pencerobohan. Terdapat tiga teknik dalam menangani masalah ketidakseimbangan data. Antaranya, pada peringkat data, peringkat algoritma, dan pendekatan hibrid. Secara ringkasnya, teknik pada peringkat data menggunakan pendekatan pensampelan semula, teknik pada peringkat algoritma menggunakan pendekatan penalaan algoritma dan pendekatan hibrid iaitu penggabungan beberapa teknik untuk membentuk pendekatan yang baharu. Teknik-teknik ini akan dihuraikan dalam Bab 2 : Kajian Literatur.

Motivasi di sebalik kajian ini adalah untuk menangani masalah ketidakseimbangan kelas terutamanya kelas minoriti dalam set data NSL-KDD dan meningkatkan peratusan pengesanan kelas minoriti. Data yang tidak seimbang

merupakan topik yang penting dan relevan dalam pembelajaran mesin, terutamanya untuk aplikasi dunia sebenar di mana data sering menunjukkan ketidakseimbangan kelas. Dalam dunia sebenar, data tidak seimbang merupakan masalah yang sentiasa wujud dalam pelbagai domain, seperti pengesanan penipuan, diagnosis perubatan dan pengesanan anomali. Mengabaikan ketidakseimbangan kelas boleh membawa kepada model yang berat sebelah terhadap kelas majoriti, menjadikannya tidak berkesan dalam aplikasi praktikal (Japkowicz et al. 2002). Dalam sesetengah aplikasi, membuat ralat dalam kelas minoriti menjadikan kos ralat mahal berbanding kelas majoriti. Sebagai contoh, dalam diagnosis perubatan, salah mendiagnosis penyakit yang jarang berlaku boleh membawa akibat yang teruk. Memberi tumpuan pada data yang tidak seimbang membantu dalam mengurangkan ralat yang mahal ini (Elkan 2001).

Menangani ketidakseimbangan kelas adalah langkah penting dalam mengurangkan bias dalam model pembelajaran mesin. Jika tidak dikendalikan dengan betul, data yang tidak seimbang boleh membawa kepada ramalan berat sebelah (Chawla et al. 2004). Algoritma yang direka khusus untuk mengendalikan data tidak seimbang, seperti Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE) dan pembelajaran kos sensitif (*Cost-Sensitive Learning*), boleh meningkatkan prestasi model pembelajaran mesin pada set data tidak seimbang dengan ketara (Chawla et al. 2002). Model yang dilatih mengenai data tidak seimbang boleh menjadi lebih teguh dalam mengendalikan variasi dan kawasan luaran *(outlier)* dalam senario dunia sebenar, kerana mereka telah belajar untuk membuat generalisasi daripada contoh terhad dalam kelas minoriti (Kubat et al. 1997). Selain itu, menganalisis dan menangani ketidakseimbangan kelas boleh membawa kepada pemahaman yang lebih mendalam tentang taburan data dan faktor yang menyumbang kepada ketidakseimbangan, dan boleh menghasilkan fitur dan reka bentuk model yang lebih baik (Batista et al. 2004).

Pematuhan terhadap kawal selia memerlukan isu ketidakseimbangan data ditangani. Dalam sesetengah industri seperti kewangan dan penjagaan kesihatan, terdapat keperluan kawal selia untuk memastikan model tidak berat sebelah terhadap mana-mana kumpulan tertentu. Menangani ketidakseimbangan kelas adalah langkah penting dalam memenuhi piawaian pematuhan ini (Obermeyer et al. 2019). Ringkasnya, memberi tumpuan kepada data yang tidak seimbang bukan sekadar daripada

pertimbangan teknikal, ianya mempunyai implikasi dunia sebenar dari segi keadilan, keberkesanan kos dan prestasi model. Kajian terkini secara konsistennya menekankan kepentingan menangani ketidakseimbangan kelas untuk membina model pembelajaran mesin yang lebih berkesan (Aburbeian et al. 2023; Ameur et al. 2023; Napoli et al. 2023).

1.3 PERMASALAHAN KAJIAN

Menangani ketidakseimbangan kelas adalah cabaran dalam membangunkan sistem pengesanan pencerobohan kerana algoritma pembelajaran mesin tradisional cenderung memberikan prestasi yang lemah pada set data yang tidak seimbang disebabkan kecenderungannya terhadap kelas majoriti. Pengagihan kelas yang tidak seimbang sering mengakibatkan prestasi ramalan yang kurang optimum, di mana Sistem Pengesanan Pencerobohan (IDS) cenderung untuk mengutamakan kelas majoriti yang membawa kepada pengesanan yang kurang optimum bagi serangan baharu, jarang berlaku atau serangan kritikal.

Untuk mengatasi cabaran ini, kajian ini akan memberi tumpuan untuk menangani masalah ketidakseimbangan kelas dalam set data NSL-KDD dengan menggunakan teknik pensampelan semula bagi meningkatkan prestasi model pengelasan dengan mengelaskan data yang tidak seimbang.

Terdapat pelbagai pendekatan pensampelan semula yang telah dijalankan dalam kajian-kajian terdahulu. Dalam kajian yang dijalankan oleh Japkowicz dan Stephen (2002) berpendapat bahawa kaedah pensampelan lebih (*oversampling)* dan pensampelan kurang (*undersampling)* dapat memelihara dan mengekalkan maklumat kelas minoriti berbanding menggunakan kaedah penalaan algoritma. Mereka juga mendapati dengan menggunakan pendekatan penalaan algoritma, ia lebih cenderung untuk mewujudkan pemadanan berlebihan (*overfitting*) apabila berurusan dengan data yang tidak seimbang. Kaedah pensampelan semula secara langsung dapat memanipulasi set data dan memastikan maklumat yang terkandung dalam kelas minoriti dikekalkan.

Dalam menangani ketidakseimbangan kelas melalui teknik penalaan algoritma, ia mungkin tidak mencukupi untuk menghapuskan bias dalam ramalan model. Kajian oleh Liu et al. (2009) mendapati bahawa penggunaan kaedah pensampelan semula boleh mengurangkan bias dalam keputusan pengelasan dengan ketara. Sebaliknya, penyelesaian melalui kaedah algoritma akan menghasilkan ramalan yang bias jika tidak menala algoritma dengan teliti. Selain itu, pensampelan semula boleh membantu meningkatkan keupayaan generalisasi oleh model pembelajaran mesin dan mengurangkan kesan ketidakseimbangan kelas pada prestasinya.

Kajian yang dilakukan oleh Chawla et al. (2002) menunjukkan bahawa Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE) secara konsisten berupaya meningkatkan prestasi generalisasi model pengelasan terhadap set data tidak seimbang berbanding set data yang tidak melalui proses pensampelan semula. Kaedah pensampelan semula juga menawarkan lebih fleksibiliti dan kebolehsuaian kepada ciri khusus set data. Kajian oleh Batista et al. (2004) mendapati kaedah pensampelan semula boleh disesuaikan kepada tahap ketidakseimbangan yang berbeza dalam sesuatu set data dan boleh diaplikasikan dalam pelbagai senario dunia sebenar, manakala penyelesaian secara pendekatan algoritma mempunyai fleksibiliti terhad dalam hal ini.

1.3.1 Teknik Pensampelan Semula

Seperti yang telah dibincangkan dalam Seksyen 1.3, teknik pensampelan semula merupakan fokus kajian ini. Antara teknik pensampelan semula yang masih relevan digunakan sehingga kini adalah teknik pensampelan lebih (*oversampling*) iaitu Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE). SMOTE adalah teknik yang menghasilkan sampel sintetik untuk kelas minoriti dengan interpolasi antara titik data yang sedia ada. Teknik SMOTE sehingga kini masih berjaya membuktikan bahawa ia dapat memberikan prestasi pengelasan yang lebih baik, seperti yang dibuktikan dalam kajian Aryuni et al. (2023), dimana prestasi pengelas yang menggunakan teknik SMOTE memberikan hasil yang lebih tinggi berbanding dengan hanya menggunakan pengelas tunggal. Ia juga berjaya meningkatkan sensitiviti pengelasan kelas minoriti.

Teknik pensampelan kurang (*undersampling*) pula mengurangkan bilangan contoh kelas majoriti untuk mencapai taburan kelas yang lebih seimbang. Kajian oleh

Hancock et al. (2022) membuktikan bahawa dengan menggunakan sejenis teknik pensampelan rawak kurang iaitu (RUS), prestasi pengelasan model Pokok Rawak Ekstrim (ET) dan Peningkatan Kecerunan Ekstrim (XGBoost) lebih baik. Masa yang diperlukan untuk melatih model pengelasan juga berkurang apabila RUS digunakan.

Teknik gabungan pensampelan lebih dan pensampelan kurang juga digunakan dalam mengatasi ketidakseimbangan kelas. Contohnya teknik penggabungan SMOTE-RUS (*Synthetic Minority Oversampling Technique – Random Undersampling*) yang dikaji oleh (Ismail et al. 2023). Kajian ini bertujuan mengklasifikasikan set data Gangguan Spektrum Autisme (*Autism Spectrum Disorder*) yang tidak seimbang. Eksperimen menggunakan model SMOTE-RUS yang dicadangkan mencapai ketepatan sekitar 88% menggunakan pengelasan Jiran Terdekat K (KNN). Ini adalah peningkatan dari ketepatan 79% yang dicapai menggunakan teknik pensampelan rawak kurang (RUS) sahaja. Selain itu, ketika menggunakan teknik pembelajaran mesin penggabungan Peningkatan Kecerunan (GB) dengan SMOTE-RUS, model tersebut mencapai ketepatan yang lebih tinggi iaitu kira-kira 90.5%. Ini menunjukkan keberkesanan teknik SMOTE-RUS dalam menangani set data yang tidak seimbang untuk meramalkan gen Gangguan Spektrum Autisme.

Berdasarkana kajian lepas, Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE), teknik Pensampelan Rawak Kurang (RUS) dan SMOTE-RUS (*Synthetic Minority Oversampling Technique – Random Undersampling*) memberikan keputusan yang meyakinkan. Oleh itu, secara khususnya, kajian ini bertujuan untuk menyiasat keberkesanan teknik pensampelan semula tersebut dalam mengatasi ketidakseimbangan set data NSL-KDD. Kajian ini akan membandingkan kesan SMOTE, RUS dan SMOTE-RUS terhadap prestasi model ramalan dengan mengambil kira aspek seperti ketepatan (*accuracy*), kepersisan (*precision*), rekal (*recall*), skor F1 (*F1-Score*).

Dengan merangka strategi yang berkesan untuk menangani ketidakseimbangan menggunakan teknik pensampelan semula, kajian ini bertujuan untuk menyediakan penyelesaian yang boleh meningkatkan ketepatan dan kebolehpercayaan pengelas serangan sistem pengesanan pencerobohan dalam domain keselamatan siber terhadap kelas minoriti.

1.4 PERSOALAN KAJIAN

1. Adakah teknik pensampelan semula antara SMOTE, RUS, dan SMOTE-RUS, berkesan dalam mengurangkan ketidakseimbangan kelas dalam dataset NSL-KDD sekaligus berkesan meningkatkan peratusan pengesanan?

2. Adakah pendekatan hibrid SMOTE-RUS memberikan peningkatan yang ketara berbanding teknik pensampelan semula individu SMOTE dan RUS?

1.5 OBJEKTIF KAJIAN

Berikut merupakan objektif kajian ini:

1. Mengimbangi set data NSL-KDD dengan menggunakan beberapa teknik pensampelan semula SMOTE, RUS dan SMOTE-RUS.

2. Menilai kesan SMOTE, RUS dan SMOTE-RUS terhadap prestasi model.

3. Membandingkan keputusan yang diperoleh menggunakan teknik SMOTE, RUS dan SMOTE-RUS.

1.6 KEPENTINGAN KAJIAN

Kepentingan kajian ini terletak pada pelaksanaan teknik pensampelan semula, iaitu Teknik Pensampelan Minoriti Lebih (SMOTE), teknik Pensampelan Rawak Kurang (RUS), dan SMOTE-RUS, sebagai penyelesaian yang berpotensi untuk menangani isu ketidakseimbangan kelas dalam set data NSL-KDD. Ketidakseimbangan kelas adalah cabaran yang berterusan dalam pembelajaran mesin dan perlombongan data, di mana satu kelas mengatasi kelas yang lain dengan ketara. Ini berpotensi membawa kepada model yang bias dan kurang tepat. Dalam konteks set data NSL-KDD, yang digunakan secara meluas untuk pengesanan pencerobohan dalam keselamatan siber, isu ketidakseimbangan boleh menghalang pengesanan anomali dalam Sistem Pengesanan Pencerobohan (IDS) yang penting untuk memastikan keselamatan rangkaian.

Dengan membandingkan teknik pensampelan semula, kajian ini bertujuan untuk membandingkan teknik pensampelan semula dalam keberkesanannya untuk

mengurangkan masalah ketidakseimbangan kelas dalam set data NSL-KDD menggunakan pendekatan pada peringkat data. Kajian perbandingan ini boleh memberi perspektif dalam memilih strategi pensampelan semula untuk meningkatkan ketepatan sistem IDS. Memahami kaedah pensampelan semula yang paling berkesan boleh membawa kepada model yang lebih dipercayai dan teguh, dan meningkatkan sistem IDS yang lebih mantap dan pengesanan yang efektif supaya ia boleh mengurangkan negatif palsu dan positif palsu dengan ketara. Selain itu, kajian ketidakseimbangan ini bukan sahaja menyumbang dalam skop pengesanan pencerobohan tetapi juga sesuai dalam domain lain seperti diagnosis perubatan, pengesanan penipuan dan lain-lain.

1.7 SKOP DAN BATASAN KAJIAN

Skop kajian ini adalah untuk menilai dan membandingkan keberkesanan teknik-teknik pensampelan semula yang fokus kepada setiap contoh teknik dalam kategori pensampelan semula yang berbeza, iaitu teknik pensampelan lebih SMOTE, teknik pensampelan kurang RUS, dan teknik hibrid SMOTE-RUS dalam menangani isu ketidakseimbangan kelas dalam set data NSL-KDD. Set data NSL-KDD merupakan set data yang digunakan sebagai penanda aras yang terkenal dalam domain keselamatan siber, dan ia mengandungi ketidakseimbangan kelas yang ketara antara kelas serangan rangkaian dan kelas normal. Dengan menggunakan teknik pensampelan semula ini, kajian ini bertujuan untuk meningkatkan prestasi algoritma pembelajaran mesin dalam mengelaskan data serangan dan data normal serta meningkatkan pengesanan kelas minoriti. Objektif utama adalah untuk mengukur keberkesanan teknik pensampelan semula ini pada prestasi model, ketepatan (*accuracy*), kepersisan (*precision*), rekal (*recall*) dan skor F1 (*F1-Score*).

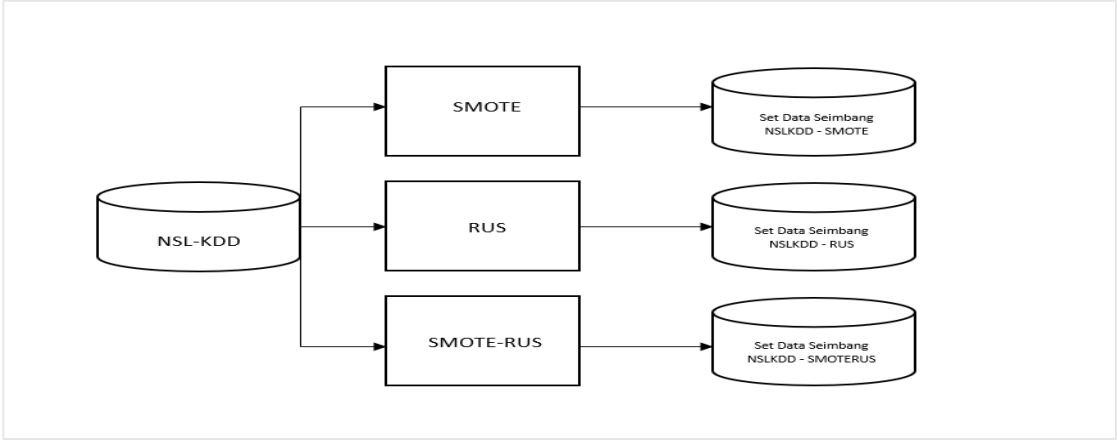
Walaubagaimanapun, beberapa batasan perlu diakui dalam kajian ini. Pertama, hasil penemuan mungkin terhad kepada set data NSL-KDD. Set data NSL-KDD mempunyai batasannya yang tersendiri dan hasil yang diperoleh mungkin tidak boleh digunakan secara langsung pada set data lain yang mempunyai ciri yang berbeza dengan set data NSL-KDD. Kedua, walaupun terdapat banyak teknik-teknik lain yang ada, kajian ini hanya memilih tiga teknik pensampelan semula dan tidak meneroka teknik-teknik yang lain yang mungkin akan memberi keputusan pengelasan yang lebih tepat dan dapat mengesan kelas minoriti dengan lebih berkesan. Selain itu, kajian ini hanya

memfokuskan pada SMOTE, RUS, dan SMOTE-RUS sebagai teknik pensampelan semula dan berkemungkinan terdapat kaedah alternatif yang menghasilkan keputusan yang berbeza. Ketiga, pilihan metrik penilaian juga terhad. Penilaian prestasi yang dipilih adalah berdasarkan set metrik yang standard, dan hasil kajian hanya terhad kepada metrik standard ini. Akhir sekali, kajian ini tidak mengambil kira kemungkinan variasi dalam prestasi pengelasan pembelajaran mesin yang digunakan bersama dengan kaedah pensampelan semula. Model pengelas yang berbeza mungkin bertindak balas secara berbeza terhadap set data yang telah di sampel semula. Batasan-batasan ini akan muncul dalam menjalankan kajian ini.

Namun begitu, kajian ini berhasrat dapat memberikan pandangan tentang kekuatan dan kelemahan setiap teknik pensampelan semula yang dipilih dalam konteks set data NSL-KDD dan dapat menyumbang idea dan cadangan dalam domain keselamatan siber yang memfokuskan kepada kepentingan keseimbangan data dalam membina model ramalan serangan yang optimum.

1.8 KESIMPULAN

Secara ringkasnya, bab pengenalan ini adalah asas permulaan untuk kajian ini dengan membincangkan kepentingan menangani ketidakseimbangan kelas dalam set data NSL-KDD dalam konteks keselamatan rangkaian. Carta alir pelaksanaan teknik pensampelan data ke atas NSL-KDD dapat dilihat dalam Rajah 1.1. Pelaksanaan teknik pensampelan data yang menggunakan teknik SMOTE, RUS dan SMOTE-RUS ke atas NSL-KDD bagi mendapatkan set data yang seimbang selari dengan objektif yang telah digariskan dalam kajian ini.



Rajah 1.1 Carta Alir Pelaksanaan Teknik Pensampelan Data

Bab ini juga telah merumuskan soalan kajian dan membincangkan skop dan batasan kajian. Dalam bab-bab seterusnya, kajian ini akan membincangkan kajian literatur, metodologi, eksperimen dan analisis yang dijalankan untuk mencapai objektif kajian ini secara menyeluruh.

BAB II

KAJIAN LITERASI

2.1 PENGENALAN

Kajian literasi ini membincangkan penemuan dan kajian yang berkaitan dengan pengesanan pencerobohan, set data NSL-KDD (*Network Security Laboratory-Knowledge Discovery in Databases*) dan masalah ketidakseimbangan kelas data. Membincangkan kaedah dan teknik dalam kajian terdahulu dalam menangani cabaran ketidakseimbangan kelas terhadap set data NSL-KDD. Selain itu, ia meneroka strategi yang digunakan untuk mengoptimumkan prestasi model pengesanan pencerobohan, dengan penekanan khusus kepada aplikasi teknik-teknik pensampelan semula iaitu Teknik Pensampelan Minoriti Lebih (SMOTE), teknik Pensampelan Rawak Kurang (RUS) dan gabungan teknik SMOTE-RUS.

2.2 KESELAMATAN SIBER DAN KEPENTINGANNYA

Keselamatan Siber merangkumi polisi dan teknologi yang direka untuk melindungi sistem komputer, rangkaian dan data daripada pencerobohan. Ia merupakan bahagian penting dalam dunia digital yang semakin berkembang dan mempunyai implikasi yang meluas kepada individu, organisasi dan kerajaan.

Keselamatan siber amat penting dalam melindungi segala aset dan maklumat daripada penceroboh. Pertama, ia melibatkan perlindungan data sensitif. Ia memastikan kerahsiaan, integriti dan ketersediaan maklumat. Kedua, keselamatan siber mempunyai implikasi ekonomi yang besar, kerana serangan siber boleh membawa kepada kerugian kewangan akibat pencerobohan data dan gangguan perkhidmatan kritikal (Saeed et al. 2023). Ketiga, soal keselamatan negara, serangan siber terhadap infrastruktur kritikal seperti grid kuasa dan sistem kesihatan akan mengancam kepada kesejahteraan negara

(Montasari 2023). Selain itu, privasi adalah aspek asas keselamatan siber, jika diceroboh, ia berpotensi menjejaskan maklumat peribadi. Akhir sekali, disebabkan rangkaian kesalinghubungan global yang berlaku atas talian, keselamatan siber telah menjadi kebimbangan global, yang memerlukan kerjasama antarabangsa (Arogundade).

2.2.1 Jenis Ancaman dan Serangan Siber

Ancaman serangan siber terdiri dalam pelbagai bentuk yang mempunyai ciri-ciri yang dan motivasi tersendiri. Perisian perosak (*malware)* yang merangkumi virus, cecacing komputer (*worm),* Trojan dan perisian tebusan (*ransomware),* direka untuk menyusup dan merosakkan sistem atau mencuri data. Serangan pancingan data (*phishing)* pula menipu individu untuk mendedahkan maklumat sensitif selalunya melalui e-mel atau laman web yang dikompromi. Serangan penafian perkhidmatan (DOS) dan penafian perkhidmatan teragih (DDOS) bertujuan untuk membanjiri sistem atau rangkaian dengan trafik yang boleh menjadikan sistem tersebut tidak boleh diakses kepada pengguna-pengguna yang sah. Ancaman orang dalam melibatkan serangan atau pencerobohan data yang dilakukan oleh individu dalam organisasi, selalunya didorong oleh hal peribadi atau kebencian. Ancaman berterusan lanjutan (APT) adalah serangan yang berpanjangan dan tersembunyi yang dijalankan oleh kumpulan penceroboh yang mahir. Ia selalunya berkaitan dengan motivasi politik atau ekonomi (Moore 2023).

2.2.2 Sistem Pengesanan Pencerobohan (IDS)

Pengesanan pencerobohan adalah komponen penting dalam strategi keselamatan siber. Sistem Pengesanan Pencerobohan (IDS) direka untuk mengenal pasti dan bertindak balas terhadap aktiviti yang mencurigakan atau berniat jahat dalam rangkaian atau sistem.

Terdapat dua jenis IDS. IDS yang berasaskan tandatangan *(signature)* dan berasaskan anomali (*anomaly*) . IDS berasaskan tandatangan bergantung kepada corak yang telah ditetapkan atau serangan yang diketahui. Manakala IDS berasaskan anomali mengesan serangan berdasarkan garis dasar yang ditetapkan. Sistem IDS menyediakan pemantauan masa nyata, menjejak trafik rangkaian dan aktiviti sistem secara berterusan

dan pengesanan segera terhadap potensi ancaman. Walau bagaimanapun, cabaran utama dalam IDS adalah meminimumkan positif palsu dan negatif palsu. Sebagai tindak balas kepada cabaran ini, teknik pembelajaran mesin telah diterapkan dalam IDS, bagi meningkatkan keupayaan untuk menyesuaikan diri dengan ancaman yang semakin berkembang dan mengurangkan positif palsu melalui pengesanan anomali (Abdulganiyu et al. 2023).

2.2.3 Peranan IDS dalam Keselamatan Siber

Sistem IDS memainkan peranan penting dalam mengekalkan keselamatan dan integriti sistem dan rangkaian komputer. Mereka bertindak sebagai mekanisme pertahanan proaktif, membolehkan organisasi bertindak balas secara berkesan terhadap ancaman siber. IDS memudahkan pengesanan ancaman awal, mengenal pasti serangan pada peringkat awalnya, membenarkan organisasi bertindak balas sebelum kerosakan yang ketara berlaku. Selain itu, log IDS menyediakan data berharga untuk analisis forensik apabila berlaku insiden keselamatan (Hassan et al. 2023). Tambahan pula, banyak rangka kerja dan piawaian kawal selia mewajibkan penggunaan IDS dalam perlindungan data dan pematuhan polisi keselamatan siber untuk menggariskan peranan penting IDS dalam strategi keselamatan siber moden (Rawindaran et al. 2023).

Kajian literatur ini memberikan pemahaman asas tentang keselamatan siber dan peranan penting sistem IDS. Dalam konteks set data NSL-KDD, ia menggambarkan senario pencerobohan rangkaian di mana pengesanan pencerobohan yang berkesan adalah penting menjadikan usaha kajian ini sangat relevan.

2.3 SET DATA NSL-KDD

Set data NSL-KDD (*Network Security Laboratory-Knowledge Discovery in Databases*) adalah set data yang sering digunakan dalam bidang keselamatan siber dan pengesanan pencerobohan rangkaian. Ia diperkenalkan pada tahun 2009 oleh Tavallaee, Bagheri et al. Ia adalah hasil penambahbaikan daripada set data asal KDD Cup 99 (*Knowledge Discovery in Databases 1999*) yang direka dengan teliti untuk menangani kekurangan dan kerumitan yang wujud pada set data ini.

Salah satu objektif utama set data NSL-KDD direka adalah untuk mengurangkan ketidakseimbangan kelas dan isu data berlebihan *(redundancy)* yang melanda set data KDD Cup 99. Set data asal menimbulkan cabaran dalam menilai sistem pengesanan pencerobohan (IDS) secara realistik disebabkan oleh perwakilan trafik rangkaian dan serangannya yang tidak tepat dalam menggambarkan serangan realiti. Hasilnya, NSL-KDD telah direka dengan teliti untuk mimik tingkah laku rangkaian dengan lebih tepat.

Salah satu ciri penting set data NSL-KDD ialah pengkategorian komprehensif trafik rangkaiannya. Ia mengklasifikasikan data kepada lima kategori berbeza: Normal, Penafian Perkhidmatan (DOS), Penerokaan (*Probe*), Remot ke Lokal (R2L) dan Pengguna ke Pangkal (U2R). Pengkategorian jenis serangan secara terperinci ini membolehkan analisis yang lebih mendalam terhadap pencerobohan rangkaian dan memberi pemahaman yang lebih mendalam tentang ancaman keselamatan.

Untuk meningkatkan kualiti set data, rekod berlebihan daripada set data asal KDD Cup 99 telah dikeluarkan dan menghasilkan set data yang lebih seimbang dan mewakili trafik rangkaian realiti. Pengurangan dalam data berlebihan ini merupakan langkah penting ke arah mewujudkan penanda aras dan piawai yang lebih dipercayai dalam penyelidikan pengesanan pencerobohan. Mimikan sebenar dalam trafik rangkaian adalah satu lagi ciri set data NSL-KDD. Dengan meniru tingkah laku rangkaian sebenar, set data ini adalah pilihan ideal untuk menilai algoritma pengesanan pencerobohan oleh kerana ia menggambarkan trafik rangkaian sebenar (Mahfouz 2021; Moustafa 2017; Zhao et al. 2022).

Set data NSL-KDD dibahagikan kepada subset latihan dan ujian, membolehkan penyelidik menilai prestasi model mereka pada data yang tidak kelihatan sebelum ini. Pengasingan data kepada set berbeza ini sejajar dengan amalan terbaik dalam pembelajaran mesin dan penyelidikan pengesanan pencerobohan. Selain itu, set data ini mempunyai pelbagai ciri yang merangkumi kedua-dua sifat numerikal dan nominal. Ciri ini mewakili pelbagai aspek trafik rangkaian, termasuk protokol (*protocol),* servis *(service)*, penanda *(flags)* dan banyak lagi. Ciri ini memberi penyelidik untuk meneroka pelbagai pemboleh ubah dan atribut yang memberikan fleksibiliti dalam

membangunkan dan menilai model pengesanan pencerobohan. Walaubagaimanapun, set data NSL-KDD masih terdapat ketidakseimbangan kelas. Ketidakseimbangan yang wujud ini membolehkan set data ini diuji menggunakan pelbagai teknik melalui teknik pensampelan, algoritma dan hibrid yang bertujuan untuk menangani isu ketidakseimbangan data dalam NSL-KDD ini (Xu et al. 2021).

Set data NSL-KDD merupakan alat penting dalam penyelidikan bidang keselamatan rangkaian dan pengesanan pencerobohan kerana pengkategorian kelasnya, struktur berasaskan kelas, pengurangan data berlebihan dan gambaran realistik trafik rangkaian. Berdasarkan Jadual 2.1, berikut merupakan jumlah saiz sampel kelas dalam set data NSL-KDD untuk data KDDTrain+ dan KDDTest+. Kelas serangan terbahagi dua iaitu kepada normal dan anomali, kelas anomali adalah kelas serangan. Ia terbahagi kepada empat kategori iaitu DOS, *Probe*, U2R dan R2L.

Jadual 2.1 Saiz Sampel Kelas NSL-KDD

Kelas	**Saiz Sampel KDDTrain+**	**Saiz Sampel KDDTest+**
Normal	67,343	9,711
Dos	45,927	7,458
Probe	11,656	2,421
R2L	995	2,654
U2R	52	200

Ketidakseimbangan data dalam NSL-KDD disebabkan tiga kelas minoriti yang telah dikenal pasti iaitu *Probe*, R2L dan U2R, manakala kelas majoriti terdiri daripada rekod normal dan DOS. Terdapat empat jenis serangan dalam set data ini iaitu:

1. Penafian Perkhidmatan (DOS): Penggodam membanjiri sistem komputer dengan menghantar trafik rangkaian sehingga sistem komputer sasaran tidak dapat diakses.

2. Penerokaan *(Probing)*: Penggodam mendapatkan maklumat sebanyak yang mungkin dengan meneroka sesuatu rangkaian atau sistem komputer untuk

melangkaui keselamatan yang wujud di dalam rangkaian atau sistem komputer itu.

3. Remot ke Lokal (R2L): Penggodam berjaya mendapatkan akses pangkal (*root*) dalam sistem komputer mangsa secara remot dengan cara mengeksploitasi sistem komputer.
4. Pengguna ke Pangkal (U2R): Penggodam menggunakan akaun mangsa dan mengeksploit kelemahan di dalam sistem untuk mendapatkan akses ke peringkat yang lebih tinggi seperti akses ke pangkal sistem.

Dalam set data NSL-KDD, terdapat 8 jenis data set yang berlainan yang terdiri daripada:

1. KDDTrain+.ARFF: Set data latihan yang penuh dengan label binari dalam format ARFF iaitu format data yang hanya digunakan untuk perisian WEKA sahaja.
2. KDDTrain+.TXT: Set data latihan penuh termasuk label serangan dan tahap kompleks serangan dalam format CSV.
3. KDDTrain+_20Percent.ARFF: Subset 20% daripada fail KDDTrain+.arff dalam format ARFF iaitu format data yang hanya digunakan untuk perisian WEKA sahaja.
4. KDDTrain+_20Percent.TXT: Subset 20% subset daripada KDDTrain+.txt file
5. KDDTest+.ARFF: Set data ujian penuh dengan label binari dalam format ARFF iaitu format data yang hanya digunakan untuk perisian WEKA sahaja.
6. KDDTest+.TXT: Set data ujian penuh termasuk label serangan dan tahap kompleks serangan dalam format CSV.
7. KDDTest-21.ARFF: Subset daripada fail KDDTest+.arff file yang tidak mempunyai tahap kompleks serangan . dalam format ARFF iaitu format data yang hanya digunakan untuk perisian WEKA sahaja.
8. KDDTest-21.TXT: Subset daripada fail KDDTest+.txt file yang tidak mempunyai tahap kompleks serangan.

Jadual 2.2 menunjukkan 41 kategori fitur dalam set data NSL-KDD.

Jadual 2.2 41 Kategori Fitur NSL-KDD

Kategori	Huraian	Jenis
duration	Jumlah masa sambungan.	*Integer*
protocol_type	Jenis protokol (cth., TCP, UDP).	*String*
service	Perkhidmatan rangkaian pada destinasi (cth., HTTP, FTP).	*String*
flag	Status masa sambungan (cth., S0, S1, S2).	*String*
src_bytes	Jumlah bait data dari sumber ke destinasi.	*Integer*
dst_bytes	Jumlah bait data dari destinasi ke sumber.	*Integer*
land	1 jika sambungan dari/ke host/port yang sama; 0 sebaliknya.	*Binary*
wrong_fragment	Jumlah fragmen "salah".	*Integer*
urgent	Jumlah paket '*urgent'*.	*Integer*
count	Bilangan sambungan ke host yang sama dalam 2 saat terakhir.	*Integer*
srv_count	Bilangan sambungan ke perkhidmatan yang sama dalam 2 saat terakhir.	*Integer*
serror_rate	Peratusan sambungan dengan kesalahan "SYN".	*Float*
srv_serror_rate	Peratusan sambungan dengan kesalahan "SYN" ke perkhidmatan yang sama.	*Float*
rerror_rate	Peratusan sambungan dengan kesalahan "REJ".	*Float*
srv_rerror_rate	Peratusan sambungan dengan kesalahan "REJ" ke perkhidmatan yang sama.	*Float*
same_srv_rate	Peratusan sambungan ke perkhidmatan yang sama.	*Float*
diff_srv_rate	Peratusan sambungan ke perkhidmatan yang berbeza.	*Float*
srv_diff_host_rate	Peratusan sambungan ke host yang berbeza.	*Float*
dst_host_count	Bilangan sambungan dari IP sumber yang sama.	*Integer*
dst_host_srv_count	Bilangan sambungan ke perkhidmatan yang sama dari IP sumber yang sama.	*Integer*
dst_host_same_srv_rate	Peratusan sambungan ke perkhidmatan yang sama dari host yang sama.	*Float*
dst_host_diff_srv_rate	Peratusan sambungan ke perkhidmatan yang berbeza.	*Float*
dst_host_srv_diff_host_rate	Peratusan sambungan ke host yang berbeza dari perkhidmatan yang sama.	*Float*
dst_host_serror_rate	Peratusan sambungan dengan kesalahan "SYN" dari host yang sama.	*Float*
dst_host_srv_serror_rate	Peratusan sambungan dengan kesalahan "SYN" ke perkhidmatan yang sama.	*Float*
dst_host_rerror_rate	Peratusan sambungan dengan kesalahan "REJ" dari host yang sama.	*Float*
dst_host_srv_rerror_rate	Peratusan sambungan dengan kesalahan "REJ" ke perkhidmatan yang sama.	*Float*
hot	Jumlah indikator "hot" dalam sambungan.	*Integer*
	bersambung...	
...sambungan		
num_failed_logins	Bilangan percubaan log masuk yang gagal.	*Integer*

logged_in	1 jika berjaya log masuk; 0 sebaliknya.	*Binary*
num_compromised	Jumlah keadaan "terkompromi".	*Integer*
root_shell	1 jika mendapat shell root; 0 sebaliknya.	*Binary*
su_attempted	1 jika percubaan perintah "su root"; 0 sebaliknya.	*Binary*
num_root	Jumlah akses root.	*Integer*
num_file_creations	Jumlah operasi pembuatan fail.	*Integer*
num_shells	Jumlah prompt shell.	*Integer*
num_access_files	Jumlah operasi pada fail kawalan akses.	*Integer*
num_outbound_cmds	Jumlah arahan keluar dalam sesi FTP.	*Integer*
is_host_login	1 jika log masuk termasuk dalam senarai "host"; 0 sebaliknya.	*Binary*
is_guest_login	1 jika log masuk adalah sebagai "tetamu"; 0 sebaliknya.	*Binary*

Dalam set data ini, 41 fitur boleh digunakan sebagai penunjuk serangan atau trafik biasa. Trafik rangkaian biasa mempunyai corak yang boleh diramal seperti menggunakan protokol biasa seperti HTTP atau FTP, kekerapan interaksi yang konsisten dan saiz data yang dijangkakan. Contoh fitur adalah seperti fitur *duration, protocol_type, services* dan *flags* membantu mengenal pasti trafik biasa. Manakala trafik serangan akan bercanggah dengan corak trafik biasa. Ia akan menunjukkan tingkah laku yang luar biasa dalam fitur seperti berapa percubaan yang berlaku untuk memasuki sesuatu sistem seperti *num_failed_logins, num_compromised* dan *su_attempted* dan peningkatan mendadak dalam percubaan sambungan daripada trafik luar atau kadar ralat yang tinggi boleh menandakan serangan.

Jadual 2.3 menunjukkan peratusan yang ada dalam KDDTrain+ dan KDDTest+.

Jadual 2.3 Jumlah Rekod NSL-KDD

Set Data	Jumlah	Normal	DOS	Probe	U2R	R2L
KDDTrain+	125973	67343 (53%)	45927 (37%)	11656 (9.11%)	52 (0.04%)	995 (0.85%)
KDDTest+	22544	9711 (43%)	7458 (33%)	2421 (11%)	200 (0.9%)	2654 (12.1%)

Jadual 2.3, rekod data Normal dan DOS dapat dilihat mendominasi set data ini dengan mengambil separuh daripada jumlah rekod dalam set data ini yang menjadikannya kelas majoriti. Manakala kelas minoriti dilihat secara ketara dipegang oleh *Probe*, U2R dan R2L. Perbezaan peratusan yang sangatlah jauh menjadi punca ketidakseimbangan kelas dalam NSL-KDD yang membawa kepada permasalahan dalam membina model pengelasan.

2.3.2 Kajian Terdahulu Menggunakan Set Data NSL-KDD

Set data NSL-KDD (*Network Security Laboratory-Knowledge Discovery in Databases*) telah digunakan secara meluas dalam bidang pengesanan pencerobohan dan keselamatan rangkaian, berfungsi untuk menilai keberkesanan pelbagai teknik pengesanan pencerobohan. Penyelidik telah memanfaatkan set data ini untuk menilai prestasi sistem pengesanan pencerobohan (IDS), meneroka fitur dan menangani isu ketidakseimbangan kelas. Bahagian ini mengetengahkan beberapa kajian yang menggunakan set data NSL-KDD.

Satu aplikasi bagi set data NSL-KDD telah menjadi penanda aras untuk menilai sistem pengesanan pencerobohan. Dalam kajian perbandingan, Rashid et al. (2020) menilai prestasi IDS menggunakan set data NSL-KDD. Keputusan eksperimen menunjukkan bahawa pengelas Jiran Terdekat K (KNN), Mesin Sokongan Vektor (SVM), Rangkaian Neural (NN), dan Rangkaian Neural Dalam (DNN) mencapai kira-kira 100% ketepatan dalam metrik penilaian prestasi apabila digunakan pada set data NSL-KDD. Sebaliknya, pengelasan KNN dan *Naïve Bayes* mencapai kira-kira 99% ketepatan apabila diuji pada set data CIDDS-001 (Rashid et al. 2020). Walaubagaimanapun, keputusan kajian tersebut berkemungkinan mempunyai pemadanan berlebihan (*overfitting*) dan generalisasi model yang lemah. Model ketepatan tinggi dalam kajian tersebut mendorong penyiasatan lanjut tentang kesempurnaan nilai keputusan yang dihasilkan apabila berhadapan dengan set data yang tidak seimbang, yang biasa berlaku dalam aplikasi dunia sebenar.

Ketidakseimbangan kelas merupakan cabaran biasa dalam penyelidikan pengesanan pencerobohan, di mana kategori serangan tertentu kurang diwakili dalam set data. Untuk menangani isu ini, Liu, Wang et al. (2020) membangunkan pendekatan teknik hibrid yang disesuaikan dengan set data yang tidak seimbang. RF, SVM, XGBoost, Memori Jangka Pendek Panjang (LSTM), AlexNet dan Mini-VGGNet dibangunkan untuk model pengelasan. Hasil kajian menunjukkan keberkesanan gabungan teknik pembelajaran dalam *Difficult Set Sampling Technique* (DSSTE) dan AlexNet berbanding model pengelasan tradisional dalam meningkatkan ketepatan pengesanan dalam set data NSL-KDD.

Pemilihan fitur telah menjadi titik fokus penyelidikan menggunakan set data NSL-KDD. Hakim, Fatma et al. (2019) menjalankan analisis tentang kepentingan ciri dan memperkenalkan kaedah pemilihan ciri (*feature selection*) dan kriteria pemisahan (s*plitting criterion), The Information Gain* dan *Gain Ratio*. Eksperimen menggunakan pemilihan ciri, *Chi-squared*, dan *ReliefF* dan kriteria pemisahan *The Information Gain* dan *Gain Ratio*. Matlamatnya adalah untuk meningkatkan prestasi model dengan mengalih keluar data yang tidak relevan atau berlebihan, mengurangkan pemasangan berlebihan dan mengurangkan masa latihan. Pendekatan ini meningkatkan kecekapan dan ketepatan model pengesanan pencerobohan mereka dengan ketara, menekankan kepentingan kejuruteraan fitur dalam mereka bentuk IDS yang berkesan (Hakim et al. 2019)

Selain itu, set data NSL-KDD juga memainkan peranan dalam penerokaan teknik pembelajaran mesin yang baharu dan keberkesanan mengesan serangan anomali di mana penyelidik menggunakan set data NSL-KDD sebagai panduan dalam menguji keberkesanan teknik-teknik yang ingin diuji seperti kajian yang dilakukan oleh Pai dan Adesh (2021). Kajian sedemikian menyumbang kepada pembangunan IDS yang mampu menyesuaikan diri dengan ancaman yang muncul dalam persekitaran rangkaian dinamik.

Kajian menggunakan set data NSL-KDD sebagai penanda aras dalam menguji prestasi pelbagai algoritma pembelajaran mesin dan teknik pengesanan pencerobohan. Dalam kajian yang dilakukan oleh Madwanna et al. (2023) memperkenalkan dua model IDS berasaskan pembelajaran dalam (DL) iaitu YARS-IDS dan YARS-IDS-II. YARS-IDS diuji pada set data UNSW-NB15 dan NSL-KDD dan menggunakan teknik SMOTE untuk meningkatkan prestasi, terutamanya dalam kelas minoriti. YARS-IDS mencapai ketepatan pengelasan masing-masing sebanyak 82.19% dan 98.87% untuk UNSW-NB15 dan NSL-KDD. YARS-IDS-II, diuji pada NSL-KDD, mencapai ketepatan pengelasan 98.8%. Keputusan ini mengatasi pendekatan yang menggunakan algoritma pembelajaran mesin tradisional seperti Mesin Sokongan Vektor (SVM).

Kesimpulannya, set data NSL-KDD memainkan peranan penting dalam membantu penyelidikan dalam pengesanan pencerobohan dan keselamatan rangkaian

sehingga kini. Penyelidik telah memanfaatkan fitur, perwakilan trafik rangkaian yang realistik, dan kategori serangan yang pelbagai. Kajian-kajian diatas membuktikan set data NSL-KDD mempunyai kredibiliti dalam menilai prestasi IDS, mengurangkan ketidakseimbangan kelas dan sekali gus dapat membantu dalam pembangunan sistem pengesanan pencerobohan yang optimum dalam domain keselamatan siber dan rangkaian.

2.4 KETIDAKSEIMBANGAN KELAS

Ketidakseimbangan kelas adalah isu yang meluas dan penting dalam bidang pembelajaran mesin dan analisis data. Dengan implikasi di pelbagai domain seperti penjagaan kesihatan, kewangan, pengesanan penipuan, dan lain-lain. Perkara ini merujuk kepada situasi di mana taburan kelas dalam set data sangat cenderung kepada satu kelas sahaja iaitu kelas majoriti dan ini akan memberi kesan kepada kelas minoriti. Dalam set data yang tidak seimbang seperti itu, kelas minoriti biasanya membentuk bahagian kecil daripada data keseluruhan, sementara kelas majoriti mendominasi. Ketidakseimbangan ini mencipta cabaran besar bagi membina model ramalan dan pengelasan kerana keberkesanan algoritma pembelajaran mesin akan cenderung untuk tidak memberikan hasil pengelasan yang optimum (Hernandez et al. 2013).

Sebagai contoh, dalam set data pengesanan penipuan kad kredit, sebahagian besar urus niaga adalah sah (kelas negatif), manakala hanya sebahagian kecil mewakili aktiviti penipuan (kelas positif). Begitu juga, dalam diagnosis perubatan, penyakit dengan kelaziman yang rendah boleh membawa kepada set data yang sangat tidak seimbang, di mana kelas majoriti terdiri daripada pesakit yang sihat, dan kelas minoriti termasuk mereka yang mempunyai penyakit dengan kelaziman yang rendah.

2.4.1 Cabaran Ketidakseimbangan Kelas Terhadap Pembelajaran Mesin

Menangani ketidakseimbangan kelas dalam pembelajaran mesin melibatkan beberapa cabaran besar. Satu cabaran utama adalah bias yang terjadi apabila satu kelas melebihi jumlah yang lain secara ketara yang membawa kepada bias algoritma, bias penilaian, bias perwakilan dan bias pensampelan. Bias algoritma timbul kerana kebanyakan algoritma mengoptimumkan ketepatan keseluruhan, namun ketepatan tersebut

kebiasaannya mengabaikan kelas minoriti. Bias penilaian adalah kebimbangan utama kerana metrik tradisional seperti ketepatan tidak menggambarkan prestasi dengan tepat pada set data tidak seimbang, memerlukan metrik alternatif seperti kepersisan, rekal dan skor F1.

Bias perwakilan berlaku kerana kelas minoriti yang kurang diwakili boleh dianggap sebagai bunyi bising, menjejaskan keupayaan model untuk mempelajari ciri-cirinya. Model-model yang dilatih dengan set data yang tidak seimbang cenderung ke kelas majoriti, dan akan menghasilkan ketepatan yang lebih tinggi walaupun prestasi ramalan model terhadap kelas minoriti kurang baik (Almogahed 2014; Johnson 2016). Bias ini boleh memudaratkan di dalam aplikasi dunia sebenar sekiranya kedua-dua kelas adalah sama pentingnya, seperti diagnosis perubatan, pengesanan pencerobohan atau penipuan dan lain-lain.

Cabaran lain adalah generalisasi terhadap kelas minoriti. Oleh kerana terdapat sedikit contoh kelas minoriti yang boleh dipelajari oleh model, ia mungkin menghadapi kesukaran dalam mengelaskan contoh secara tepat. Akibatnya, model hanya mempunyai pemahaman dan pengalaman yang kurang terhadap kelas minoriti, menyebabkan kadar kesalahan negatif palsu yang lebih tinggi untuk contoh tersebut (Thabtah et al. 2020). Dalam senario kritikal seperti diagnosis perubatan, ini boleh memberi kesan serius.

Selain itu, penilaian prestasi model menjadi rumit dengan ketidakseimbangan kelas. Metrik seperti ketepatan *(accuracy)* boleh memberikan keputusan yang tidak tepat kerana model hanya meramalkan kelas majoriti untuk semua contoh. Ia mungkin mencapai ketepatan yang tinggi tetapi pada hakikatnya tidak memberikan keputusan yang sebenar (García et al. 2015; Johnson et al. 2019). Untuk mengatasi cabaran ini, metrik alternatif seperti kepersisan *(precision),* rekal *(recall),* skor F1 *(F1-Score)* dan kawasan di bawah lengkung (AUC) perlu digunakan supaya dapat memberikan penilaian yang lebih menyeluruh terhadap keupayaan model untuk mengklasifikasikan kedua-dua kelas majoriti dan minoriti (Raeder et al. 2012).

Pembelajaran corak dalam data yang tidak seimbang juga adalah cabaran yang sukar. Perwakilan terhadap kelas minoriti yang terhad membuatkan model sukar untuk memahami kerumitan yang ada dalam contoh minoriti. Hasilnya, model mungkin tidak berprestasi dengan baik pada kelas minoriti(Dou et al. 2023; Petersen et al. 2023).

Untuk mengurangkan cabaran ini, pelbagai teknik boleh digunakan, termasuk teknik pensampelan semula seperti pensampelan kurang dan pensampelan lebih, menggunakan metrik penilaian yang berbeza, memilih algoritma yang sesuai, pembelajaran berasaskan kos sensitif (*cost-sensitive learning),* penghasilan data sintetik, dan kaedah-kaedah hibrid. Walau bagaimanapun, pemilihan teknik atau gabungan teknik yang sesuai bergantung kepada ciri-ciri set data yang spesifik dan objektif penggunaan teknik pembelajaran mesin. Mengatasi ketidakseimbangan kelas adalah penting untuk memastikan bahawa model memberikan ramalan yang adil untuk semua kelas.

2.4.2 Teknik – Teknik Menangani Ketidakseimbangan Kelas

Ketidakseimbangan kelas merupakan cabaran yang sentiasa ada dalam pembelajaran mesin di mana taburan contoh kelas yang berat sebelah, yang membawa kepada prestasi model yang tidak optimum. Untuk menangani isu ini, beberapa teknik telah dibangunkan dan dikaji secara meluas. Teknik tersebut dikategorikan kepada pendekatan pada peringkat data, algoritma dan pendekatan hibrid.

1. **Pendekatan pada peringkat data adalah dengan menggunakan teknik pensampelan semula**. Ia bertujuan untuk mengimbangi taburan kelas dengan mengubah suai set data latihan. Terdapat tiga jenis utama teknik pensampelan semula:
    a) Pensampelan Lebih *(Oversampling)*: Meniru kejadian daripada kelas minoriti untuk memadankan saiz kelas majoriti. Walaupun kaedah ini meningkatkan perwakilan kelas minoriti, kaedah ini boleh menyebabkan pemadanan berlebihan (*overfitting*). Contoh: Teknik Pensampelan Minoriti Lebih (SMOTE).

b) Pensampelan Kurang *(Undersampling)*: Mengalih keluar kejadian daripada kelas majoriti untuk memadankan saiz kelas minoriti. Pendekatan ini mengurangkan saiz data latihan tetapi boleh mengakibatkan kehilangan maklumat berharga. Contoh: Pensampelan Rawak Kurang (RUS).

c) Teknik Hibrid : Kaedah pensampelan semula hibrid menggabungkan kedua-dua pensampelan lebih dan pensampelan kurang untuk mengimbangi set data. Contohnya SMOTENN dan SMOTE-Tomek.

i. SMOTENN (*SMOTE-Edited Nearest Neighbour*). Teknik ini menggunakan SMOTE untuk mencipta sampel sintetik kelas minoriti, dan kemudian menggunakan algoritma *Disunting Jiran Terdekat* (ENN) untuk membersihkan set data dengan mengalih keluar kejadian yang berkemungkinan bunyi bising atau kawasan luar (*outlier)*.

ii. SMOTE-Tomek (*SMOTE-Tomek Links*): Sama seperti SMOTENN, ia menggabungkan SMOTE dengan Tomek Links, yang merupakan pasangan contoh jiran terdekat dari kelas bertentangan. Kejadian daripada kelas majoriti dalam Tomek Links dialih keluar untuk menyediakan set data yang lebih bersih dan berasingan.

2. **Pendekatan peringkat algoritma menggunakan pendekatan algoritma dengan menyesuaikan algoritma pembelajaran untuk menangani ketidakseimbangan.** Pendekatan peringkat algoritma melibatkan pengubahsuaian algoritma pembelajaran untuk mempertimbangkan ketidakseimbangan kelas semasa latihan. Antara contoh:

a) Mesin Sokongan Vektor (SVM) dengan pemberat kelas melaraskan proses mengoptimumkan SVM untuk mengambil kira ketidakseimbangan kelas.

b) Pembelajaran sensitif kos *(Cost-Sensitive Learning):* Pembelajaran sensitif kos adalah cara mengajar algoritma untuk membuat keputusan

sambil mengambil kira bahawa akan ada beberapa kesilapan yang mempunyai kos akibat yang besar berbanding yang lain. Dalam pembelajaran sensitif kos, boleh menetapkan kos atau penalti yang berbeza untuk setiap kesilapan. Pembelajaran sensitif kos berguna dalam situasi di mana setiap kesilapan mendapat penalti yang berbeza, dan model pembelajaran mesin akan memberi lebih perhatian untuk mengurangkan penalti kepada kesilapan yang mempunyai kos yang mahal semasa proses latihannya.

3. **Pendekatan hibrid menggabungkan pelbagai teknik untuk meningkatkan pengendalian ketidakseimbangan kelas.** Contohnya:

a) Kaedah penggabungan berbilang pengelas asas untuk meningkatkan prestasi ramalan keseluruhan. Untuk mengendalikan ketidakseimbangan kelas, kaedah hibrid akan tertumpu dalam meningkatkan kelas minoriti. Teknik yang terkenal termasuk Hutan Rawak Seimbang dan Pelengkap Mudah.

i. Hutan Rawak Seimbang *(Balanced Random Forests):* Pengubahsuaian Hutan Rawak ini memberikan pemberat yang berbeza kepada kelas berdasarkan tahap ketidakseimbangannya, dengan itu memberi lebih kepentingan kepada kelas minoriti.

ii. *Pelengkap Mudah (EasyEnsemble)*: Kaedah pelengkap ini mencipta berbilang subset bagi kelas majoriti dan menggabungkannya dengan keseluruhan kelas minoriti untuk membentuk set latihan yang seimbang, kemudian melatih pengelas asas pada set ini.

Jadual 2.4 merupakan jadual perbandingan antara teknik data, teknik algoritma dan teknik hibrid.

Jadual 2.4 Perbandingan Teknik Data, Algoritma dan Hibrid

Kaedah Data	**Kaedah Algoritma**	**Kaedah Hibrid**
• Teknik-teknik resampling mengubah taburan kelas dataset dengan oversampling minoriti atau undersampling majoriti.	• Kaedah algoritma mengubah algoritma pembelajaran untuk mengendalikan ketidakseimbangan kelas.	• Kaedah hibrid menggabungkan teknik resampling dan algoritma.
• Mudah untuk dilaksanakan • Tiada perubahan kepada algoritma pembelajaran. • Berkesan untuk set data tertentu yang tidak seimbang.	• Disesuaikan untuk algoritma tertentu. • Lebih sofistikated mengurus data yang tidak seimbang.	• Menggabungkan kelebihan kedua-dua kaedah resampling dan algoritma. • Prestasi yang lebih baik dalam sesetengah kes.
• Potensi kehilangan maklumat disebabkan oleh *undersampling*. • Risiko *overfitting* dalam *oversampling*.	• Mungkin memerlukan penyesuaian tertentu yang berkaitan dengan algoritma. • Tidak semestinya berkesan untuk semua algoritma.	• Kompleksiti dalam pelaksanaan dan penalaan.

Berdasarkan perbandingan dalam Jadual 2.4, kaedah data lebih mudah dilaksanakan kerana ia menangani masalah data tidak seimbang daripada data itu sendiri berbanding kaedah algoritma yang menangani masalah data tidak seimbang melalui penalaan algoritma. Manakala kaedah hibrid pula melalui gabungan dua kaedah lebih kompleks.

Jadual 2.5 menunjukkan perbandingan kajian lepas dalam menyeimbang data tidak seimbang NSL-KDD menggunakan tiga teknik iaitu teknik data, algoritma dan hibrid. Perbandingan tiga kajian ini menunjukkan bahawa peratusan ketepatan pengelas tidak begitu ketara dalam pengelasan data serangan. Namun begitu, melalui perbandingan yang dilakukan dalam Jadual 2.4 dan Jadual 2.5, teknik pensampelan data masih lagi berada dalam peratusan kedudukan yang meyakinkan untuk eksperimen dalam kajian ini dilaksanakan.

Jadual 2.5 Perbandingan Kajian Teknik Data, Algoritma dan Hibrid

Teknik	Kajian	Keputusan
Data (Massaoudi et al. 2022)	Perbandingan peratusan model pengelasan untuk mengatasi ketidakseimbangan NSL-KDD dengan teknik SMOTE	Hasil eksperimen memberi peratusan ketepatan yang tinggi (99.79%) apabila menggunakan model SMOTE dengan ERT berbanding dengan menggunakan model tradisional tunggal seperti SVM dan MLP.
Algoritma (Venkata Abhiram et al. 2023)	Menggunakan teknik gabungan SVM dan ANN untuk mengatasi ketidakseimbangan NSL-KDD	Hasil eksperimen menunjukkan teknik asing SVM (97%) dan ANN (95%) masih memberi ketepatan yang lebih kurang berbanding teknik gabungan SVM-ANN (95%).
Hibrid (Abdelkhalek et al. 2023)	Menggunakan teknik ADASYN dan Tomek-Links bersama DL untuk mengatasi ketidakseimbangan NSL-KDD (Abdelkhalek et al. 2023)	Hasil eksperimen menunjukkan peningkatan yang signifikan dalam kadar pengesanan kelas minoriti, mencapai ketepatan sebanyak 99.8% dalam klasifikasi binari dan 99.98% dalam klasifikasi pelbagai kelas.

Kesimpulannya, menangani ketidakseimbangan kelas adalah penting untuk membina model pembelajaran mesin yang tepat dan boleh dipercayai, terutamanya dalam domain yang mempunyai banyak contoh kelas minoriti. Penyelidik telah membangunkan pelbagai teknik untuk mengurangkan ketidakseimbangan kelas, masing-masing dengan kekuatan dan kelemahannya. Pilihan teknik bergantung pada set data khusus dan masalah yang dihadapi, dan eksperimen perlu dijalankan untuk menentukan pendekatan yang paling berkesan. Kajian ini memfokus untuk teknik pensampelan data semula bagi menangani masalah asas iaitu mengimbangi set data yang tidak seimbang.

2.5 TEKNIK PENSAMPELAN SEMULA

Teknik pensampelan semula yang digunakan secara meluas dalam statistik dan pembelajaran mesin memainkan peranan penting dalam memanipulasi atau menjana sampel data baharu daripada set data sedia ada. Tujuan utama adalah untuk menangani pelbagai cabaran yang berkaitan dengan data, seperti set data tidak seimbang dan pemadanan berlebihan (*overfitting). S*eperti yang dihuraikan dalam Seksyen 2.4.2, terdapat beberapa pendekatan untuk menangani ketidakseimbangan data dan kajian ini akan menggunakan teknik pensampelan semula. Antara teknik pensampelan semula yang biasa digunakan dalam analisis data dan pembelajaran mesin, iaitu teknik pensampelan lebih, pensampelan kurang dan hibrid. Kajian ini menggunakan Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE) untuk pensampelan lebih, untuk teknik pensampelan kurang pula adalah Pensampelan Rawak Kurang (RUS*)*, manakala teknik hibrid yang dipilih adalah SMOTE dan RUS (SMOTE-RUS). Teknik SMOTE-RUS melibatkan penjanaan tambahan sampel sintetik untuk kelas minoriti dan kemudian memperhalusi set data melalui teknik pensampelan kurang.

Pilihan teknik pensampelan semula tertentu bergantung pada ciri set data dan algoritma pembelajaran mesin yang digunakan. Adalah penting untuk menilai dengan teliti dan memilih teknik yang paling sesuai untuk mengelak daripada memperkenalkan bias atau pemadanan berlebihan (*overfitting).* Untuk mengelak daripada penghasilan data bias, adalah penting untuk memastikan bahawa proses pensampelan semula tidak mengubah taburan data asal. Secara ringkasnya, teknik pensampelan semula adalah salah satu teknik untuk menangani cabaran berkaitan data dalam pembelajaran mesin, terutamanya dengan set data yang tidak seimbang. Pertimbangan, percubaan dan penilaian yang teliti diperlukan untuk memilih pendekatan pensampelan semula yang paling sesuai untuk masalah tertentu sambil meminimumkan risiko bias dan pemadanan berlebihan (*overfitting).*

2.6 TEKNIK PENSAMPELAN SINTETIK MINORITI LEBIH (SMOTE)

Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE) adalah teknik penambahan data yang direka oleh Chawla Nitesh et al. (2002) pada tahun 2002 untuk menangani ketidakseimbangan kelas, masalah yang sering berlaku dalam pembelajaran mesin di mana satu kelas (kelas majoriti) mendominasi kelas lain (kelas minoriti) dalam set data. Ketidakseimbangan kelas boleh membawa kepada ramalan model yang bias, di mana kelas majoriti mendominasi dan kelas minoriti sering diabaikan. SMOTE bertujuan untuk menangani isu ini dengan mencipta sampel sintetik kelas minoriti, dengan itu mencapai set data yang lebih seimbang.

2.6.1 Kelebihan dan Kelemahan Teknik SMOTE

Jadual 2.6 merupakan kelebihan dan kelemahan teknik SMOTE. Kelebihan dan kelemahan teknik SMOTE ini telah dikaji terlebih dahulu dalam kajian-kajian lepas (Sihag et al. 2021; Wongvorachan et al. 2023).

Jadual 2.6 Kelebihan dan Kelemahan SMOTE

Aspek	Kelebihan	Kelemahan
Menyeimbangkan Kelas	Membantu menyeimbangkan pengedaran kelas dalam dataset yang tidak seimbang.	Tidak semestinya membawa kepada peningkatan prestasi model.
Meningkatkan Latihan Model	Meningkatkan keupayaan model untuk mempelajari kelas minoriti.	Boleh memperkenalkan bunyi atau kecenderungan berlebihan, terutamanya apabila digunakan secara berlebihan.
Tiada Kehilangan Data	Menghasilkan sampel sintetik dari data yang sedia ada, jadi tiada data yang hilang.	Mungkin tidak sesuai untuk semua dataset, terutamanya apabila kelas minoriti adalah berbeza secara inheren.
Sokongan Meluas	Dilaksanakan dalam pelbagai perpustakaan dan alat pembelajaran mesin.	Penalaan parameter sering diperlukan untuk mengoptimumkan prestasinya.
Aplikasi	Boleh diaplikasi dalam pelbagai algoritma pembelajaran mesin.	Mungkin tidak berfungsi dengan baik untuk dataset yang sangat tidak seimbang dengan nisbah ketidakseimbangan kelas yang besar.

2.7 PENSAMPELAN RAWAK KURANG (RUS)

Teknik Pensampelan Rawak Kurang (RUS) diperkenalkan oleh Batista et al. (2004) yang merupakan satu teknik yang digunakan dalam bidang pengelasan yang tidak seimbang untuk menangani masalah set data, di mana satu kelas mendominasi dengan ketara berbanding yang lain. Dalam teknik ini, satu subset dari contoh kelas majoriti dipilih secara rawak dan dikeluarkan dari set data untuk menyeimbangkan taburan kelas. Ini membantu mencegah model pembelajaran mesin dari menjadi bias kepada kelas majoriti dan boleh meningkatkan prestasi dalam hal pengelasan kelas minoriti dengan betul.

2.7.1 Kelebihan dan Kelemahan Teknik RUS

Berikut Jadual 2.7 merupakan kelebihan dan kelemahan teknik RUS (Liang et al. 2012; Liu 2004). Antara kelebihan dan kelemahan teknik RUS adalah seperti berikut:

Jadual 2.7 Kelebihan dan Kelemahan Teknik RUS

Aspek	Kelebihan	Kekurangan
Penyeimbangan Taburan Kelas	Pensampelan kurang membantu menyeimbangkan taburan kelas, memudahkan model pembelajaran mesin untuk memahami pola dalam kelas minoriti.	Mengurangi jumlah sampel kelas majoriti dapat menyebabkan kehilangan maklumat berharga, membuat model kurang kukuh.
Peningkatan Prestasi Model	Ini dapat menghasilkan peningkatan prestasi model, terutama ketika ketidakseimbangan kelas kritikal dan prestasi kelas majoriti mendominasi semasa latihan model.	Dalam situasi yang ada sedikit data, pensampelan kurang dapat lebih lanjut mengurangi size set data, yang dapat mengurangi prestasi model.
Pelatihan Lebih Cepat	Dengan sedikit rekod untuk diproses, proses latihan jadi lebih cepat, iaitu lebih bermanfaat untuk set data yang besar juga untuk model yang memerlukan proses komputasi yang intensif.	
Sederhana dan Mudah Diterapkan	Ini adalah teknik yang mudah untuk diimplementasikan dan difahami, menjadikannya Teknik yang baik untuk mengatasi ketidakseimbangan kelas.	

Walau bagaimanapun, RUS mempunyai risiko kehilangan maklumat yang berharga yang ada dalam sampel kelas majoriti yang telah dibuang, yang boleh mengakibatkan prestasi keseluruhan model yang berkurangan. Pertimbangan dan penalaan yang teliti diperlukan apabila melaksanakan pengurangan rawak untuk mencapai keseimbangan kelas dan mengekalkan integriti data.

2.8 TEKNIK PENSAMPELAN SINTETIK MINORITI LEBIH – PENSAMPELAN RAWAK KURANG (SMOTE-RUS)

Teknik Pensampelan Sintetik Minoriti Lebih – Pensampelan Rawak Kurang (SMOTE-RUS), menggabungkan teknik SMOTE dan teknik RUS yang diperkenalkan oleh Ubaya et al. (2020). Ia adalah pendekatan hibrid dalam pembelajaran mesin untuk menangani ketidakseimbangan kelas dalam masalah pengelasan. SMOTE berfungsi dengan mencipta sampel sintetik daripada kelas minoriti. Sebaliknya, RUS mengurangkan saiz kelas majoriti dengan menghapuskan sampel secara rawak. Dengan penggabungan ini, SMOTE-RUS mengimbangi set data dengan menambah kelas minoriti dan mengurangkan kelas majoriti, memastikan perwakilan kelas yang lebih saksama untuk latihan model dan ketepatan ramalan yang lebih baik.

2.8.1 Kelebihan dan Kelemahan Teknik SMOTE-RUS

Berikut Jadual 2.8 merupakan kelebihan dan kelemahan teknik SMOTE-RUS (Gad et al. 2023). Antara kelebihan dan kelemahan teknik SMOTE-RUS adalah seperti berikut:

Jadual 2.8 Kelebihan dan Kelemahan Teknik SMOTE-RUS

Aspek	Kelebihan	Kekurangan
Keseimbangan Data	Berkesan dalam menyeimbangkan pengedaran kelas dengan mengaugmentasi kelas minoriti dan mengurangkan kelas majoriti.	Kemungkinan kehilangan maklumat penting dari kelas majoriti akibat pensampelan kurang secara rawak.
Pemadanan Berlebihan	Mengurangkan risiko pemadanan berlebihan (*overfitting)*, masalah biasa apabila menggunakan teknik pensampelan lebih.	Penghasilan data sintetik kadangkala boleh membawa kepada generalisasi berlebihan jika tidak diurus dengan teliti.
Ketepatan Model	Boleh meningkatkan ketepatan model dengan menyediakan set data yang lebih seimbang untuk latihan.	Sifat sintetik data SMOTE mungkin tidak mewakili kompleksiti dunia sebenar.

bersambung...

...sambungan

Pelaksanaan	Agak mudah dilaksanakan dengan menggunakan pelbagai *library* dan perisian data sains.	Memerlukan penalaan yang teliti untuk mengelakkan bias dalam proses penghasilan data sintetik.
Kepelbagaian Data	Meningkatkan kepelbagaian kelas minoriti melalui penghasilan sampel sintetik.	Pensampelan bawah secara rawak mungkin mengakibatkan kurangnya perwakilan pola tertentu dari kelas majoriti.
Kos Komputasi	Secara umumnya lebih efisien daripada kaedah pengumpulan sampel atau data yang kompleks.	Kos komputasi yang lebih tinggi berbanding dengan teknik pensampelan yang biasa disebabkan oleh penghasilan data sintetik.

Dengan menggabungkan teknik SMOTE dan RUS, ia menawarkan keseimbangan yang baik antara meningkatkan perwakilan kelas minoriti dan mengurangkan dominasi kelas majoriti. Walaupun ia membawa risiko tertentu seperti kehilangan maklumat dari kelas majoriti dan keperluan untuk penalaan yang teliti, kelebihannya dalam meningkatkan kepelbagaian data dan mengurangkan pemadanan terlebih (*overfitting)* menjadikannya pilihan yang bagus untuk set data yang tidak seimbang.

2.8.2 Kajian Terdahulu Dengan Teknik Pensampelan SMOTE, RUS dan SMOTE-RUS

Kajian yang dilakukan oleh Bagui et al. (2023) menyiasat keberkesanan dua teknik pensampelan lebih, Sempadan SMOTE (BSMOTE) dan Mesin Sokongan Vektor - SMOTE (SVM-SMOTE), dalam analisis data keselamatan siber. Ia menumpukan pada penentuan nisbah pensampelan semula optimum untuk teknik ini untuk mempertingkatkan pengenalpastian kejadian serangan siber yang jarang berlaku dalam set data tidak seimbang, khususnya set data UNSW-NB15, yang mengandungi sebahagian kecil data serangan berbanding data normal. Eksperimen direka bentuk untuk menilai kesan perubahan peratusan pensampelan lebih dari 10% hingga pada ketepatan pengelasan. Nilai berbeza Jiran Terdekat K (KNN) juga telah diuji untuk menilai kesannya terhadap prestasi teknik pensampelan lebih. Keputusan menunjukkan bahawa kadar pensampelan berlebihan 10% menghasilkan prestasi yang lebih baik untuk kedua-dua BSMOTE dan SVM-SMOTE untuk pelbagai metrik seperti kepersisan, rekal dan skor-F.

Kajian oleh Mohammad Nasrul Aziz (2021) menyediakan analisis terperinci tentang keberkesanan RUS berasaskan kluster dalam meningkatkan sistem pengesanan pencerobohan. Kajian ini menggunakan pelbagai pengelas seperti *Naive Bayes*, Jiran Terdekat K (KNN), Mesin Sokongan Vektor (SVM), Perseptron Berbilang Lapisan (MLP), dan Hutan Rawak (RF) bersama set data NSL-KDD dan UNSW-NB15. Ia melaporkan peningkatan dalam ketepatan pengelasan, kepersisan dan metrik rekal dengan 95.6% menggunakan teknik RUS.

Berdasarkan kertas kerja "Penilaian Prestasi Kaedah Pengelasan dengan Pensampelan Hibrid untuk Data Tidak Seimbang: Kajian Simulasi Perbandingan," prestasi teknik SMOTE-RUS, walaupun relevan untuk pensampelan semula dalam konteks data tidak seimbang, menunjukkan hasil yang bercampur-campur jika dibandingkan dengan teknik pensampelan hibrid yang lain.

Kajian oleh Malek dan Yaacob (2021) menilai keberkesanan SMOTE-RUS terhadap kaedah lain seperti ROS-RUS (*Random Oversampling – Random Undersampling*), RSYN dan RACOG-RUS (*Rapidly Converging Gibbs sampler – Random Undersampling*) yang dicadangkan dan diuji dengan pengelas berbeza seperti Hutan Rawak (RF), Pokok Keputusan (DT) dan Peningkatan Kecerunan (GB). Bagi pengelas RF, SMOTE-RUS mencapai ralat salah pengelasan yang lebih rendah (0.2833) berbanding beberapa teknik, tetapi ia bukan kaedah berprestasi terbaik. Kaedah cadangan kajian, RACOG-RUS , mengatasi prestasi SMOTE-RUS dalam konteks ini. Untuk pengelas DT, SMOTE-RUS mempunyai ralat salah klasifikasi yang jauh lebih tinggi (0.6315) berbanding teknik lain, menunjukkan keberkesanan yang kurang dengan pengelas ini. Pengelas GB pula, SMOTE-RUS mencapai ralat salah yang kompeten berbanding kaedah lain. Walaubagaimanapun, teknik SMOTE-RUS masih kompeten dalam mengatasi ketidakseimbangan kelas walaupun keberkesanannya berbeza-beza bergantung kepada pengelas yang digunakan dan ciri khusus set data.

Oleh itu, berdasarkan kajian-kajian terdahulu, teknik pensampelan SMOTE, RUS dan SMOTE-RUS adalah teknik yang relevan dan boleh digunakan dalam mengatasi ketidakseimbangan kelas.

2.9 MESIN SOKONGAN VEKTOR (SVM)

Model pembelajaran mesin berguna untuk pelbagai aplikasi, tetapi keberkesanannya selalunya bergantung pada kualiti dan taburan data latihan. Dalam banyak senario dunia sebenar, set data boleh menjadi tidak seimbang, bermakna satu kelas dengan ketara mengatasi kelas yang lain. Ketidakseimbangan ini boleh membawa kepada prestasi model yang lemah, kerana algoritma cenderung memihak kepada kelas majoriti. Untuk mengurangkan isu ini, kaedah pensampelan semula boleh digunakan untuk mengimbangi set data. Walau bagaimanapun, adalah penting untuk memilih model pembelajaran mesin yang sesuai digunakan bersama teknik pensampelan semula ini.

Mesin Sokongan Vektor (SVM) adalah algoritma pembelajaran mesin yang efektif yang digunakan untuk tugas pengelasan dan regresi. SVM diperkenalkan oleh Vladimir N. Vapnik dan Alexey Ya. Chervonenkis pada tahun 1964. Ia telah mendapat popularitiyang meluas kerana keberkesanannya dalam pelbagai aplikasi dunia sebenar. Konsep asas SVM berkisar tentang mencari satah satah hiper (*hyperplane*) dalam ruang N-dimensi yang paling baik memisahkan titik data ke dalam kategori masing-masing. Satah hiper (*hyperplane*) ini bertindak sebagai sempadan keputusan yang mengkategorikan titik data baharu berdasarkan sisi satah yang mana ia jatuh. Satah hiper (*hyperplane)* ini memaksimumkan margin antara dua kelas, yang membantu meningkatkan keupayaan generalisasi algoritma (Gandhi 2018).

SVM boleh menyesuaikan kepada data bukan linear. Walaupun SVM pada mulanya direka untuk klasifikasi linear, ia boleh dilanjutkan untuk menangani masalah bukan linear dengan menggunakan teknik seperti kernel. Konsep kernel diperkenalkan oleh Bernhard Boser, Isabelle Guyon dan Vladimir Vapnik pada tahun 1992, yang sejak itu telah menjadi aspek asas SVM. Ini membolehkan SVM mengelaskan data yang tidak boleh dipisahkan secara linear, menjadikannya sangat berkesan daripada regresi logistik (LR).

Tambahan lagi, SVM menawarkan keteguhan kepada pemadanan berlebihan (*overfitting),* yang merupakan masalah biasa dalam pembelajaran mesin. Ini dicapai dengan mengoptimumkan margin antara kelas, dengan berkesan mengurangkan risiko menghafal data latihan dan meningkatkan generalisasi kepada data baharu yang tidak

kelihatan (Tanveer et al. 2022). Atribut ini telah didokumentasikan dengan baik dalam kajian dan telah menyumbang kepada populariti berterusan SVM dalam komuniti pembelajaran mesin.

Dalam kajian yang dilakukan oleh Hong, Horng et al. (2021), beliau membandingkan dengan enam algoritma iaitu Jiran Terdekat K (KNN), Pokok Keputusan (DT), Hutan Rawak (RF), *Naives Bayes*, Pokok Pokok Peningkatan Pencerunan (*Gradient Boosted Trees)* dan SVM ke atas set data NSL-KDD, dan mendapati algoritma SVM menduduki tangga ke-empat selepas DT,RF dan *Naives Bayes* dengan peratusan 98.2%. Berdasarkan kedudukan SVM, ia menunjukkan model pengelas SVM masih relevan walaupun ia adalah model pengelasan yang tertua yang mula di rekodkan dalam kajian pada 1964 (Vapnik et al. 1964) .

Keberkesanan SVM didokumentasikan dengan baik dalam banyak kajian dan aplikasi dunia sebenar. Satu aspek penting SVM ialah keteguhannya dalam mengendalikan data berdimensi tinggi. Ini amat penting dalam bidang seperti pengecaman imej dan genomik, di mana set data boleh mempunyai beribu-ribu malah berjuta-juta ciri (Prajapati et al. 2023) . SVM telah konsisten mengatasi algoritma pengelasan lain dalam konteks ini. Sebagai contoh, dalam kajian Rajpal et al. (2023), beliau mengakui bahawa SVM merupakan penanda aras yang terbaik untuk kajiannya kerana SVM menghasilkan keputusan yang sangat baik pada pelbagai set data penanda aras, mewujudkan keberkesanannya dalam pengkategorian teks, pengecaman digit tulisan tangan, diagnosis perubatan, pengelasan imej, analisis sentimen dan banyak lagi.

Dalam penyelidikan perubatan, SVM telah digunakan untuk mengelaskan pesakit berdasarkan data genetik, membantu mengenal pasti penanda penyakit dan meramalkan hasil pesakit (Myszczynska et al. 2020). Dalam klasifikasi imej, SVM telah digunakan untuk membezakan objek dalam imej, menyumbang kepada kemajuan dalam kenderaan autonomi, pengecaman muka dan banyak lagi (Myszczynska et al. 2020). Selain itu, dalam analisis sentimen, SVM telah cemerlang dalam menentukan sentimen data teks (Sadhasivam et al. 2019), menjadikannya alat yang berharga untuk memahami maklum balas pelanggan dan ulasan pengguna dalam talian.

Berurusan dengan data yang tidak seimbang melalui kaedah pensampelan semula adalah cabaran biasa dalam pembelajaran mesin. Pilihan model pembelajaran mesin yang betul adalah penting untuk mencapai hasil yang tepat dan boleh dipercayai. Pertimbangan sifat set data dan kaedah pensampelan semula khusus yang digunakan semasa memilih model. Dengan memadankan pilihan model dengan teliti dengan data dan teknik pensampelan semula, cabaran yang ditimbulkan oleh set data yang tidak seimbang dapat ditangani dengan membina model pembelajaran mesin yang mantap.

Kesimpulannya, SVM adalah algoritma pembelajaran mesin yang serba boleh dan sangat berkesan. Kekukuhannya dalam ruang berdimensi tinggi, kebolehsuaian kepada data bukan linear dan menunjukkan hasil yang optimum dalam pelbagai aplikasi dunia sebenar menjadikan mereka alat yang berharga dalam bidang data sains. Kajian lepas telah mengesahkan keberkesanannya, dan SVM terus menjadi algoritma yang penting dalam landskap pembelajaran mesin, menawarkan penyelesaian kepada masalah pengelasan dan regresi yang mencabar.

2.9.1 Kajian Terdahulu Dengan Mesin Sokongan Vektor (SVM)

Kajian oleh Manikandan et al. (2023) memberi tumpuan kepada peningkatan ketepatan dalam mendiagnos peringkat utama karsinoma sel skuamosa menggunakan pemilihan fitur dan teknik SMOTE. Empat pengelas, termasuk Pengemaskinian *Naives Bayes* (*Updatable Naive Bayes)*, Perseptron Berbilang Lapisan (MLP), Jiran Terdekat K (KNN), dan SVM digunakan untuk meramal diagnosis kanser mulut. Hasil kajian menunjukkan bahawa SVM sebanyak 96% mengatasi kaedah lain apabila digabungkan dengan pemilihan ciri dan SMOTE semasa pra pemprosesan.

Kajian oleh Yaqin et al. (2022) mengkaji ramalan tamat pengajian awal bagi pelajar dalam program pengajian sistem maklumat dan informatik di Universiti XYZ di Indonesia, berdasarkan peraturan dari 2014. Mereka mempertimbangkan faktor seperti Purata Nilai Gred (GPA), Jantina dan Umur untuk membangunkan model ramalan , termasuk Rangkaian Neural Buatan (ANN), Jiran Terdekat K (KNN) dan SVM. Cabarannya adalah menangani data yang tidak seimbang, yang ditangani menggunakan Teknik SMOTE. Selepas menggunakan SMOTE, ANN menunjukkan ketepatan ujian

terbaik pada 70.5%, manakala KNN mencapai 69.3% dan SVM 69.8%. Peningkatan paling ketara dalam nilai ingat semula dilihat dalam ANN, mencapai 71.3%.

Kajian oleh Krawczyk et al. (2021) menangani cabaran untuk belajar daripada data yang tidak seimbang, terutamanya dalam senario berbilang kelas. Penulis mencadangkan pendekatan baru yang melibatkan kaedah pensampelan dua langkah. Pertama, pengelas satu kelas dilatih untuk setiap kelas untuk mendapatkan vektor sokongan, yang kemudiannya digunakan sebagai wakil untuk kelas. Dalam langkah kedua, pendekatan pensampelan kurang digunakan pada vektor sokongan ini untuk menyeimbangkan lagi set data latihan. Teknik ini mengurangkan masa pengiraan dan meningkatkan ketepatan berbanding kaedah sedia ada. Keputusan eksperimen mengesahkan keberkesanan pendekatan ini dalam mengendalikan data tidak seimbang berbilang kelas.

Dalam kajian yang dilakukan oleh Qaddoura et al. (2022) pula, penulis menyiasat pengesanan penipuan kad kredit, menekankan kepentingannya untuk institusi kewangan dalam mencegah caj yang salah kepada pelanggan. Untuk menangani cabaran set data tidak seimbang, kajian ini membandingkan lima teknik pensampelan berlebihan iaitu SMOTE, Pendekatan Pensampelan Sintetik Adaptif (ADASYN), Sempadan 1 SMOTE (borderline1 SMOTE), Sempadan 2 SMOTE (borderline2 SMOTE) dan SVM-SMOTE. Teknik ini bertujuan untuk meningkatkan prestasi model pembelajaran mesin dalam mengesan penipuan. Penilaian itu merangkumi lima algoritma pembelajaran mesin yang berbeza, termasuk Regresi Logistik (LR), Hutan Rawak (RF), Jiran Terdekat K (KNN), *Naives Bayes*, gabungan teknik SVM dan RF. Keputusan menunjukkan bahawa teknik pensampelan berlebihan secara amnya meningkatkan prestasi model, tetapi pilihan teknik terbaik bergantung pada algoritma pembelajaran mesin khusus yang digunakan. Sebagai contoh, DT tidak mendapat manfaat daripada pensampelan berlebihan, manakala KNN menunjukkan peningkatan yang ketara apabila menggunakan pensampelan berlebihan SMOTE. Dalam perbandingan algoritma ini, SVM masih berada di antara kedudukan yang baik dan meyakinkan.

Kajian-kajian lepas membuktikan bahawa algoritma SVM berjaya menghasilkan keputusan prestasi yang tidak lebih dan tidak kurang hebat dari model-model yang lain.

2.10 KESIMPULAN

Kesimpulannya, kajian literatur ini membincangkan kaedah teknik pensampelan menangani taburan data yang tidak sekata dalam set data NSL-KDD. Bab ini membincangkan penggunaan tiga teknik, SMOTE, RUS dan SMOTE-RUS untuk menangani isu ini. Walau bagaimanapun, untuk menjayakan eksperimen ini, kajian ini perlu melaraskan tetapan dengan teliti dan memastikan pengujian model juga dijalankan dengan teliti. Kaedah eksperimen akan dibincangkan dalam bab seterusnya.

BAB III

KAEDAH KAJIAN

3.1 PENGENALAN

Kajian ini dijalankan berdasarkan objektif yang telah dinyatakan dalam Bab 1. Kajian ini bertujuan untuk membandingkan keberkesanan antara Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE), Teknik Pensampelan Rawak Kurang (RUS) dan teknik hibrid Teknik SMOTE-RUS dengan algoritma Mesin Sokongan Vektor (SVM) keatas set data tidak seimbang NSL-KDD (*Network Security Laboratory-Knowledge Discovery in Databases*) dengan membina model pengelasan bagi menguji ketepatan dalam mengelaskan data serangan dengan data normal. Oleh kerana set data NSL-KDD kekurangan contoh kelas minoriti, teknik SMOTE, RUS dan SMOTE-RUS dicadangkan untuk menangani isu ini serta membuat perbandingan antara teknik mana yang lebih berkesan dan membangunkan model ramalan bersama algoritma SVM. Perjalanan eksperimen ini ditunjukkan dalam Rajah 3.1, carta aliran kaedah metodologi kajian ini.

Rajah 3.1 Carta Alir Metodologi Kajian

Secara rumusnya, berdasarkan Rajah 3.1, aliran eksperimen ini akan berpandukan carta alir metodologi ini. Eksperimen ini akan bermula dengan pengumpulan data NSL-KDD yang diperolehi di laman sesawang *Kaggle*. Seterusnya, pra-pemprosesan data

akan dilaksanakan sebelum pelaksanaan pensampelan berlaku. Setelah pra-pemprosesan data dilakukan, pelaksanaan pensampelan semula SMOTE, RUS dan SMOTE-RUS dilaksanakan dan dibangunkan bersama model SVM. Pembangunan ini akan menggunakan set latihan KDDTrain+. Apabila keputusan metrik diperolehi, penalaan SVM parameter akan dilakukan untuk mendapatkan nilai parameter yang boleh meningkatkan peratusan metrik penilaian. Selepas itu, pengujian model sebenar akan diuji keatas set KDDTest+ dimana set data ini mempunyai contoh data serangan yang tidak wujud dalam set data KDDTrain+ untuk menguji prestasi sebenar model yang dibangunkan. Akhir sekali, analisa keputusan metrik dan perbandingan dilakukan dan dibentangkan dalam Bab 5.

3.2 PENGUMPULAN DATA NSL-KDD

Data yang digunakan dalam kajian ini adalah NSL-KDD yang diperkenalkan oleh Tavallaee et al. (2009) yang diperoleh dari laman web Kaggle. Set data NSL-KDD merupakan penambahbaikan daripada kekurangan yang ada dalam set data sebelumnya iaitu DARPA (*Defense Advanced Research Projects Agency*) dan KDD99 *(Knowledge Discovery in Databases 1999)*. Set data DARPA merupakan penanda aras untuk pembangunan set data seperti KDD99, NSL-KDD dan CSE-CIC-IDS2018. Set data DARPA terbuka kepada umum untuk diperoleh bagi kajian dan penambahbaikan. Set data NSL-KDD adalah antara set data yang sering digunakan untuk membina model ramalan bagi mengesan serangan rangkaian. Kelebihan pada set data NSL-KDD adalah kebolehan untuk menganalisis set data penuh NSL-KDD, terbahagi kepada set latihan dan ujian yang akan menghasilkan keputusan yang konsisten walaupun diuji berapa banyak kali, tiada rekod berulang dan boleh memberi kadar pengesanan yang baik untuk algoritma pembelajaran (Bala 2019).

3.3 PRA PEMPROSESAN DATA

Pra pemprosesan data adalah langkah penting dalam menangani ketidakseimbangan kelas dalam model pembelajaran mesin. Ketidakseimbangan kelas berlaku apabila satu jumlah contoh kelas mengatasi kelas yang lain, yang boleh membawa kepada model

bias yang akan memberi prestasi rendah pada kelas minoriti. Dalam kajian ini, set data NSL-KDD akan melalui pra-pemprosesan data sebelum data tersebut sebelum melaksanakan teknik pensampelan semula.

3.3.1 Pengkategorian Data

Data asal NSL-KDD mempunyai pelbagai jenis serangan yang berasingan seperti *satan* dan *portsweep*. Sebelum melaksanakan tugas pra-pemprosesan yang lain, jenis serangan yang berasingan ini akan dikategorikan dalam kelas serangan DOS, Penerokaan (*Probe)*, R2L atau U2R degan merujuk kepada kertas kajian Guha et al. (2016) untuk pengkategorian jenis-jenis serangan.

3.3.2 Pengekodan Nilai Nominal

Set data NSL-KDD mempunyai 6 kategori nominal iaitu protocol_type, service, land, logged in, is_host_login dan is_guest_login. Keenam-enam kategori nominal ini akan dikodkan ke nilai numerikal bagi memudahkan algoritma SVM membacanya. Contoh teknik adalah pengekodan *One-hot* atau pengekodan secara label.

3.3.3 Penskalaan Data

Nilai rekod dalam set data mempunya nilai yang besar dan jurang yang ketara. Oleh itu, perbezaan jurang nilai ini akan ditukar supaya nilai tersebut hanya berada dalam julat nilai yang ditetapkan. Proses penskalaan ini akan menggunakan teknik penskalaan min-maks dengan NSL-KDD.

3.4 PENSAMPELAN DATA DENGAN TEKNIK SMOTE

SMOTE merupakan teknik pensampelan lebih yang direka untuk menangani ketidakseimbangan kelas dengan menghasilkan sampel sintetik untuk kelas minoriti. Ia berfungsi dengan mencipta contoh data baharu antara contoh kelas minoriti sedia ada. Pemilihan parameter nilai K, dalam teknik SMOTE untuk kajian ini adalah berdasarkan Aziz dan Ahmad (2021). Berikut merupakan langkah-langkah melaksanakan teknik pensampelan SMOTE. Langkah-langkah penjanaan teknik pensampelan SMOTE seperti berikut:

1. Pilih secara rawak contoh kelas minoriti (A) daripada set data.
2. Kenal pasti K-jiran terdekat A daripada kelas yang sama.
3. Pilih secara rawak satu jiran (B) daripada K-jiran.
4. Hasilkan contoh sintetik dengan menggabungkan A dan B.
5. Ulang langkah sehingga mendapat bilangan contoh sintetik yang diinginkan.
6. Gabung data sintetik dengan data asal untuk mendapatkan data yang seimbang

3.5 PENSAMPELAN DATA DENGAN TEKNIK RUS

RUS direka untuk mengurangkan data majoriti bagi mencipta keseimbangan kelas dan memilih contoh yang ingin dibuang secara rawak. Berikut merupakan langkah-langkah melaksanakan teknik pensampelan RUS. Langkah-langkah penjanaan teknik algoritma RUS seperti berikut:

1. **Tentukan Nisbah.**
   Tentukan nisbah yang diinginkan antara kelas minoriti dan majoriti selepas pengurangan. Sebagai contoh, jika nisbah 1:1 dipilih, pengurangan bilangan sampel kelas majoriti sehingga sepadan dengan bilangan contoh kelas minoriti.
2. **Pemilihan Rawak.**
   Pilih secara rawak satu set contoh daripada kelas majoriti untuk mencapai nisbah yang diinginkan.
3. **Kemaskini set data.**
   Cipta set data baru dengan contoh minoriti yang dipilih dan contoh majoriti yang dipilih secara rawak. Set data baru ini akan mempunyai taburan kelas yang seimbang.

3.6 PENSAMPELAN DATA DENGAN TEKNIK SMOTE-RUS

SMOTE-RUS adalah gabungan dua teknik, SMOTE dan RUS yang digunakan untuk menangani ketidakseimbangan kelas dalam set data. Gabungan ini bertujuan untuk mengimbangi pengagihan kelas dengan kedua-dua pensampelan lebih untuk kelas

minoriti dan pensampelan kurang untuk kelas majoriti. Teknik SMOTE-RUS adalah berdasarkan kajian oleh Ismail et al. (2023). Berikut adalah langkah-langkah yang terlibat dalam penjanaan set data SMOTE-RUS:

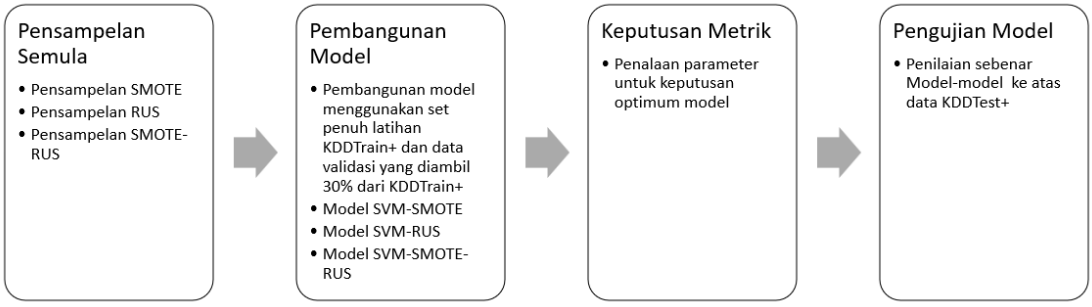
1. Kenal pasti Kelas Minoriti dan Majoriti: Langkah pertama ialah menentukan kelas mana yang minoriti (kurang diwakili) dan yang mana majoriti (lebih diwakili) dalam set data.
2. Melaksanakan SMOTE ke Kelas Minoriti.
3. Melaksanakan RUS ke Kelas Majoriti.
4. Menggabungkan set data kelas minoriti yang telah di-SMOTE dan kelas majoriti yang telah dikurangkan untuk membentuk set data seimbang yang baharu.

3.7 MODEL RAMALAN SUPPORT VECTOR MACHINE (SVM)

Untuk menjalani kajian ini, platform WEKA dan KNIME menjadi pilihan dalam pembangunan model ramalan. Penalaan parameter untuk algoritma SVM adalah berdasarkan kajian oleh Aamir and Ali Zaidi 2021, Kuyoro, Alimi et al. (2022).

3.7.1 Pembangunan Model Ramalan

Keberkesanan sesuatu model ramalan dapat ditentukan dengan hasil keputusan yang diperoleh. Sebelum model ramalan dibangunkan, latihan keatas model perlulah dilaksanakan terlebih dahulu untuk mengajar model ramalan supaya ia dapat memberikan output ramalan yang optimum apabila diuji menggunakan data yang tidak dapat dilihat semasa latihan.



Rajah 3.2 Pembangunan Model Ramalan

Berdasarkan Rajah 3.2, set data NSL-KDD akan diseimbangkan menggunakan teknik SMOTE, RUS dan SMOTE-RUS. Model asas dan model teknik pensampelan semula akan dilatih oleh algoritma SVM menggunakan set data validasi. Set validasi ini mengambil 30% daripada set data penuh KDDTrain+. Selepas latihan model telah berlangsung, berdasarkan keputusan metrik, penalaan akan berlaku kepada parameter SVM sehingga mendapatkan keputusan yang paling optimum. Seterusnya pengujian model ke atas set data ujian dimana data yang tidak pernah dilihat oleh model. Pada akhir eksperimen ini, keberkesanan model ramalan ini akan diukur melalui nilai metrik.

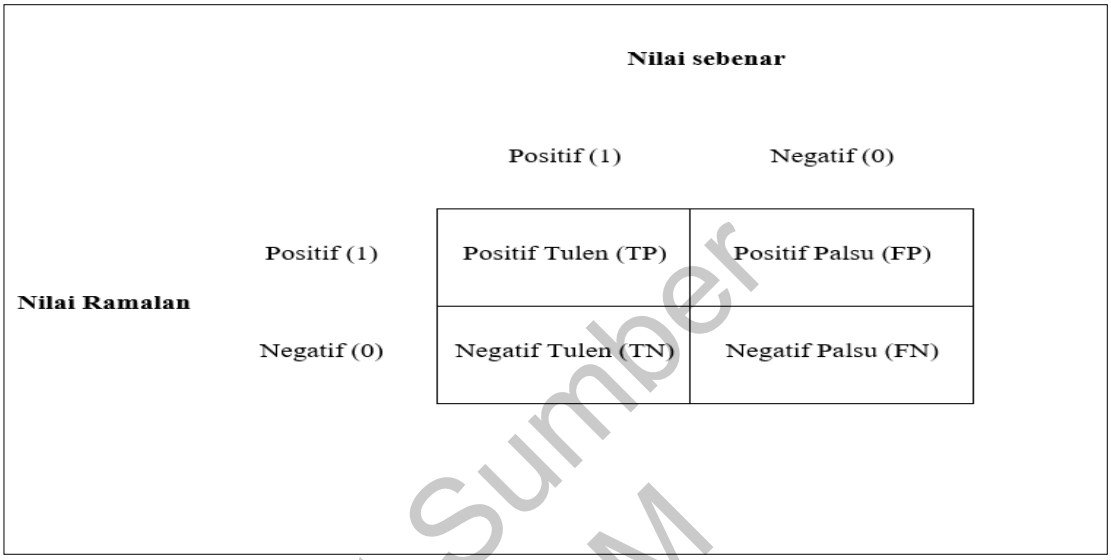
Bagi membangunkan model pengelasan, perisian WEKA dan KNIME akan digunakan dalam pembangunan ini. Model akan dibangunkan menggunakan algoritma pengelasan SVM. Jumlah set data yang akan digunakan adalah empat set data iaitu set data asal, SMOTE, SMOTE-RUS, dan RUS. Model asas dan model pensampelan semula akan dibangunkan. Prestasi model-model yang dibangunkan ini akan diuji dengan penilaian metrik seperti *(accuracy),* kepersisan *(precision),* rekal *(recall),* dan skor F1 *(F1-Score)* dan membandingkan prestasi model-model menggunakan tiga teknik pensampelan semula ini.

3.8 PENILAIAN METRIK

Prestasi model pengelasan yang telah dibangunkan boleh dianalisis berdasarkan metrik penilaian prestasi. Antara metrik penilaian prestasi yang digunakan dalam kajian ini adalah *(accuracy),* kepersisan *(precision),* rekal *(recall)* dan skor F1 *(F1-Score).* Berikut merupakan metrik penilaian prestasi yang digunakan dalam kajian ini:

1. Metrik Kekeliruan

Metrik kekeliruan digunakan sebagai penilaian prestasi model ramalan dalam pengelasan. Ia menunjukkan sekiranya model ramalan memberi pengelasan yang tepat terhadap data. Rajah 3.3 merupakan matriks kekeliruan dimana ketepatan ramalan yang dilakukan oleh model pengelasan akan diukur berdasarkan empat output iaitu nilai positif tulen, positif palsu, negatif tulen dan negatif palsu.



Rajah 3.3 Metrik Kekeliruan

a) Positif Tulen (TP): Nilai sebenar adalah sama dengan nilai ramalan iaitu 1.

b) Negatif Tulen (TN): Nilai sebenar adalah sama dengan nilai ramalan iaitu 0.

c) Positif Palsu (FP): Nilai sebenar adalah 0 manakala nilai ramalan adalah 1.

d) Negatif Palsu (FN): Nilai sebenar adalah 1 manakala nilai ramalan adalah 0.

2. Ketepatan *(Accuracy)*

Ketepatan adalah nilai pengelasan yang tepat untuk positif tulen dan negatif tulen yang diperoleh oleh model pengelasan. Ketepatan yang baik akan diperoleh sekiranya pengelasan menempatkan kelas di tempat yang sepatutnya.

3. Kepersisan *(Precision)*

Kepersisan adalah nilai sebenar positif tulen di antara nilai ramalan positif tulen. Nilai ramalan positif tulen ini merangkumi positif tulen dan positif palsu. Sekiranya tiada nilai positif palsu, model pengelasan dianggap mempunyai 100% kepersisan.

4. Rekal / Sensitiviti *(Recall / Sensitivity)*

Rekal bertujuan menghasilkan kadar ketepatan yang tinggi. Ia mencari nilai sebenar positif tulen diantara jumlah sampel positif tulen dan negatif palsu yang diramal oleh model.

5. Skor-F1 (F1-Score)

Skor-F1 merupakan gabungan nilai kepersisan dan rekal untuk menyeimbang nilai kepersisan dan rekal. Ia saling melengkapi tanpa mengabaikan salah satu nilai. Sebagai contoh, sekiranya fokus adalah kepada nilai kepersisan dimana ia ingin mengurangkan nilai positif palsu, nilai negatif palsu tetap diberi fokus yang sama. Manakala, sekiranya fokus adalah kepada nilai rekal dimana ia ingin mengurangkan nilai negatif palsu, nilai positif palsu juga akan diberi fokus yang sama.

3.9 KESIMPULAN

Bab ini menerangkan kaedah atau metodologi yang digunakan dalam kajian ini. Ia menggariskan teknik dan proses untuk mencapai objektif untuk menyelesaikan permasalahan kajian dalam Bab 1. Pelaksanaan, pengujian dan dapatan kajian akan dijelaskan dalam bab seterusnya.

BAB IV

PELAKSANAAN EKSPERIMEN

4.1 PENGENALAN

Bab ini membincangkan persediaan eksperimen yang direka untuk berfungsi sebagai tanda aras dalam melaksanakan eksperimen ini. Set data yang dipilih merupakan tujuan utama eksperimen dijalankan, iaitu set data tidak seimbang NSL-KDD (*Network Security Laboratory-Knowledge Discovery in Databases*) dan teknik pensampelan semula iaitu Teknik Pensampelan Sintetik Minoriti Lebih (SMOTE), Pensampelan Rawak Kurang (RUS), dan Teknik Pensampelan Sintetik Minoriti Lebih - Pensampelan Rawak Kurang (SMOTE-RUS) digunakan untuk menyeimbang set data ini. Seterusnya langkah pra-pemprosesan untuk membantu dalam membuat analisis. Setelah itu, teknik pensampelan semula untuk mengimbangi data yang tidak seimbang. Model asas digunakan dalam eksperimen ini sebagai titik rujukan penting. Model asas ini tidak ada sebarang campur tangan pensampelan semula kerana ia bertujuan untuk menetapkan titik permulaan analisis perbandingan ini. Algoritma pengelasan yang dipilih, tanpa pensampelan berlebihan atau pensampelan terkurang, berfungsi sebagai penanda aras yang terhadapnya kesan teknik pensampelan semula akan diukur. Metrik seperti ketepatan, kepersisan, rekal dan skor-F1 menjadi penanda untuk memahami tentang kesan data seimbang atau tidak seimbang pada prestasi model pengelasan.

4.2 PERSEDIAAN EKSPERIMEN

Berikut merupakan persediaan eksperimen:

1. Set data KDDTrain+, KDDTest+ dan set data validasi NSL-KDD (KDDTrain+ 30%).

2. Pelaksanaan pra-pemprosesan data iaitu pengkategorian kelas serangan, pengekodan *One-Hot*, penskalaan data dan pembahagian set data.

3. Penggunaan teknik pensampelan semula SMOTE, RUS dan SMOTE-RUS.

4. Pembangunan model asas sebagai penanda aras untuk model pensampelan data.

5. Matriks penilaian yang dipilih iaitu Ketepatan, Kepersisan, Rekal, dan Skor F1.

6. Menggunakan perisian KNIME 5.1.1, WEKA 3.8.6 dan Microsoft Excel 2021.

7. Perkakasan komputer iaitu, Windows 11 Home Edition 64-bit, 12th Gen Intel(R) Core(TM) i7-1255U 1.70 GHz dan RAM 8.00 GB.

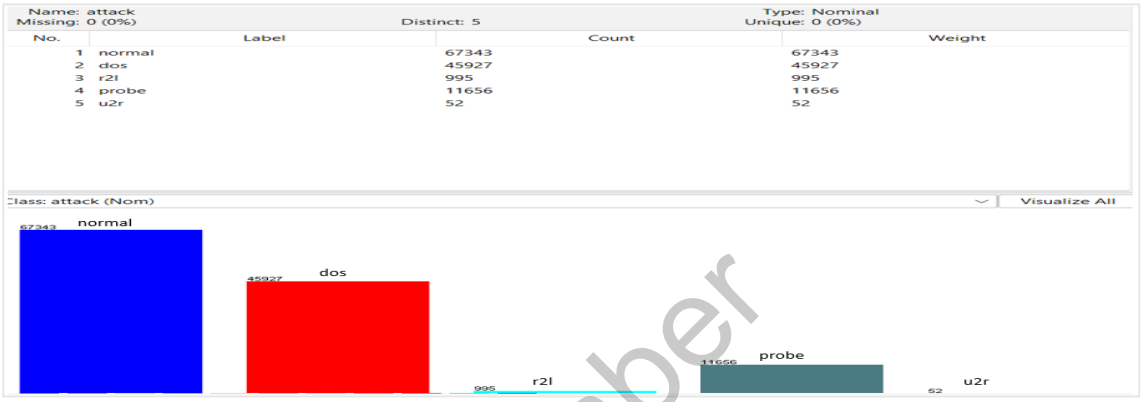
4.3 PRA-PEMPROSESAN DATA

Dalam proses penyediaan data kajian ini, perisian KNIME (Abualkibash 2019; Kanagavalli et al. 2023) dan Microsoft Excel 2021 telah digunakan untuk menyediakan data sebelum diberi kepada model pengelasan SVM untuk pembangunan. Perisian KNIME digunakan untuk pra-pemprosesan, teknik pensampelan, pembangunan model dan pengujian model, manakala untuk perisian Excel digunakan untuk mengkategorikan pelbagai jenis serangan ke dalam kelas serangan yang dilakukan secara manual.

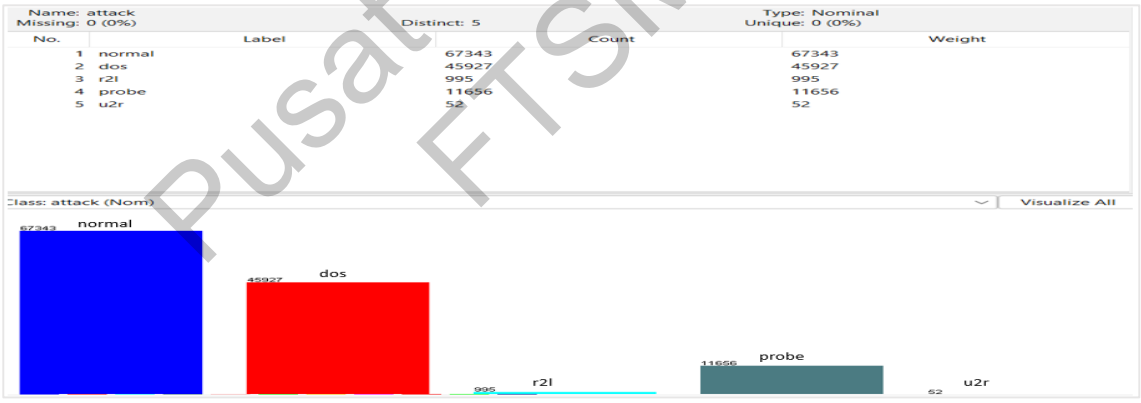
Kajian ini melalui empat peringkat pra-pemprosesan data iaitu mengkategorikan jenis-jenis serangan ke dalam satu kelas serangan, pengekodan nilai nominal, penskalaan dan pembahagian set data kepada set data latihan dan set data validasi. Setelah pra-pemprosesan dilaksanakan keatas set data NSL-KDD yang asal, set data yang tadi akan digunakan bagi menjalankan eksperimen untuk teknik pensampelan SMOTE, RUS dan SMOTE-RUS.

4.3.1 Pengkategorian Pelbagai Serangan Dalam Kelas Serangans

Set data asal NSL-KDD mempunyai pelbagai serangan antaranya adalah seperti Satan dan Portsweep. Oleh itu, jenis-jenis serangan yang pelbagai ini akan dikategorikan ke dalam kelas serangan yang bersesuaian sama ada DOS, Probe, U2R atau R2L yang berpandukan daripada kertas kajian oleh Guha et al. (2016).



Rajah 4.1 Rajah 4.1 menunjukkan kelas Normal dan kelas serangan iaitu DOS, *Probe*, U2R dan R2L.



Rajah 4.1 Pengkategorian Serangan

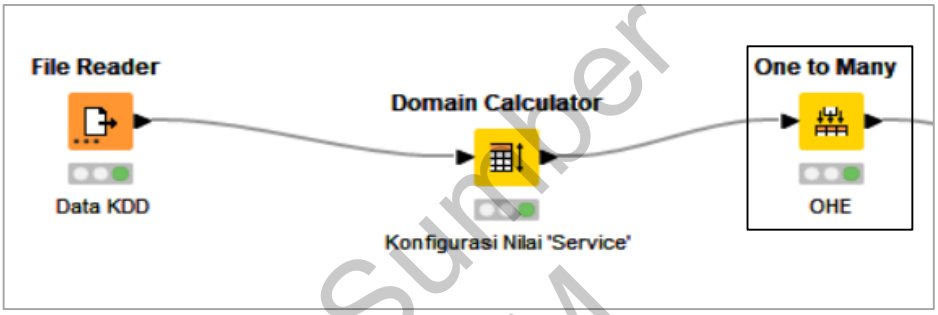
Setelah pengkategorian telah dilaksanakan, wujud kelas empat jenis kelas serangan iaitu kelas serangan DOS, *Probe*, U2R atau R2L.

4.3.2 Pengekodan *One-Hot*

Dalam set data NSL-KDD, atribut *protocol_type*, *service* dan *flag* merupakan nilai nominal. Nilai nominal tersebut perlu ditukar ke nilai numerikal kerana SVM memerlukan input berbentuk numerikal, menjadikan pengekodan *One-Hot* penting

untuk mengendalikan nilai nominal. Proses ini menterjemahkan kategori ke dalam vektor binari, memastikan SVM tidak salah menginterpretasikan adanya urutan atau magnitud di antara kategori. Ini membolehkan SVM memproses dan mengklasifikasikan kategori ini dengan tepat bagi mengelakkan kekeliruan dan membantu dalam pemisahan dan pengelasan titik data yang efektif dalam ruang ciri.

Rajah 4.2 memperlihatkan atribut *protocol_type* mewakili nilai nominal. Pertukaran nilai nominal tersebut menggunakan nod *One to Many* untuk pengekodan *One-Hot*. Pengekodan ini dilaksanakan keatas KDDTrain+ dan KDDTest+.



protocol_type	tcp	udp	icmp
tcp	1	0	0
udp	0	1	0
tcp	1	0	0
tcp	1	0	0
tcp	1	0	0

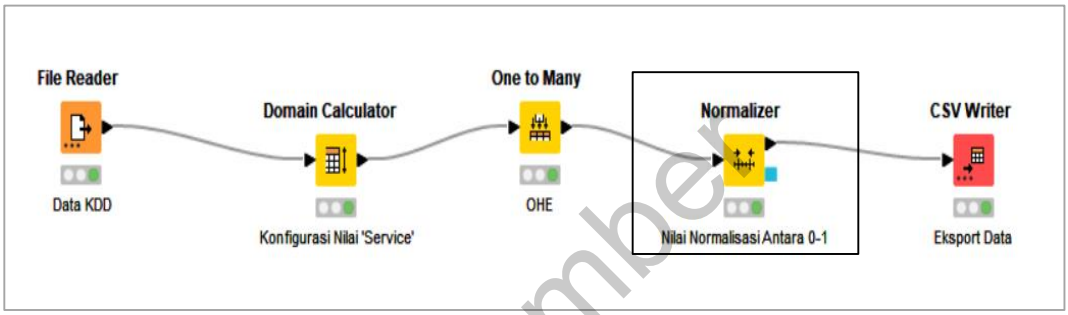
Rajah 4.2 Nilai Nominal ke Numerikal *protocol_type*

Selepas proses pengekodan *One-Hot* berlaku, atribut *protocol_type*, ia bertukar menjadi wakil nilai numerikal dengan menggunakan nilai binari iaitu 0 atau 1. Berdasarkan Rajah 4.2, tcp, udp dan icmp diwakilkan menggunakan nilai 0 dan 1. Contohnya, jika atribut tersebut merupakan tcp, ia akan mewakilkannya dengan nilai 1 dan selain itu adalah nilai 0 dan begitu juga dengan udp dan icmp.

4.3.3 Penskalaan Data

Penskalaan data membantu dalam penukaran nilai ke skala standard. Kaedah yang digunakan dalam kajian ini ialah penskalaan Min-Max, menggunakan nod *Normalizer*

di mana nilai ditukar kepada julat antara 0 sehingga 1. Oleh kerana nilai asal data dalam NSL-KDD besar, penskalaan ini dilaksanakan untuk memberi nilai yang lebih seimbang dan memudahkan algoritma SVM untuk memproses data dengan lebih cepat. Berdasarkan Rajah 4.3, set data penuh KDDTrain+ dan KDDTest+ akan dimasukkan ke dalam nod File Reader untuk melalui proses penskalaan data menggunakan nod Normalizer. Selepas proses penskalaan Min-Max dilaksanakan, nilai atribut *count* dan *srv_count* berada dalam julat antara 0 hingga 1.



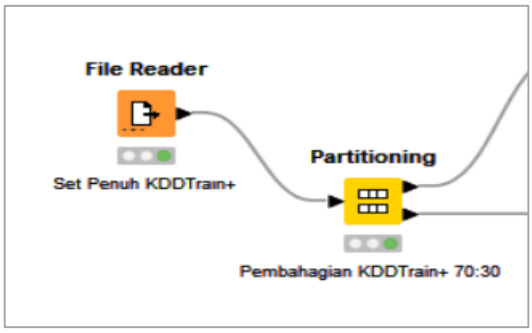
count	srv_count	count	srv_count
2	2	0.004	0.004
13	1	0.025	0.002
123	6	0.241	0.012
5	5	0.01	0.01
30	32	0.059	0.063

Rajah 4.3 Penskalaan Min-Max

Dengan ini, fasa pra-pemprosesan data telah dilaksanakan keatas set data KDDTrain+ dan KDDTest+. Langkah seterusnya adalah melaksanakan pensampelan data menggunakan teknik SMOTE, RUS dan SMOTE-RUS. Langkah ini akan dibincangkan dalam seksyen berikutnya.

4.3.4 Pembahagian Set Data NSL-KDD (KDDTrain+)

Set penuh data KDDTrain+ akan dibahagikan kepada 70% sebagai set latihan untuk model pengelas dan baki 30% digunakan sebagai set validasi. Kemudian, pensampelan data akan dilaksanakan untuk menyeimbangkan data sebelum membina model pengelasan bersama algoritma SVM.



Rajah 4.4 Pembahagian Set Data

Berdasarkan Rajah 4.4 di atas, pembahagian set penuh KDDTrain+ untuk latihan dan validasi dilakukan melalui nod *Partitioning*. Jadual 4.1 menunjukkan jumlah data selepas pembahagian untuk set latihan dan set validasi daripada set penuh NSL-KDD.

Jadual 4.1 Pembahagian Set Data

Data	**Asal**	**SMOTE**	**RUS**	**SMOTE-RUS**
Set Latihan (KDDTrain+ 70%)	88,181	47,141	17,636	41,913
Set Validasi (KDDTrain+ 30%)	37,792	37,792	37,792	37,792

Pelaksanaan teknik pensampelan semula hanya berlaku pada set KDDTrain+ dan bukan set KDDTest+ supaya dapat mencegah kebocoran data, memastikan set validasi dan set ujian kekal tidak bias semasa proses latihan (Bahrami 2023). Dengan mengekalkan integriti set validasi dan ujian, model dinilai berdasarkan data sebenar yang tidak pernah dilihat sebelumnya, memberikan penilaian yang boleh dipercayai terhadap prestasi generalisasinya. Selain itu, membolehkan set validasi dan ujian menjalankan penilaian keupayaan model yang sebenar (Kumar et al. 2023).

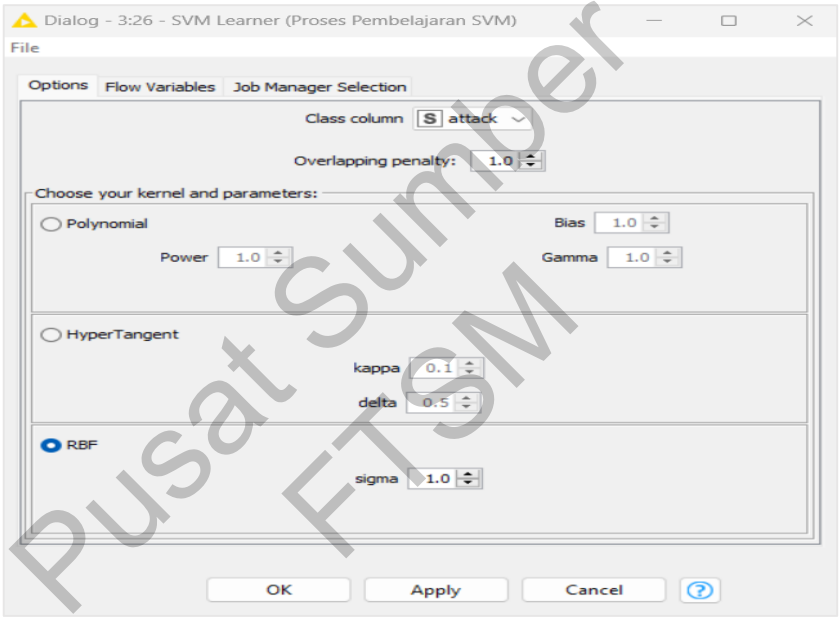
Menggunakan teknik pensampelan pada set validasi dan ujian boleh menyebabkan pemadanan berlebihan (*overfitting)* pada ciri-ciri tertentu yang diperkenalkan oleh teknik pensampelan dan mengubah kemampuan generalisasi sebenar model. Ini bertujuan supaya dapat mengekalkan kebolehpercayaan penilaian model pada data yang tidak pernah dilihat sebelumnya (Montesinos López et al. 2022).

4.4 PEMBANGUNAN MODEL KE ATAS SET DATA NSL-KDD

Dalam seksyen ini, tetapan parameter SVM, pembangunan model dan pelaksanaan teknik pensampelan SMOTE, RUS dan SMOTE-RUS akan dibincangkan.

4.4.1 Parameter Support Vector Machine (SVM)

Tetapan parameter SVM akan dilaksanakan sebelum membina model ramalan. Dalam perisian KNIME, terdapat nod *SVM Learner* dan nod *SVM Predictor* yang digunakan untuk membina model ramalan dan meramalkan output.



Rajah 4.5 Tetapan SVM

Berdasarkan Rajah 4.5, dua parameter yang perlu ditetapkan adalah fungsi Penalti bertinding (*Overlapping penalty)* dan jenis kernel dengan nilainya. Nilai parameter yang dipilih untuk penalti bertindih dan kernel RBF adalah berdasarkan kajian oleh Aamir et al. (2021) dan (Kuyoro et al. 2022).

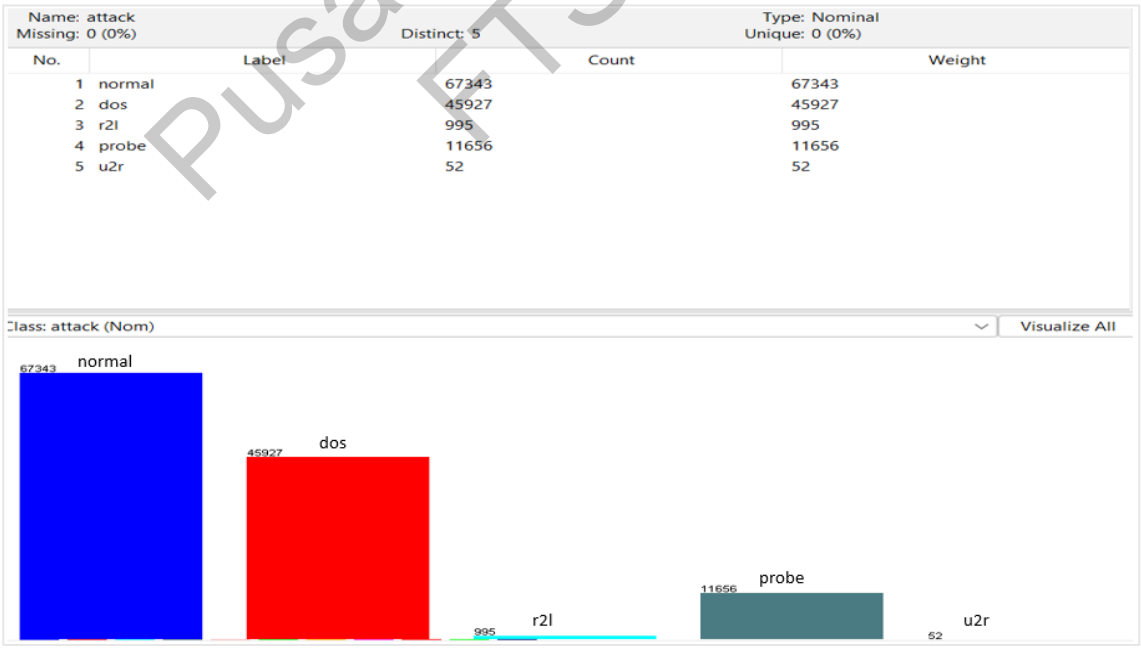
Fungsi penalti bertindih ditetapkan ke nilai 1.0. Penalti bertindih berfungsi untuk mengendalikan data yang tidak dapat dipisahkan dengan sempurna ke dalam kelas yang berbeza. Ia membolehkan SVM bertolak ansur dengan beberapa pengelasan yang salah yang membolehkan model mencari keseimbangan antara mengelaskan titik

data majoriti dengan tepat dalam masa yang sama mengekalkan sempadan keputusan yang mudah dan boleh digeneralisasikan (Abe 2005).

Kernel yang dipilih dalam eksperimen ini adalah kernel RBF dan nilai sigma menggunakan 1.0. Kernel RBF (*Radial Basis Function*) mengubah data yang kompleks dan bercampur-campur menjadi format yang lebih mudah untuk diasingkan. Fleksibilitinya datang daripada parameter boleh laras, membolehkannya mengendalikan kedua-dua corak data ringkas dan kompleks dengan cekap. Ini menjadikan kernel RBF sangat serba boleh dan sesuai untuk pelbagai tugas daripada mengesan corak, peramalan dan lain-lain (Abe 2005).

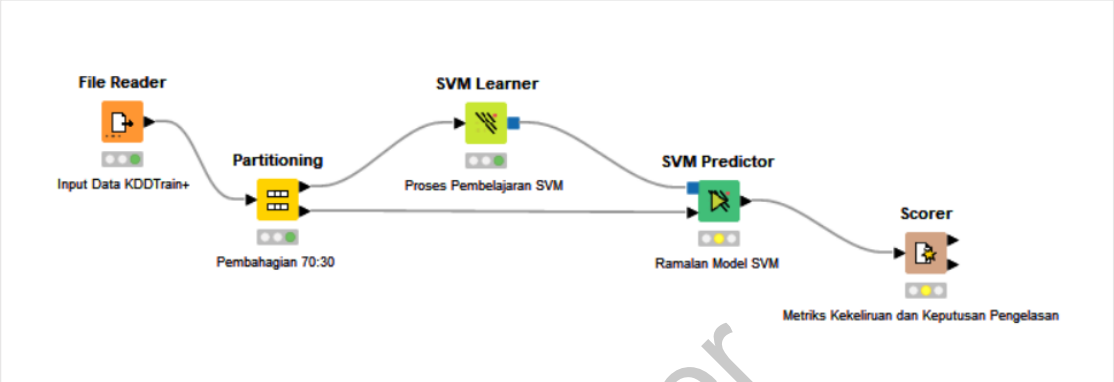
4.4.2 Pembangunan Model Asas

Perlaksanaan model asas ini adalah sebagai penanda aras awal, yang mewakili prestasi pengelasan yang dilatih pada set data NSL KDD asal yang tidak seimbang. Ini berfungsi sebagai titik rujukan untuk membandingkan model yang dilatih dengan teknik pensampelan semula yang berbeza, seperti SMOTE, RUS dan SMOTE-RUS.



Rajah 4.6 Data Model Asas

Rajah 4.6 menunjukkan data asal yang tidak menjalani pensampelan semula. Ia juga menunjukkan bahawa ketidakseimbangan data yang sangat ketara lebih-lebih lagi dalam kelas minoriti iaitu R2L, *Probe* dan U2R.



Rajah 4.7 Pembangunan Model Asas

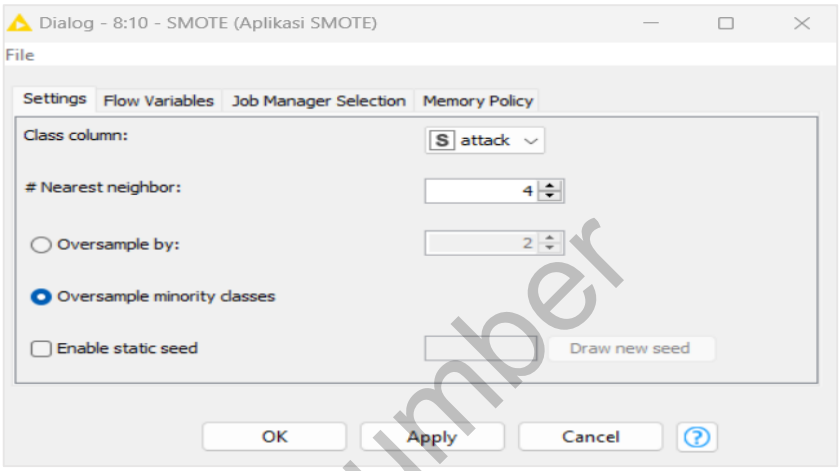
Rajah 4.7 menunjukkan proses pembangunan model asas menggunakan data asal NSL-KDD. Proses ini bermula dengan kemasukan data penuh KDDTrain+ ke dalam *File Reader*. Selepas data dibaca oleh nod tersebut, data ini akan dibahagikan kepada 70:30 menggunakan nod *Partitioning*, dimana 70% adalah untuk set latihan, manakala 30% adalah untuk set validasi. Set latihan 70% ini akan diberi kepada nod *SVM Learner* dan baki 30% data validasi digunakan oleh nod *SVM Predictor* untuk membuat ramalan model. Nod *Scorer* akan menghasilkan output hasil pembangunan model tadi.

4.4.3 Pembangunan Model SMOTE

Dalam perisian KNIME, nod SMOTE digunakan untuk menjana sampel SMOTE bagi mendapatkan data seimbang. Berdasarkan Rajah 4.8, penalaan parameter SMOTE tertumpu pada *Nearest neighbour (*K*)* dan *Oversample minority classes*. Nilai K dipilih berdasarkan kertas kajian oleh Aziz dan Ahmad (2021) yang menggunakan nilai 4.

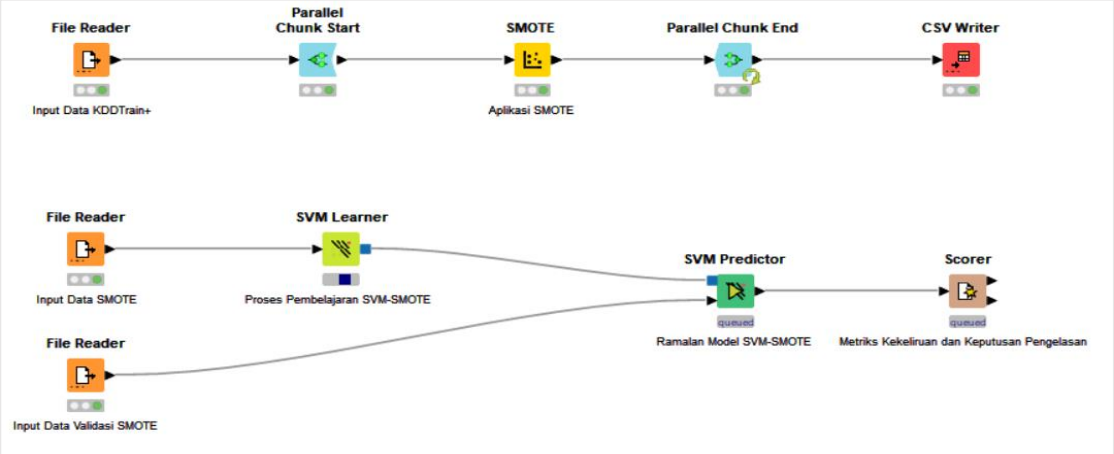
Dalam KNIME, SMOTE menjana sampel baharu dengan teknik interpolasi. Nilai K ditetapkan kepada 4 supaya struktur asal kelas minoriti boleh dikekalkan dengan menggunakan nilai K yang lebih kecil. Dengan pemilihan nilai K = 4, SMOTE

mengenal pasti sampel kelas minoriti dan, untuk setiap satu data, mencari 4 jiran terdekat dalam kelas yang sama untuk menginterpolasi dan mencipta sampel sintetik baharu. Interpolasi ini melibatkan penjanaan sampel baharu dengan mengubah sedikit nilai ciri antara sampel asal dan jirannya, dan prosesnya berulang sehingga kelas minoriti mencapai kiraan yang dikehendaki.



Rajah 4.8 Tetapan SMOTE

Parameter *Oversample minority classes* dipilih supaya SMOTE akan meningkatkan kiraan kelas DOS, *Probe*, R2L, U2R agar sepadan dengan kelas Normal, mengimbangi taburan kelas dalam set data dengan berkesan dengan menjana 47,141 sampel sintetik untuk setiap kelas DOS, Probe, R2L, U2R.



Rajah 4.9 Pembangunan Model SMOTE

Berdasarkan Rajah 4.9, pembangunan model SMOTE bermula dengan memasukkan data KDDTrain+ ke dalam *File Reader*. Data asal dibahagikan kepada 70% set latihan dan 30% data validasi. 70% data ini akan dilaksanakan teknik pensampelan SMOTE menggunakan nod *SMOTE* kemudian. Setelah proses SMOTE dilaksanakan, set data SMOTE yang telah dijana akan dihantar ke nod *SVM Learner* sebagai data pembelajaran. Data validasi 30% dihantar ke nod *SVM Predictor* sebagai data ramalan. Setelah model ramalan SVM-SMOTE dibina, hasil keputusan akan direkodkan dalam nod *Scorer*. Rajah 4.10 menunjukkan jumlah data SMOTE yang telah dijana.

Rajah 4.10 Data Model SMOTE

Selepas SMOTE dilaksanakan, setiap kelas mempunyai rekod sebanyak 47,141 seperti yang ditunjukkan dalam Jadual 4.2. Hasil penjanaan 47,141 sampel sintetik untuk setiap kelas minoriti menggunakan SMOTE, sememangnya hasil yang dijangkakan apabila objektif kajian ini adalah untuk mengimbangi semua kelas agar sepadan dengan kiraan kelas majoriti dalam set data.

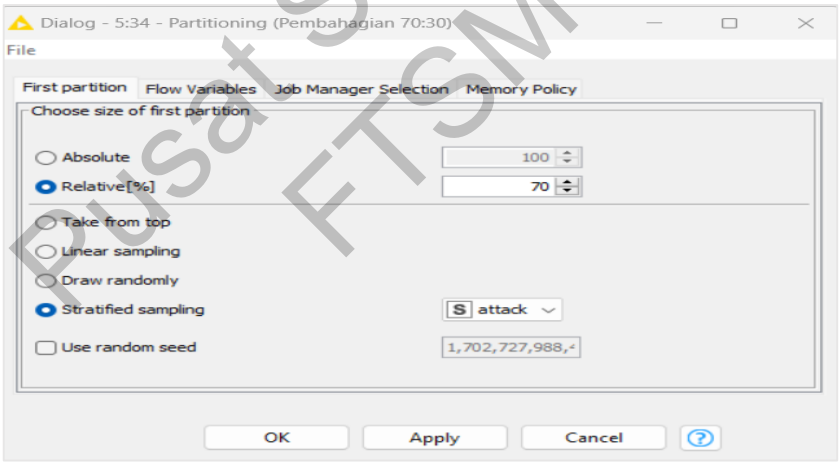
Jadual 4.2 Data SMOTE

Data	**Normal**	**DOS**	**Probe**	**R2L**	**U2R**
Asal	47,141	32,149	8,159	696	36
SMOTE	47,141	47,141	47,141	47,141	47,141

Daripada jumlah set data asal 125,974, menjadi 235,705 hasil SMOTE.

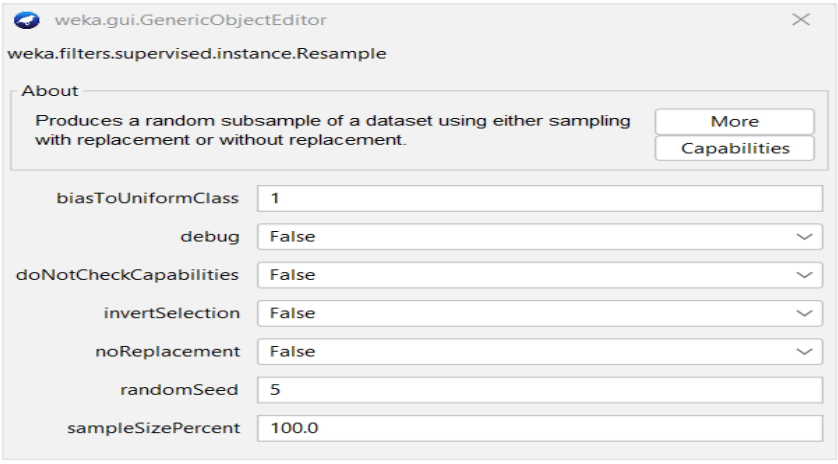
4.4.4 Pembangunan Model RUS

Untuk melaksanakan teknik pensampelan RUS, perisian WEKA digunakan bagi menghasilkan data daripada teknik RUS. Data asal KDDTrain+ dibahagikan terlebih dahulu kepada 70% set latihan dan 30% set validasi. Pecahan ini dibahagikan dalam perisian KNIME menggunakan nod *Partitioning* seperti dalam Rajah 4.11. Set latihan 70% yang dipecahkan ini merupakan set latihan yang akan dilaksanakan teknik RUS menggunakan perisian WEKA.



Rajah 4.11 Pembahagian Data RUS

Data asal dipecahkan kepada 70% dan fungsi *Stratified sampling* digunakan. Dalam tetapan RUS berdasarkan Rajah 4.12, satu parameter yang diubah iaitu biasToUniformClass ditukar ke nilai 1.0 untuk memberikan nilai yang sama rata untuk semua kelas serangan termasuk kelas normal.



Rajah 4.12 Tetapan RUS

Parameter nilai *randomSeed* ditetapkan ke nilai standard iaitu 5. Parameter *sampleSizePercent* adalah saiz peratusan subsampel daripada set data asal. Nilai 100% dipilih supaya tiada pengurangan subsampel yang ketara dan jumlah rekod asal kekal.

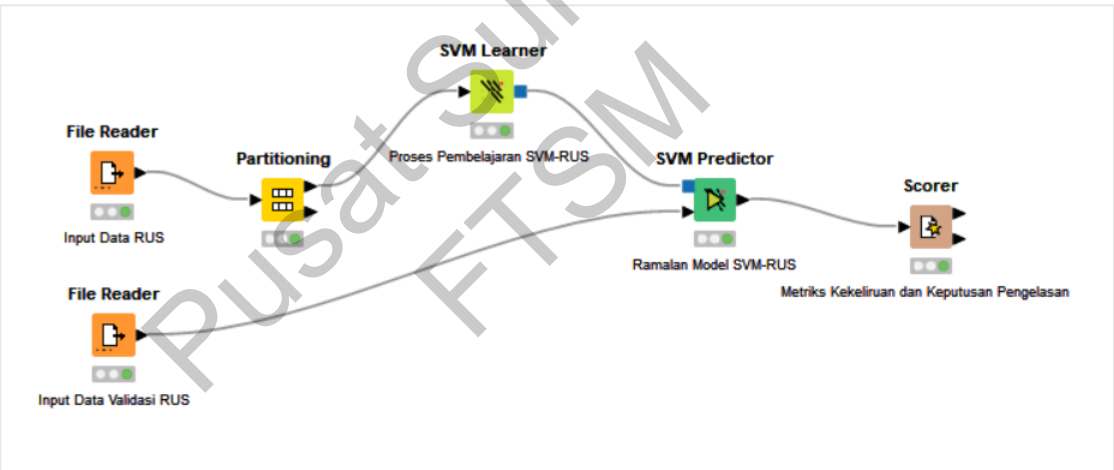
Rajah 4.13 Data Model RUS

Rajah 4.13 menunjukkan hasil tetapan parameter dan penjanaan data RUS. Jumlah bilangan data hasil pengurangan RUS adalah 88,180.

Jadual 4.3 Data RUS

Data	**Normal**	**DOS**	**Probe**	**R2L**	**U2R**
Asal	47,141	32,149	8,159	696	36
RUS	17,636	17,636	17,636	17,636	17,636

Berdasarkan Jadual 4.3, bilangan contoh untuk semua kelas berkurangan dan diseratakan menjadikan bilangan contoh setiap kelas adalah 17,636. Hasil daripada menggunakan RUS pada set data, yang membawa kepada bilangan sampel yang sama merentas semua kategori (17,636), dijangka dan sejajar dengan objektif utama RUS iaitu untuk mengurangkan ketidakseimbangan kelas dengan mengurangkan saiz secara rawak kelas majoriti untuk memadankan kelas minoriti, dengan itu mencapai pengagihan yang lebih seimbang. Daripada jumlah set data asal 125,974, menjadi 88,180 hasil RUS.

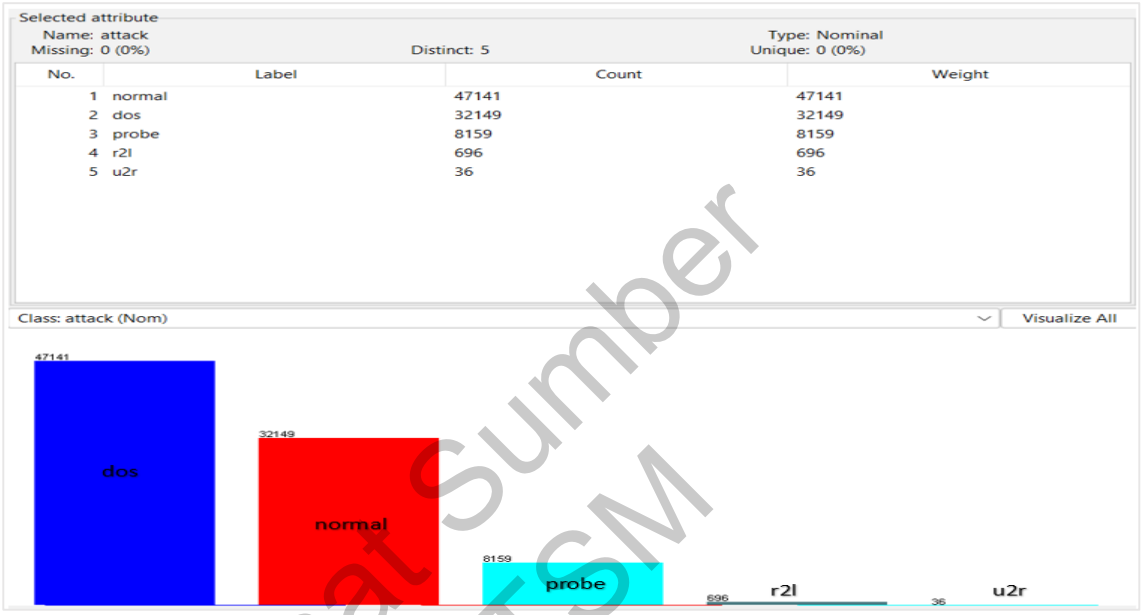


Rajah 4.14 Pembangunan Model RUS

Berdasarkan Rajah 4.14, pembangunan model RUS bermula dengan memasukkan data RUS dan data validasi ke dalam File Reader secara berasingan. 70% data RUS dari *File Reader* ini kemudian akan dihantar ke nod *SVM Learner* sebagai data pembelajaran. Data validasi 30% daripada nod *File Reader* dihantar ke nod *SVM Predictor* sebagai data ramalan. Setelah model ramalan SVM-RUS dibina, hasil keputusan akan direkodkan dalam nod *Scorer*.

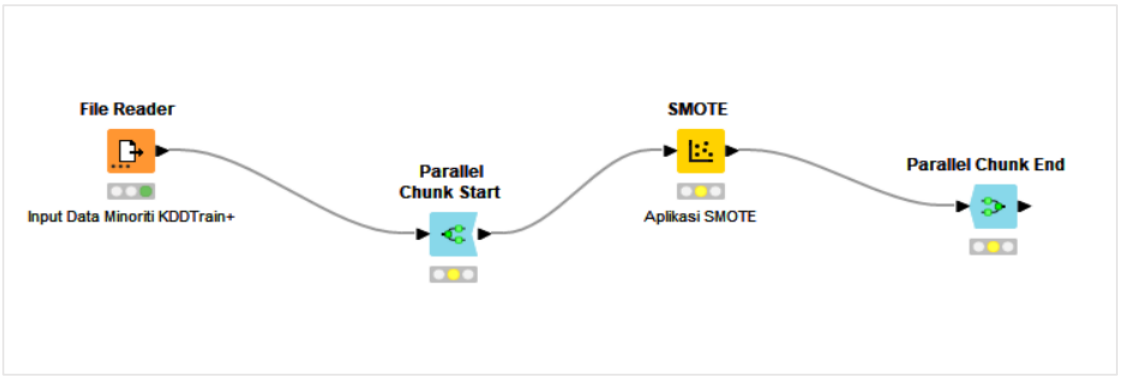
4.4.5 Pembangunan Model SMOTE-RUS

Untuk teknik SMOTE-RUS. Pecahan kelas majoriti dan kelas minoriti dilaksanakan terlebih dahulu sebelum melaksanakan pensampelan data SMOTE pada kelas minoriti dan pensampelan RUS pada kelas majoriti. Rajah 4.15 menunjukkan jumlah rekod set latihan asal.



Rajah 4.15 Rekod Set Latihan Asal

Berdasarkan Rajah 4.16, pelaksanaan pensampelan SMOTE dilaksanakan ke atas kelas minoriti iaitu *Probe*, R2L dan U2R.



Rajah 4.16 Pensampelan SMOTE Untuk Kelas Minoriti

Setelah data untuk kelas minoriti dimasukkan ke dalam nod *FileReader* dan seterusnya dihantar ke nod *SMOTE* untuk proses pensampelan SMOTE.

Rajah 4.17 Jumlah Kelas Minoriti Selepas SMOTE

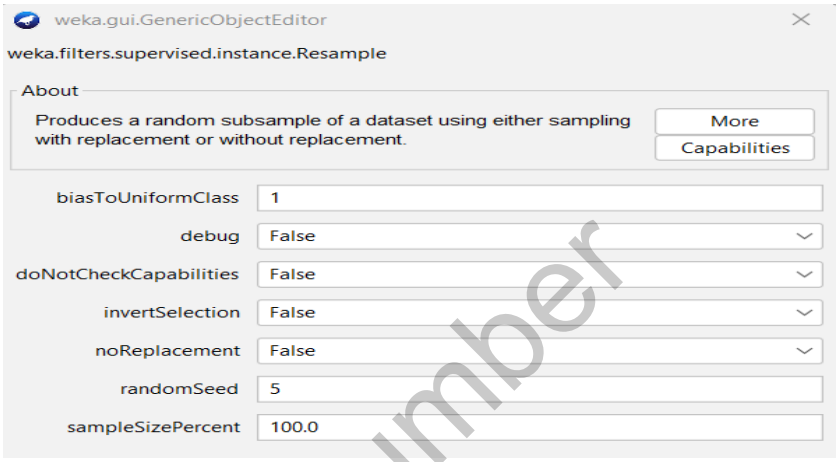
Selepas pelaksanaan SMOTE, jumlah rekod untuk kelas minoriti adalah 24,471 berbanding jumlah rekod kelas minoriti sebelum SMOTE iaitu 8,891. Jadual 4.4 menunjukkan rekod asal dan selepas SMOTE untuk kelas minoriti.

Jadual 4.4 Rekod Data Minoriti Selepas SMOTE

	Probe	**R2L**	**U2R**
Asal	8,159	696	36
SMOTE	8,157	8,157	8,157

Penjanaan dan penyeimbangan jumlah SMOTE untuk kelas minoriti dilakukan dalam perisian KNIME dengan tetapan parameter yang sama seperti dalam seksyen 4.4.3. Seterusnya, teknik pensampelan RUS dilaksanakan pada kelas majoriti iaitu kelas Normal dan DOS. Teknik pensampelan RUS dilaksanakan dalam perisian WEKA.

Untuk pensampelan RUS untuk kelas majoriti, pengurangan sampel adalah 22% daripada data asal rekod majoriti. Ini bertujuan untuk membawa jumlah rekod majoriti samar-samar dengan jumlah rekod minoriti. Setelah jumlah rekod telah dikurangkan, berdasarkan Rajah 4.18, parameter ditetapkan. Parameter *biastoUniformClass* ditetapkan kepada 1 agar jumlah rekod untuk kelas Normal dan DOS adalah sama-rata.



Rajah 4.18 Tetapan Parameter SMOTE-RUS

Parameter nilai *randomSeed* ditetapkan ke nilai standard iaitu 5. Oleh kerana pengurangan jumlah subsampel telah dilakukan terlebih dahulu, nilai 100% dipilih dalam parameter *sampleSizePercent* supaya jumlah asal rekod adalah tetap dan tiada pengurangan subsampel berlaku.

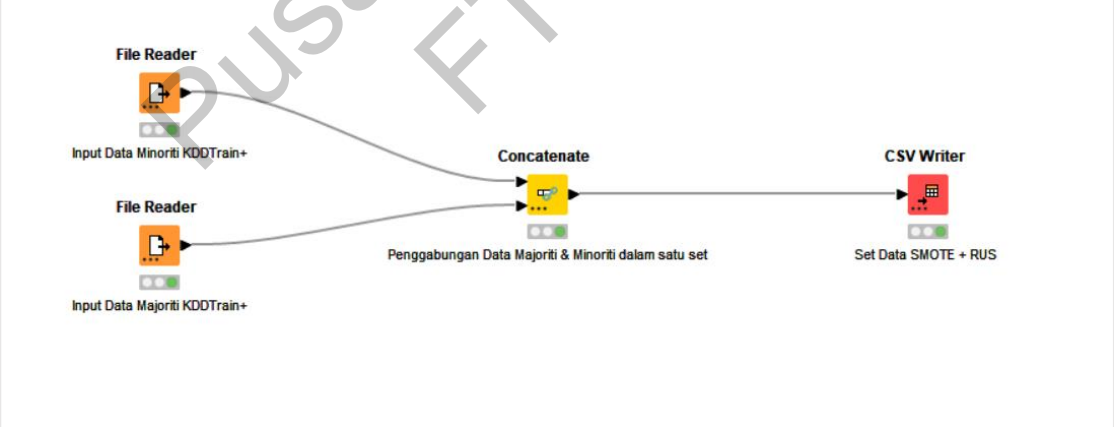
Rajah 4.19 Jumlah Rekod RUS Kelas Majoriti

Jumlah rekod kelas majoriti selepas pensampelan RUS dapat dilihat dalam Rajah 4.19. Jumlah rekod kelas majoriti dikurangkan dan kemudian diseimbangkan menggunakan fungsi *Resample* dalam WEKA dan menjadi 8,721 untuk kedua-dua kelas Normal dan DOS.

Jadual 4.5 Rekod Data Majoriti Selepas RUS

	Normal	**DOS**
Asal	47,141	32,149
Pengurangan 22%	10,371	7,072
RUS	8,721	8,721

Seperti dalam Jadual 4.5, jumlah rekod kelas Normal dan DOS pada asalnya adalah 79,290. Pengurangan sebanyak 22% menjadi 8,721 adalah untuk mendapatkan jumlah rekod yang hampir sama dengan kelas *Probe*, U2R dan R2L. Seterusnya, data majoriti dan minoriti digabungkan menggunakan nod *Concatenate* seperti dalam Rajah 4.20.

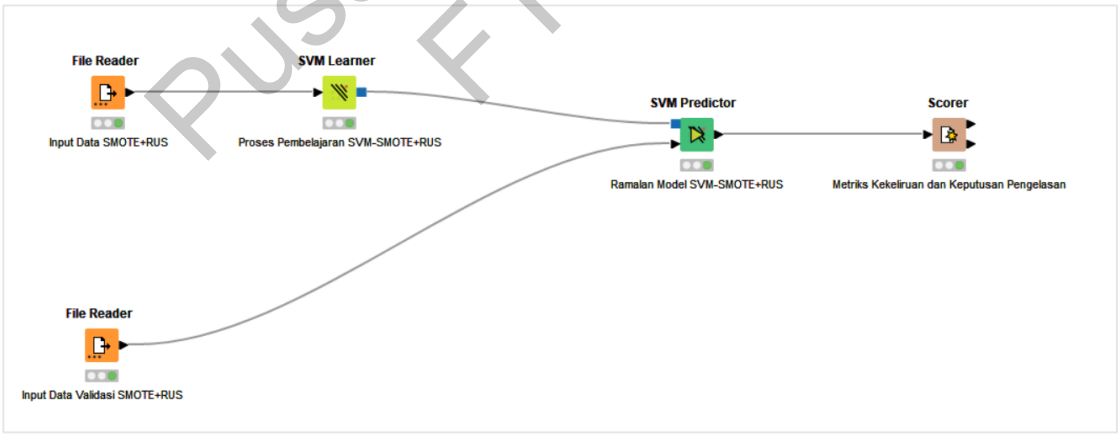


Rajah 4.20 Gabungan Data Majoriti dan Minoriti

Penggabungan data majoriti dan data minoriti menghasilkan jumlah rekod seperti dalam Rajah 4.20.

Rajah 4.21 Jumlah data SMOTE-RUS

Rajah 4.21 menunjukkan jumlah rekod data selepas teknik pensampelan SMOTE-RUS. Jumlah data SMOTE-RUS adalah 41,913.



Rajah 4.22 Pembangunan Model SMOTE-RUS

Berdasarkan Rajah 4.22, pembangunan model SMOTE-RUS bermula dengan memasukkan data SMOTE-RUS dan data validasi ke dalam File Reader secara berasingan. 70% data SMOTE-RUS dari *File Reader* ini kemudian akan dihantar ke

nod *SVM Learner* sebagai data pembelajaran. Data validasi 30% daripada nod *File Reader* dihantar ke nod *SVM Predictor* sebagai data ramalan. Setelah model ramalan SVM-SMOTE-RUS dibina, hasil keputusan akan direkodkan dalam nod *Scorer*.

4.5 KESIMPULAN

Dalam Bab 4, tumpuan diberikan kepada persiapan eksperimen untuk menguji keberkesanan teknik pembelajaran mesin dalam mengesan pencerobohan menggunakan set data NSL-KDD. Ia melibatkan pra-pemprosesan data yang merangkumi pengekodan One-Hot, penskalaan dan pembahagian set data. Bab ini juga menerangkan proses pembangunan SVM dan pelaksanaan teknik pensampelan seperti SMOTE, RUS, dan kombinasi kedua-duanya, iaitu SMOTE-RUS untuk mengatasi ketidakseimbangan data. Hasil ketiga-tiga teknik pensampelan tersebut menjana jumlah set data yang baharu seperti yang ditunjukkan dalam Jadual 4.5.

Jadual 4.6 Jumlah Set Data Keseluruhan

Set Data	Normal	DOS	Probe	R2L	U2R
Set Asas	47,141	32,149	8,159	696	36
Set SMOTE	47,141	47,141	47,141	47,141	47,141
Set RUS	17,636	17,636	17,636	17,636	17,636
Set SMOTE-RUS	8,721	8,721	8,157	8,157	8,157

Persediaan data ini penting untuk membangunkan model yang efektif dalam mengesan pencerobohan rangkaian. Seterusnya, set data ini akan dibangunkan bersama model pengelas SVM. Hasil dan analisis akan dibincangkan dalam bab seterusnya, Bab 5.

BAB V

DAPATAN KAJIAN

5.1 PENGENALAN

Bab ini membincangkan tentang dapatan kajian serta keputusan dan hasil penemuan kajian berdasarkan pelaksanaan pembangunan model ramalan pengelasan. Hasil output yang diperolehi berdasarkan penilaian model ramalan pengelasan seperti rekal, kepersisan, sensitiviti, skor-F dan ketepatan akan dibincangkan.

5.2 KEPUTUSAN PEMBANGUNAN DAN PENGUJIAN MODEL ASAS

Hasil eksperimen untuk pembangunan model asas adalah seperti dalam Jadual 5.1 di bawah. Model ini adalah model yang dibangunkan menggunakan 70% daripada KDDTrain+ dan 30% set validasi daripada set KDDTrain+.

Jadual 5.1 Keputusan Pembangunan Model Asas

	Positif Tulen	**Positif Palsu**	**Rekal**	**Kepersisan**	**Sensitiviti**	**Skor-F**	**Ketepatan**
Normal	20106	267	0.995	0.987	0.995	0.991	0.99
Dos	13614	25	0.988	0.998	0.988	0.993	0.999
Probe	3463	41	0.99	0.988	0.99	0.989	0.998
R2L	231	38	0.773	0.859	0.773	0.813	0.997
U2R	0.025	0	0.438	1	0.438	0.609	0.999
Ketepatan Keseluruhan				0.996			

Model asas ini merupakan model yang menggunakan set data asal NSL-KDD dan dilatih dengan SVM. Model asas SVM menggunakan parameter fungsi penalti bertindih (*overlapping penalty*) yang ditetapkan ke nilai 1.0. Manakala, kernel yang dipilih adalah kernel RBF dan nilai sigma menggunakan 1.0. Penetapan ini telah dibincangkan dalam Bab 4. Model asas dibangunkan sebagai penanda aras awal, yang mewakili prestasi pengelasan yang dilatih pada set data NSL KDD asal yang tidak seimbang. Ini berfungsi sebagai titik rujukan untuk membandingkan model yang dilatih dengan teknik pensampelan semula yang berbeza, seperti SMOTE, RUS dan SMOTE-RUS.

Berdasarkan Jadual 5.1, berikut merupakan keputusan pembangunan model asas SVM. Hasil keputusan menunjukkan ketepatan keseluruhan adalah 99.6% dengan menggunakan set validasi. Berdasarkan keputusan ketepatan keseluruhan sebanyak 99% berkemungkinan boleh dikaitkan dengan beberapa faktor kritikal di luar keberkesanan model. Isu seperti berat sebelah taburan kelas, potensi pemadanan berlebihan (*overfitting*) dan kebocoran ciri semasa dalam set validasi yang digunakan mungkin secara tidak sengaja mempengaruhi ketepatan ini (James et al. 2013; Kuhn et al. 2013). Namun begitu, pengujian model asas SVM ke atas set KDDTest+ memberi peratusan yang sebenar akan model ini dan akan dibincangkan berdasarkan Jadual 5.2.