การเรียนรู้ของเครื่องจักรเพื่อการตรวจจับการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย

วิทยานิพนธ์

ของ

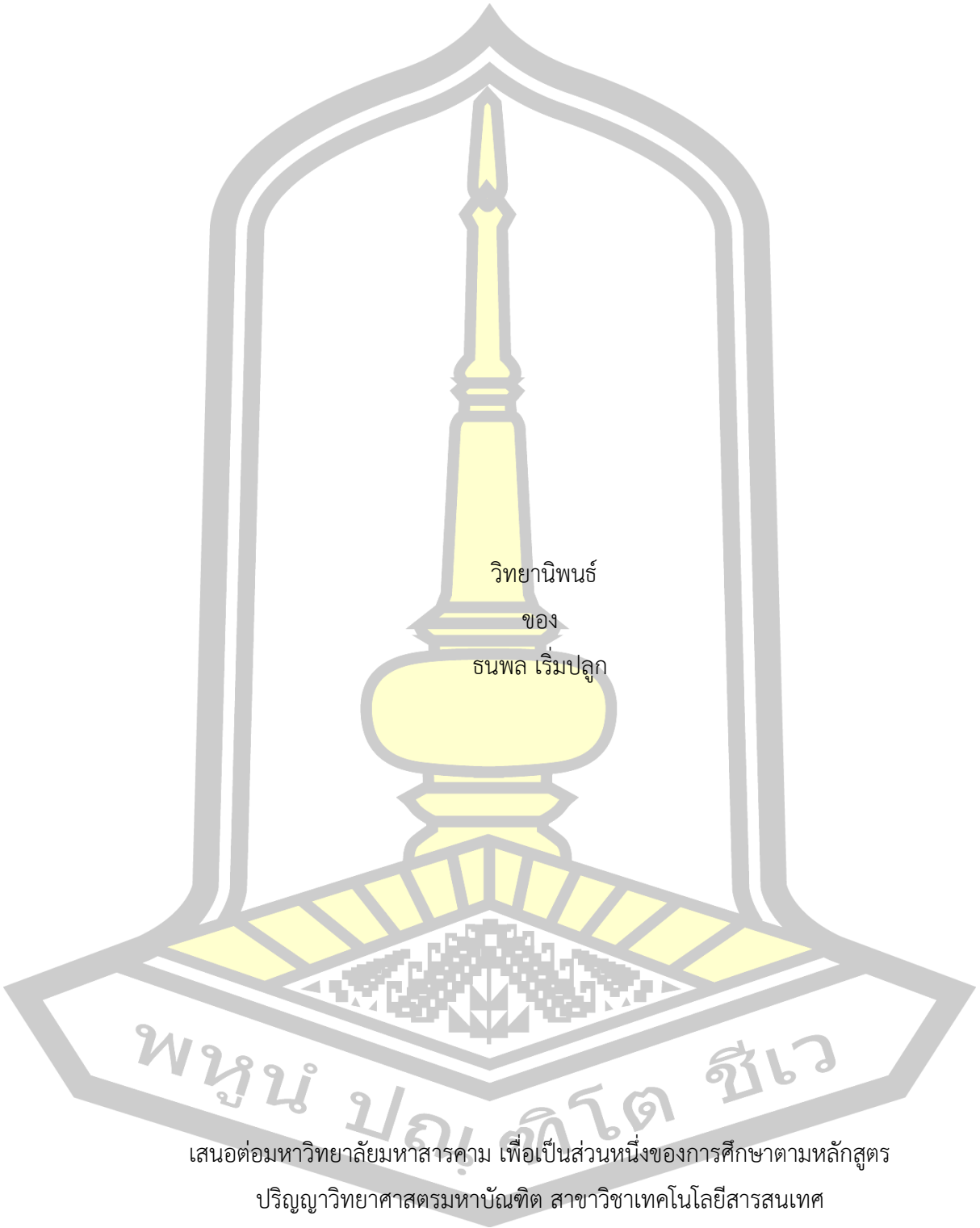
ธนพล เริ่มปลูก

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

สิงหาคม 2562

การเรียนรู้ของเครื่องจักรเพื่อการตรวจจับการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย



วิทยานิพนธ์

ของ

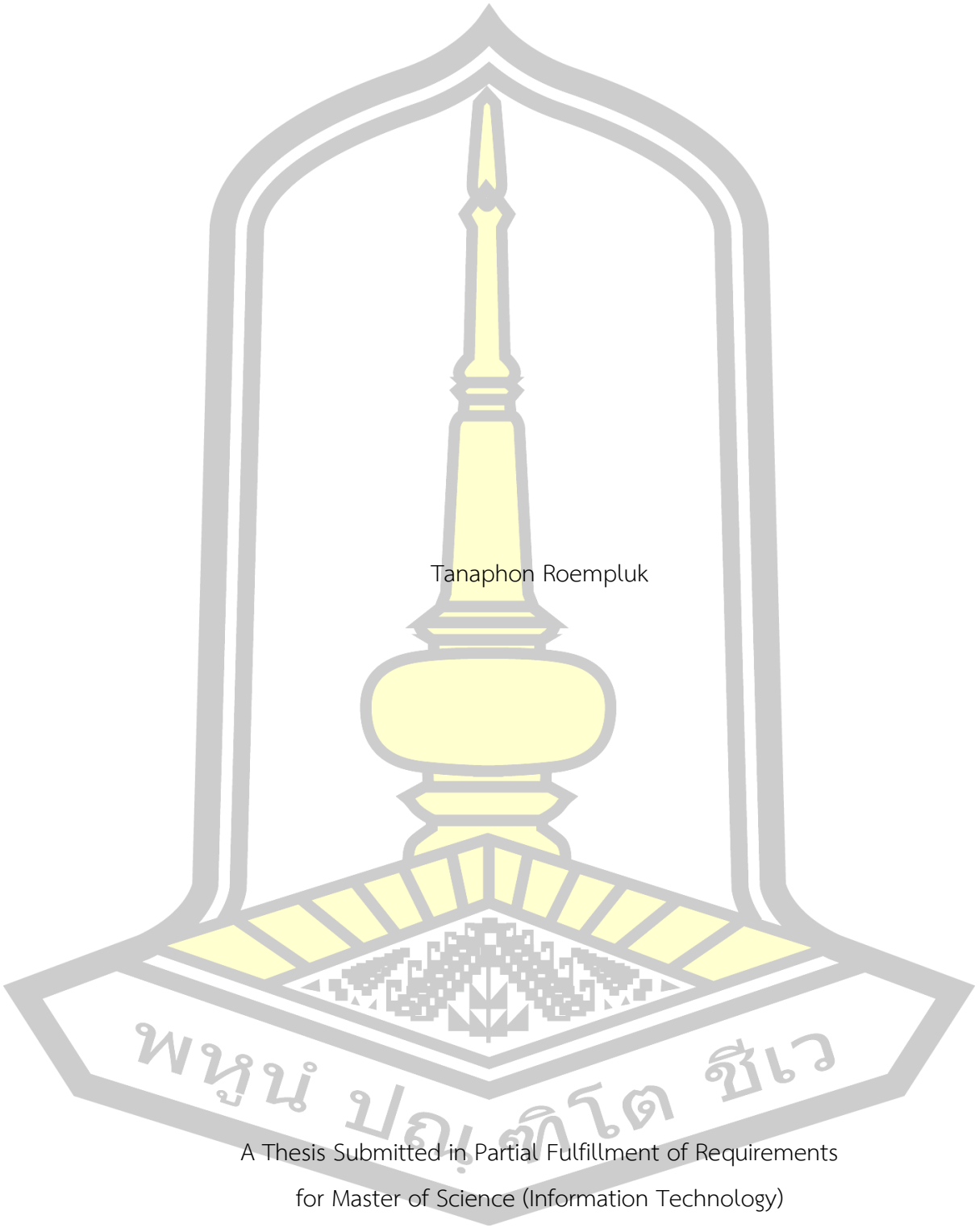
ธนพล เริ่มปลูก

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

สิงหาคม 2562

A Machine Learning Approach for Detecting Distributed Denial of Service Attacks



Tanaphon Roempluk

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

August 2019

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายธนพล เริ่มปลูก แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. ฉัตรตระกูล สมบัติธีระ)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. โอฬาริก สุรินต๊ะ)

.....กรรมการ
(ดร. สาธิต แสงประดิษฐ์)

.....กรรมการผู้ทรงคุณวุฒิภายนอก
(รศ. ดร. สิทธิชัย บุษหมั่น)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(ผศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเรียนรู้ของเครื่องจักรเพื่อการตรวจจับการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย

ผู้วิจัย ธนพล เริ่มปลูก

อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. โอฬาริก สุรินต๊ะ

ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ

มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

งานวิจัยฉบับนี้ได้นำเสนอวิธีการจำแนกการโจมตีแบบ DDoS โดยทดสอบกับ Benchmark Dataset จำนวน 2 ชุด ได้แก่ KDD CUP 1999 และ NSL-KDD โดยได้ตรวจสอบข้อมูลและลบข้อมูลที่ซ้ำกันออก ทำให้ข้อมูลชุด KDD Cup 1999 ที่มีจำนวน 4,898,431 ลดลงเหลือ 529,655 ชุด (record) และข้อมูล NSL-KDD ที่มีข้อมูลทั้งสิ้น 125,973 ชุด ลดลงเหลือเพียง 12,354 ชุด เท่านั้น ทั้งนี้เนื่องจาก การโจมตีแบบ DDoS จะเป็นการส่งข้อมูลโจมตีในรูปแบบเดิมซ้ำๆ ไปยังเครื่องเซิร์ฟเวอร์ จากนั้นจึงแปลงข้อมูลที่เป็นตัวอักษรให้อยู่ในรูปแบบของตัวเลข และส่งไปเรียนรู้ (Training) ด้วยวิธี K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP) และ Support Vector Machine (SVM) จากผลการทดลองสรุปได้ว่า วิธี KNN สามารถจำแนกการโจมตีแบบ Distributed Denial of Service (DDoS) ได้ถูกต้องแม่นยำที่สุด

คำสำคัญ : การโจมตีโดยปฏิเสธการให้บริการแบบกระจาย, วิธีการเพื่อนบ้านใกล้ที่สุด, ซัพพอร์ตเวกเตอร์แมชชีน, โครงข่ายประสาทเทียม

TITLE A Machine Learning Approach for Detecting Distributed Denial of Service Attacks

AUTHOR Tanaphon Roempluk

ADVISORS Assistant Professor Olarik Surinta , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

This research aims to present the method for identifying distributed denial of service (DDoS) attacks. Two benchmark dataset, including KDD CUP 1999 and NSL-KDD, were used. The dataset was checked and deleted duplicate data. After the process, the number of records of KDD Cup 1999 dataset was decreased from 4,898,431 records to 529,655 records, and the number of records of NSL-KDD dataset was decreased from 125,373 to only 12,354 records. The reduction of the records always happened because of the characteristics of DDoS attacks which send repeated data to the victims' server. The researchers converted alphabet data to numeric data, then training by K-nearest neighbor (KNN), multi-layer perceptron and support vector machine. The result showed that KNN was the best method to identify the DDoS attacks.

Keyword : Distributed denial of service (DDoS) attack, Multi-layer percetron (MLP), Support vector machine (SVM), K-nearest neighbor (KNN)

กิตติกรรมประกาศ

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.โอฬาริก สุรินต๊ะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่เสียสละเวลาช่วยเหลือให้คำปรึกษา คำแนะนำ แนวคิด ชี้แนะข้อบกพร่องและร่วมแก้ไขปัญหาติดตามความก้าวหน้าของงานวิจัยรวมทั้งฝึกฝนให้ผู้วิจัยมีทักษะทางด้านการคิด การอ่าน การเขียนและการนำเสนอผลงานทางวิชาการ ซึ่งเป็นประโยชน์อย่างมากในการพัฒนาตนเอง อีกทั้งให้ความเอาใจใส่ทำให้ผู้วิจัยสามารถดำเนินงานวิจัยจนประสบผลสำเร็จลุล่วงไปด้วยดี

ขอขอบคุณทุนสนับสนุนการศึกษาบุคลากรสายสนับสนุนจากมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณบิดามารดาและครอบครัวของข้าพเจ้าที่ได้ให้ชีวิตและโอกาสทางการศึกษา คอยเป็นกำลังใจและให้ความห่วงใยเสมอมา ตลอดจนคุณครูและอาจารย์ทุกท่านที่กรุณาประสิทธิ์ประสาทวิชาความรู้อันเป็นประโยชน์แก่ผู้วิจัย คุณค่าและประโยชน์อันพึงมาจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ธนพล เริ่มปลูก

สารบัญ

หน้า

สารบัญภาพ

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ความเป็นมาของการวิจัย

ในแต่ละองค์กรมีข้อมูลสารสนเทศ (Information) ที่ถูกเก็บรักษาอยู่ภายในระบบเครือข่ายที่สามารถเข้าถึงได้จากระบบเครือข่ายอินเตอร์เน็ตที่มีความเสี่ยงต่อการถูกโจมตีหากไม่มีการควบคุมหรือการป้องกันการโจมตีที่ดีทำให้การดูแลรักษาข้อมูลหรือระบบสารสนเทศเป็นสิ่งที่สำคัญมากในยุคของข้อมูลข่าวสาร (Information Age) ดังนั้น หากไม่มีระบบหรือวิธีการรักษาความปลอดภัยของข้อมูลที่ดีพอย่อมมีความเสี่ยงและอาจส่งผลเสียต่อองค์กรหากข้อมูลเกิดการรั่วไหลหรืออาจถูกโจมตีจากอาชญากรรมทางคอมพิวเตอร์ การโจมตีระบบเกิดขึ้นได้จากหลายวิธี เช่น จากผู้ใช้งานทั่วไปที่ขาดความรู้ จากผู้ใช้งานภายนอกองค์กร จากไวรัส (Virus) จากเวิร์ม (Worm) และจากม้าโทรจัน (Trojan Horse) โดยใช้ช่องทางในการโจมตีเช่น ระบบเครือข่ายภายใน อีเมล์ หรือทางบริการที่องค์กรเปิดให้สามารถเข้าถึงระบบได้ ผู้ที่ไม่หวังดีจึงใช้ช่องทางเหล่านี้เพื่อการโจมตีระบบสารสนเทศ เช่น เข้าใช้ระบบโดยไม่ได้รับอนุญาต(Access Attack) พยามแก้ไขข้อมูลระบบ (Modification Attack) การทำให้ข้อมูลเป็นเท็จ (Repudiation Attack) และทำให้ไม่สามารถเข้าถึงข้อมูลได้

จากการโจมตีด้วยวิธีต่างๆ ผู้โจมตีอาจมีเป้าหมายในการทำให้ระบบเครือข่ายคอมพิวเตอร์ทำงานช้าลงหรือหยุดการให้บริการ (Deny of Service Attack: DoS) ซึ่งเป็นการโจมตีด้วยโครงข่าย (Network Base Attack) หรือการโจมตีด้วยขนาด (Volumetric Base Attack) ผู้โจมตีจะส่งข้อมูลที่มีปริมาณมหาศาลไปที่เป้าหมายเพื่อทำให้การรับและส่งข้อมูลเกิดการติดขัด (Congestion) จนไม่สามารถติดต่อสื่อสารกับผู้ใช้งานทั่วไปได้ โดยใช้เทคนิคคือ SYN Flood, UDP Flood, Ping of Death, Reflection และ Amplification เป็นต้น ซึ่งการโจมตีประเภทนี้จะพบบ่อยและเกิดขึ้นได้ง่ายและสามารถป้องกันได้ยากหรือใช้การโจมตีด้วยแอพพลิเคชั่น (Application Base Attack) เพื่อมุ่งเน้นไปให้แอพพลิเคชั่นหยุดทำงาน ซึ่งการโจมตีชนิดนี้จะอยู่ในระดับที่สูงกว่าการโจมตีด้วยโครงข่ายจะต้องอาศัยความเชี่ยวชาญของผู้ดูแลระบบเพื่อคัดกรองแยกแยะข้อมูลที่ส่งเข้ามายังระบบ เทคนิคที่อยู่ในประเภทนี้เช่น HTTP flood, SSL Floodและ Slowloris

ผู้โจมตีสามารถโจมตีผ่านทางช่องโหว่ของระบบหรือลักลอบติดตั้งโปรแกรมที่ใช้สำหรับการโจมตีรูปแบบอื่นๆ ซึ่งอาจเกิดผลกระทบที่รุนแรงตามมา [1] ดังนั้น ระบบสารสนเทศภายในองค์กรจึงมีความจำเป็นต้องมีระบบรักษาความปลอดภัยที่ดีและสามารถตรวจจับภัยคุกคาม (Threat) ที่โจมตีระบบในรูปแบบที่หลากหลายและหลากหลายช่องทางหรือการใช้ช่องทางที่องค์กรเปิดให้บริการใน

การโจมตีในรูปแบบดังกล่าวทำให้การตรวจสอบเป็นไปได้ยากเนื่องการมีรูปแบบการโจมตีเหมือนการใช้งานแบบปกติ

การรักษาความปลอดภัยของระบบเครือข่ายคอมพิวเตอร์ในปัจจุบันไม่มีระบบใดที่สามารถป้องกันจากภัยคุกคามและช่องโหว่ (Vulnerability) ได้ทั้งหมด ผู้ดูแลระบบต้องบริหารจัดการความเสี่ยง (Risk) ที่อาจเกิดขึ้นและการวิเคราะห์การรักษาความปลอดภัย (Security) ว่ามีความปลอดภัยหรือไม่โดยการวิเคราะห์จากคุณสมบัติ 3 ด้านคือ ความลับของข้อมูล (Confidentiality) การรักษาความถูกต้องและความคงสภาพ (Integrity) และการพร้อมใช้งาน (Availability) [2]

งานวิจัยฉบับนี้มุ่งเน้นศึกษาการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย (Distributed Denial of Service: DDoS) [3–5] เป็นการโจมตีที่จะทำการเชื่อมต่อ (Connection) จากเครื่องที่ใช้ในการโจมตีตั้งแต่สองเครื่องขึ้นไป เพื่อทำให้มีการใช้งานจนถึงขีดจำกัดของเครื่องคอมพิวเตอร์ที่เปิดให้บริการทำให้ผู้ใช้งาน (Client) ไม่สามารถเชื่อมต่อหรือใช้บริการได้ทำให้ประสิทธิภาพลดลงหรือทำให้ระบบคอมพิวเตอร์หยุดทำงาน เนื่องจากการโจมตีจะทำการเพิ่มปริมาณการเชื่อมต่อจำนวนมากทำให้ปริมาณการใช้แบนด์วิดธ์และการประมวลผลมีปริมานสูงขึ้นมากผิดปกติการโจมตีแบบนี้อาจใช้โปรโตคอล (Protocol) ที่ใช้งานทั่วไปบนระบบเครือข่ายคอมพิวเตอร์ เช่น Transmission Control Protocol (TCP) หรือ Internet Control Message Protocol (ICMP) ที่เป็นจุดอ่อนและสามารถใช้ในการโจมตีระบบ หรือใช้ช่องโหว่ของระบบรักษาความปลอดภัยเป็นช่องทางในการโจมตีระบบได้ และหารูปแบบการโจมตีแบบ DDoS โดยใช้วิธีการเรียนรู้ของเครื่องจักร (Machine Learning) เพื่อการจำแนกการข้อมูลแบบปกติและข้อมูลที่ถูกโจมตีแบบ DDoS ที่ผ่านเข้ามาในระบบเครือข่ายคอมพิวเตอร์

1.2 วัตถุประสงค์ของการวิจัย

เพื่อนำเสนอกระบวนการวิเคราะห์การโจมตีแบบ DDoS ด้วยการเรียนรู้ของเครื่องจักรภายใต้แนวคิดของการจำแนกข้อมูล

1.3 ความสำคัญของการวิจัย

1.3.1 ทำให้ทราบถึงวิธีการการจำแนกการโจมตีแบบ DDoS ที่เกิดขึ้นภายในระบบเครือข่ายคอมพิวเตอร์ เพื่อเลือกคุณลักษณะที่จำเป็นในการตรวจสอบการโจมตี

1.3.2 ทำให้ทราบถึงกระบวนการเลือกคุณลักษณะที่ใช้ในการจำแนกการการโจมตีแบบ DDoS

3. สามารถนำวิธีการในการจำแนกการโจมตีแบบ DDoS เพื่อพัฒนาวิธีการป้องกันระบบเครือข่ายคอมพิวเตอร์

1.4 ขอบเขตในการวิจัย

1.4.1 ศึกษาและวิเคราะห์การโจมตีแบบ DDoS เท่านั้น

1.4.2 นำเสนอวิธีการจำแนกการโจมตีแบบ DDoS โดยวิธีการเรียนรู้ของเครื่องจักรอย่างน้อย 2 เพื่อเปรียบเทียบ

1.5 นิยามศัพท์เฉพาะ

1.5.1 การโจมตีปฏิเสธการให้บริการแบบกระจาย (Distributed Denial of Service) หมายถึง การโจมตีที่คอมพิวเตอร์จำนวนมากจะทำการเชื่อมต่อจำนวนมากในความถี่สูงจนถึงขีดจำกัดของเครื่องแม่ข่ายคอมพิวเตอร์ส่งผลให้เครื่องไม่สามารถทำงานได้หรือไม่สามารถทำการเชื่อมต่อใหม่ได้

1.5.2 การจำแนก (Classification) หมายถึง การจำแนกข้อมูลการจราจรบนระบบเครือข่ายคอมพิวเตอร์แบบปกติและการโจมตีแบบ DDoS ที่เชื่อมต่อเข้ามาในระบบเครือข่ายคอมพิวเตอร์โดยใช้ข้อมูลที่ถูกจัดเก็บได้จริงในระบบเครือข่ายคอมพิวเตอร์

1.5.3 บันทึกข้อมูลการใช้งานระบบเครือข่าย (Logfile) หมายถึง ข้อมูลการจารจรภายในระบบเครือข่ายคอมพิวเตอร์ที่เกิดขึ้นจากระบบเครือข่ายภายในและระบบเครือข่ายภายนอก

1.5.4 การจำลองการโจมตีระบบเครือข่าย หมายถึง การโจมตีระบบเครือข่ายที่สร้างขึ้นด้วยเครื่องคอมพิวเตอร์และใช้โปรแกรมสร้างการโจมตีแบบ DDoS โจมตีเครื่องแม่ข่ายคอมพิวเตอร์ที่เปิดให้บริการของหน่วยงานที่เปิดให้บริการจริง

1.5.5 ระบบเครือข่ายภายนอก หมายถึง ระบบเครือข่ายคอมพิวเตอร์ที่เชื่อมต่อเข้าสู่ระบบเครือข่ายขององค์กร เพื่อเข้าถึงระบบเครือข่ายภายในหรือบริการขององค์กร

1.5.6 ระบบเครือข่ายภายใน หมายถึง ระบบเครือข่ายคอมพิวเตอร์ที่เชื่อมต่อสื่อสารกันภายในองค์กรเพื่อให้มีความเร็วหรือสิทธ์การเข้าถึงข้อมูลสารสนเทศ

1.5.7 การโจมตี หมายถึง ผู้ที่ประสงค์ร้ายหรือผู้ที่ไม่มีสิทธิ์พยายามที่จะบุกรุกเครือข่ายเพื่อลักลอบเข้าถึงข้อมูลที่สำคัญหรือพยายามแก้ไขข้อมูลหรือการทำให้ระบบไม่สามารถใช้การได้โดยไม่ได้รับอนุญาต

1.5.8 ผู้โจมตีระบบเครือข่าย (Attacker) หมายถึง บุคคลที่ไม่มีสิทธิ์หรือไม่ได้รับอณุญาติให้เข้าถึงข้อมูลหรือเข้าถึงระบบเครือข่ายคอมพิวเตอร์

1.5.9 เป้าหมายการโจมตี (Attack Target) หมายถึง หมายถึง เครื่องคอมพิวเตอร์หรือเครื่องแม่ข่ายคอมพิวเตอร์ที่เป็นเป้าหมายในการโจมตี

1.5.10 เครื่องแม่ข่ายคอมพิวเตอร์ (Server) หมายถึง เครื่องคอมพิวเตอร์ที่เปิดให้บริการต่างๆในระบบเครือข่าย

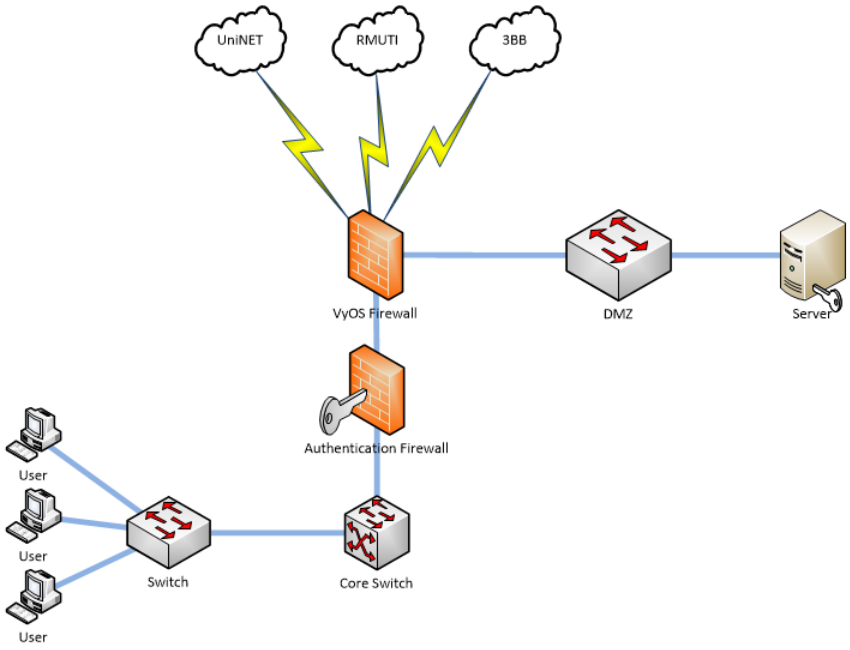
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้จะกล่าวถึงงานวิจัยและทฤษฎีที่เกี่ยวกับ โครงสร้างระบบเครือข่าย (Infrastructure) การสื่อสารในระบบเครือข่ายคอมพิวเตอร์ รูปแบบการโจมตีแบบ การตรวจสอบการโจมตีโดยปฏิเสธการให้บริการแบบกระจาย ความผิดปกติที่เกิดขึ้นจากกการโจมตี เพื่อใช้เป็นแนวทางในการตรวจสอบรูปแบบความผิดปกติและรูปแบบการใช้งานปกติที่เกิดขึ้นซึ่งมีงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดดังต่อไปนี้

2.1 ระบบเครือข่ายคอมพิวเตอร์ (Computer Network)
2.2 การสื่อสารในระบบเครือข่ายคอมพิวเตอร์ (Data Communications)
2.3 วิธีการโจมตีระบบ (Type of Network Attacks)
2.4 รูปแบบการโจมตีแบบ DDoS (Type of DDoS Attacks)
2.5 การแปลงชุดขอมูล (Data Transformation)
2.6 การคัดเลือกคุณลักษณะ (Feature Selection)
2.7 วิธีการลดมิติข้อมูล (Dimensionality Reduction)
2.8 เทคนิคการเรียนรู้ของเครื่องจักร (Machine Learning)
2.9 การทดสอบประสิทธิภาพ (Performance Evaluation)
2.10 งานวิจัยที่เกี่ยวข้อง (Related Work)

2.1 ระบบเครือข่ายคอมพิวเตอร์ (Computer Network)

ระบบเครือข่ายคอมพิวเตอร์ หมายถึง ระบบเครือข่ายของคอมพิวเตอร์ที่เชื่อมต่อกันตั้งแต่สองเครื่องขึ้นไปโดยผ่านตัวกลางหรือสื่อในรูปแบบต่างๆและสามารถสามารถใช้ทรัพยากร (Resources) ร่วมกันได้ เช่น ฮาร์ดดิสก์ สแกนเนอร์ เครื่องพิมพ์ ซีดีรอม เป็นต้น สามารถแลกเปลี่ยนข้อมูลได้ ถือว่าเป็นระบบเครือข่ายคอมพิวเตอร์ เช่น ระบบเครือข่ายของ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์ดัง ภาพประกอบที่ 2.1 แสดงถึงการเชื่อมต่อกลุ่มของคอมพิวเตอร์หลายๆกลุ่มเข้าด้วยกันจนเกิดระบบเครือข่ายภายในองค์กรขึ้นโดยการเชื่อมต่อของระบบคอมพิวเตอร์มีการเชื่อมต่อกับระบบเครือข่ายภายในและระบบเครือข่ายภายนอกและมีการเปิดให้บริการระบบของหน่วยงานซึ่งเชื่อมต่อเครื่องคอมพิวเตอร์แม่ข่ายของระบบเข้าสู่ระบบเครือข่ายคอมพิวเตอร์



ภาพประกอบที่ 2.1 โครงสร้างระบบเครือข่ายภายในมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน

2.1.1 ระบบเครือข่ายภายนอก (WAN Zone) กลุ่มนี้เชื่อมต่อผ่านระบบเครือข่ายคอมพิวเตอร์ของกระทรวงเทคโนโลยีสารสนเทศ (UniNET) โดยมีสื่อกลาง คือ ใยแก้วนำแสงที่มีความเร็วในการเชื่อมต่อ 1 Gbps และมีช่องทางสำรอง เชื่อมต่อระบบเครือข่ายของผู้ให้บริการ (Internet Service Provider: ISP) คือ 3BB โดยมีความเร็วในการเชื่อมต่อ 50 Mbps เป็นกลุ่มเครือข่ายทำหน้าที่เป็นกลางในการเชื่อมต่อระบบเครือข่ายขององค์กรกับระบบเครือข่ายภายนอกอื่นๆ

2.1.2 ระบบเครือข่ายระหว่างวิทยาเขต (Campus Zone) กลุ่มนี้เชื่อมต่อระบบเครือข่ายคอมพิวเตอร์ของ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ผ่านผู้ให้บริการเครือข่าย ISP โดยมีสื่อกลาง คือ ใยแก้วนำแสงมีความเร็วในการเชื่อมต่อ 100 Mbps เชื่อมต่อระบบ 4 วิทยาเขต ดังนี้

- ระบบเครือข่ายของมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน นครราชสีมา
- ระบบเครือข่ายของวิทยาเขตสกลนคร
- ระบบเครือข่ายของวิทยาเขตขอนแก่น
- ระบบเครือข่ายของวิทยาเขตสุรินทร์

2.1.3 ระบบเครือข่ายเครื่องแม่ข่ายคอมพิวเตอร์ (Demilitarized Zone: DMZ) กลุ่มนี้เชื่อมต่อระบบเครือข่ายคอมพิวเตอร์ของเครื่องแม่ข่ายคอมพิวเตอร์ที่ติดตั้งและให้บริการ เช่น Web Server, Data Base Server, DHCP Server และ DNS Server ถูกควบคุมการติดต่อสื่อสารด้วยไฟล์วอล

2.1.4 ระบบเครือข่ายภายใน (Internal Zone) กลุ่มนี้เชื่อมต่อระบบเครือข่ายภายในมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์ โดยมีการควบคุมอยู่ในระดับต่ำทุกๆชุดข้อมูลสามารถติดต่อสื่อสารโดยไม่ถูกไฟล์วอลควบคุม มีการเชื่อมต่อกลุ่มของเครือข่ายภายในองค์กรจำนวนมากที่ถูกสร้างขึ้น

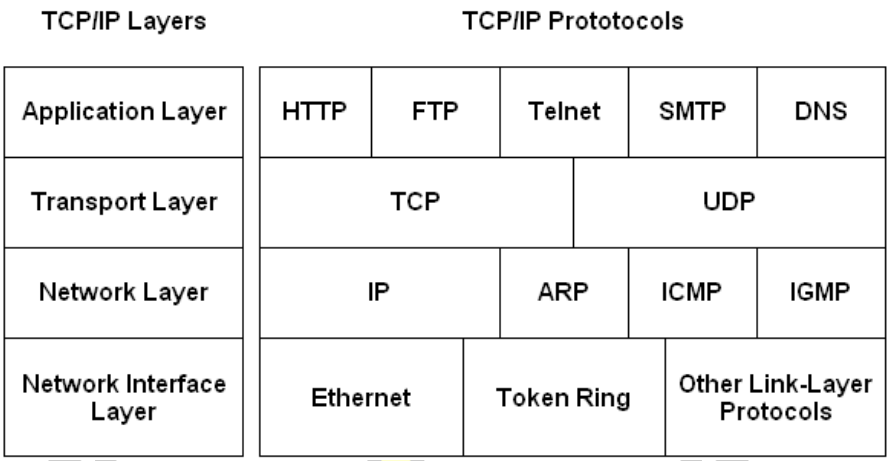
2.2 การสื่อสารในระบบเครือข่ายคอมพิวเตอร์ (Data Communications)

ภายในระบบเครือข่ายคอมพิวเตอร์มีการแลกเปลี่ยนข้อมูลระหว่างผู้ส่งข้อมูล (Sender) ไปสู่ผู้รับข้อมูล (Receiver) โดยอาศัยสื่อนำสัญญาณในรูปแบบสายหรือไร้สายและถูกควบคุมด้วยระบบปฏิบัติการเครือข่าย (Network Operating System: NOS) มีหน้าที่บริหารจัดการและควบคุมการใช้ทรัพยากรของระบบเครือข่าย ให้การสื่อสารนั้นเกิดขึ้นได้อย่างถูกต้องและข้อมูลครบถ้วน

2.2.1 การสื่อสารข้อมูล (Transmission) การสื่อสารข้อมูล คือ การแลกเปลี่ยนข้อมูลระหว่างต้นทางหรือผู้ส่งข้อมูลกับปลายทางหรือผู้รับข้อมูลโดยเครื่องคอมพิวเตอร์ซึ่งมีตัวกลาง เช่น ซอฟต์แวร์คอมพิวเตอร์ (Software) สำหรับควบคุมการส่งและการไหลของข้อมูลจากต้นทางไปยังปลายทาง โดยข้อมูลที่อยู่ต้นทางจะต้องจัดเตรียมนำเข้าสู่อุปกรณ์สำหรับส่งข้อมูลผ่านระบบเครือข่ายคอมพิวเตอร์ซึ่งข้อมูลเหล่านี้จะถูกทำให้เป็นชิ้นเล็กๆหรือหน่วยย่อยของข้อมูล (Package) ที่สามารถส่งข้อมูลได้ ข้อมูลที่ถูกส่งจากต้นทางเมื่อไปถึงปลายทางก็ดำเนินการรวบรวมหน่วยย่อยของข้อมูลเหล่านั้นเพื่อนำไปใช้ประโยชน์ การส่งข้อมูลผ่านคอมพิวเตอร์จำเป็นต้องมีโปรแกรมสำหรับดำเนินการและควบคุมการส่งข้อมูลเพื่อให้ได้ข้อมูลตามที่กำหนดไว้ ได้แก่ Linux, UNIX และ Windows Server

2.2.2 ตัวกลางหรือสื่อกลางทำหน้าที่เชื่อมต่อเครื่องคอมพิวเตอร์ในรูปแบบต่างๆจากผู้ส่งไปยังผู้รับซึ่งมีหลายรูปแบบ เช่น สายบิดคู่ตีเกลียว (Twisted Pairs) สายไฟเบอร์ออฟติก (Fiber Optic) ตัวกลางอาจจะอยู่ในรูปของคลื่นที่ส่งผ่านทางอากาศเช่น คลื่นไมโครเวฟ คลื่นดาวเทียม หรือคลื่นวิทยุ เป็นต้น ซึ่งการเลือกตัวกลางจะขึ้นอยู่กับประเภทและปริมาณการรับและการส่งข้อมูล (Data Transfer)

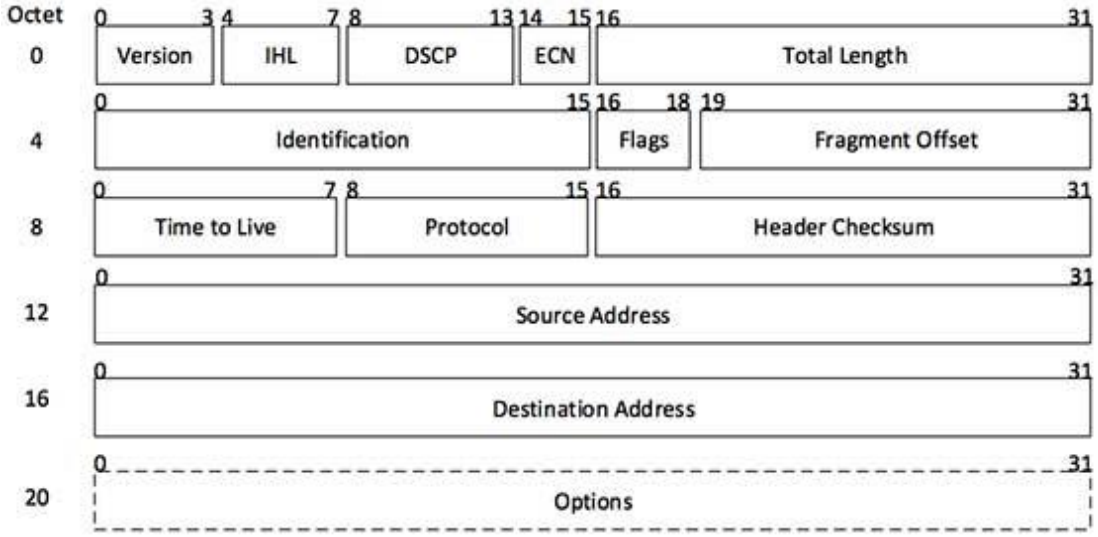
2.2.3 โปรโตคอล (Protocol) คือ กฎระเบียบหรือวิธีการใช้เป็นข้อกำหนดสำหรับการสื่อสารข้อมูลของคอมพิวเตอร์ที่ใช้ในการสื่อสารกัน เพื่อให้ผู้รับและผู้ส่งเข้าใจกันได้ซึ่งมีหลายชนิดให้เลือกใช้ เช่น TCP/IP, X.25 และ SDLC เป็นต้น ซึ่งโปรโตคอลที่ใช้ในระบบอินเตอร์เน็ตจะใช้ภาษาสื่อสารมาตรฐานที่ชื่อว่า Transmission Control Protocol/Internet Protocol (TCP/IP) เป็นภาษาหลักในการเชื่อมโยงเข้าสู่อินเทอร์เน็ตโดยมีวัตถุประสงค์เพื่อให้สามารถใช้สื่อสารจากต้นทางข้ามเครือข่ายไปยังปลายทางได้และสามารถหาเส้นทางที่จะส่งข้อมูลไปได้เองโดยอัตโนมัติ ดังภาพประกอบที่ 2.2 แสดงถึงชั้นของ TCP/IP ในแต่ละชั้น



ภาพประกอบที่ 2.2 โครงสร้างของโปรโตคอล TCP/IP [6]

2.2.4 หน้าที่ของ Transmission Control Protocol การแยกข้อมูลเป็นหน่วยย่อยของข้อมูลและส่งออกไปส่วน TCP ปลายทางจะทำการรวบรวมหน่วยย่อยของข้อมูลเข้าด้วยกันเพื่อนำไปประมวลผลโดยระหว่างการรับส่งข้อมูลนั้นก็จะมีการตรวจสอบความถูกต้องของข้อมูลด้วยถ้าเกิดผิดพลาด TCP ปลายทางก็จะขอไปยัง TCP ต้นทางให้ส่งข้อมูลมาใหม่อีกครั้งจนกว่าการสื่อสารจะสิ้นสุดลง

2.2.5 หน้าที่ของ Internet Protocolทำหน้าที่ในการจัดส่งข้อมูลจากเครื่องต้นทางไปยังเครื่องปลายทางโดยอาศัยหมายเลขประจำเครื่องคอมพิวเตอร์ (Internet Protocol Address: IP Address) ในระบบเครือข่ายที่ใช้โปรโตคอลแบบ TCP/IP จำเป็นจะต้องมีหมายเลข IP Address กำหนดไว้ให้กับคอมพิวเตอร์และอุปกรณ์ที่ต้องการในการแลกเปลี่ยนข้อมูล ทำให้ทราบถึงตำแหน่งของเครื่องที่เราต้องการส่งข้อมูลไปซึ่งประกอบด้วยตัวเลข 4 ชุด มีเครื่องหมายจุดขั้นระหว่างชุด เช่น 172.22.0.0 หรือ 203.158.199.0 เป็นต้น โดยหมายเลข IP Address ของเครื่องคอมพิวเตอร์จะมีค่าไม่ซ้ำกันตัวเลข 4 ชุดนี้จะแบ่งออกเป็น 2 ส่วนคือ Network ID กับ Host ID จาก IP Address สามารถบอกได้ว่าคอมพิวเตอร์ A และ B อยู่ใน Network ID เดียวกันหรือเปล่าโดยการเปรียบเทียบ Network ID ของ IP Address ถ้ามี Network ID ตรงกันก็แสดงว่าอยู่ใน Network เดียวกัน เช่น คอมพิวเตอร์ A มี IP Address 192.168.1.1/24 จะอยู่ใน Network วงเดียวกับคอมพิวเตอร์ B ซึ่งมี IP Address 192.168.1.200/24 เนื่องจากมี Network ID ตรงกันดังภาพประกอบที่ 2.3 แสดงข้อมูลภายใน IP Header ดังตารางที่ 2.1



ภาพประกอบที่ 2.3 แสดงข้อมูล IP Header [7]

ตารางที่ 2.1 ข้อมูลภายใน IP Header

ชื่อ	ความหมาย
IHL	ความยาวของส่วนหัว IP ทั้งหมด
DSCP	นี่คือประเภทของบริการ
ECN	ข้อมูลเกี่ยวกับความแออัดที่เห็นในเส้นทาง
Total Length	ความยาวของ IP Packet ทั้งหมด
Identification	หมายเลข IP ของชิ้นส่วนทั้งหมดของแพ็คเก็ตระหว่างการส่งข้อมูล
Flags	สามารถแยกส่วนแพ็คเก็ตได้หรือไม่
Fragment Offset	บอกตำแหน่งที่แน่นอนของชิ้นส่วนของแพ็คเก็ต
Time to Live	จำนวนของเราเตอร์ที่แพ็คเก็ตสามารถส่งผ่านได้ เมื่อค่าถึงศูนย์แพ็คเก็ตจะถูกยกเลิก
Protocol	บอกเลเยอร์เครือข่ายที่ปลายทาง ยกตัวอย่างเช่นจำนวนของโปรโตคอล ICMP คือ 1, TCP คือ 6 และ UDP คือ 17
Header Checksum	ค่าการตรวจสอบของส่วนหัวที่ใช้แล้วเพื่อตรวจสอบว่าแพ็คเก็ตที่ได้รับปราศจากข้อผิดพลาด
Source Address	ที่อยู่ 32 บิตของผู้ส่ง
Destination Address	ที่อยู่ 32 บิตของผู้รับ
Options	ฟิลด์ตัวเลือกซึ่งจะใช้ถ้าค่าของ IHL

2.3 วิธีการโจมตีระบบ (Type of Network Attacks)

วิธีการตรวจจับการโจมตีตามลักษณะการทำงานได้ 2 วิธีการตรวจจับการใช้งานที่ผิด (Misuse Detection) วิธีนี้ต้องทราบถึงสัญลักษณ์ (Signature) ของการโจมตี ถ้าตรวจพบสัญญาณดังกล่าวในระบบเครือขายคอมพิวเตอร ก็จะส่งสัญญาณเตือนว่ามีการโจมตีเกิดขึ้น วิธีการตรวจจับแบบนี้มีข้อจำกัดตรงที่ตรวจจับได้เฉพาะการโจมตีที่รู้จักสัญลักษณและจำเป็นต้องมีการปรับปรุงสัญลักษณเมื่อมีการโจมตีแบบใหม่เกิดขึ้นและวิธีตรวจจับการใช้งานที่ผิดปกติ (Anomaly Detection) วิธีนี้ต้องเรียนรู้และจำลองพฤติกรรมการใช้งานที่ปกติก่อนและจัดเก็บไว้เป็นข้อมูลอ้างอิง เมื่อตรวจพบการใช้งานที่แตกต่าง ก็ส่งสัญญาณเตือนว่ามีการโจมตีเกิดขึ้นวิธีการตรวจสอบแบบนี้มีข้อดีกว่าการตรวจสอบแบบแรกตรงที่สามารถตรวจจับการโจมตีแบบใหม่ได้ โดยไม่จำเป็นต้องรู้ถึงสัญลักษณของการโจมตี เช่น ข้อมูลการไหลของโพรโทคอลชั้นอินเทอรเน็ต (IP Flow) สามารถใช้ข้อมูลนี้ในการตรวจสอบการโจมตี โดยทำการพิจารณาจากปริมาณทราฟฟกที่สูงผิดปกติ จะสังเกตได้ง่ายเพราะปริมาณทราฟิกเพิ่มสูงขึ้นอย่างรวดเร็ว ผู้โจมตีระบบเครือข่ายคอมพิวเตอร์จึงใช้วิธีการที่หลากหลายในการโจมตีระบบเพื่อทำให้ระบบไม่สามารถใช้งานได้ พยามเข้าใช้งานระบบหรือการหวังผลอื่นๆ โดยมีวิธีการดังนี้

2.3.1 แพ็กเก็ตสนิฟเฟอร์ (Sniffer) ข้อมูลที่คอมพิวเตอร์ส่งผ่านเครือข่ายนั้นจะถูกแบ่งย่อยเป็นแพ็กเก็ตแอพพลิเคชันหลายชนิดจะส่งข้อมูลโดยไม่เข้ารหัส (Encryption) หรือในรูปแบบเคลียร์เท็กซ์ (Clear Text) ดังนั้น ข้อมูลอาจจะถูกคัดลอกและโพรเซสโดยแอพพลิเคชันอื่นก็ได้

2.3.2 ไอพีสปูฟิง (IP Spoonfing) การที่ผู้บุกรุกอยู่นอกเครือข่ายแล้วแกล้งทำเป็นว่าเป็นคอมพิวเตอร์ที่เชื่อถือได้ (Trusted) โดยอาจจะใช้ไอพีแอดเดรสเหมือนกับที่ใช้ในเครือข่ายหรืออาจจะใช้ไอพีแอดเดรสข้างนอกที่เครือข่ายเชื่อว่าเป็นคอมพิวเตอร์ที่เชื่อถือได้หรืออนุญาตให้เข้าใช้ทรัพยากรในเครือข่ายได้โดยปกติและทำการเปลี่ยนแปลงหรือเพิ่มข้อมูลเข้าไปในแพ็กเก็ตที่รับส่งระหว่างไคลเอนท์และเซิร์ฟเวอร์ หรือคอมพิวเตอร์ที่สื่อสารกันในเครือข่ายการที่จะทำอย่างนี้ได้ผู้บุกรุกจะต้องปรับเร้าติ้งเทเบิล (Routing Table) ของเราท์เตอร์เพื่อให้ส่งแพ็กเก็ตไปยังเครื่องของผู้บุกรุกหรืออีกวิธีหนึ่งคือการที่ผู้บุกรุกสามารถแก้ไขให้แอพพลิเคชันส่งข้อมูลที่เป็นประโยชน์ต่อการเข้าถึงแอพพลิเคชันนั้นผ่านทางอีเมลล์หลังจากนั้นผู้บุกรุกก็สามารถเข้าใช้แอพพลิเคชันได้โดยใช้ข้อมูลดังกล่าว

2.3.3 การโจมตีรหัสผ่านการโจมตีรหัสผ่าน (Password Attacks) หมายถึงการโจมตีที่ผู้บุกรุกพยายามเดารหัสผ่านของผู้ใช้คนใดคนหนึ่งซึ่งวิธีการเดานั้นก็มีหลายวิธี เช่น บรู๊ทฟอร์ช (Brute-Force) เป็นต้น

2.3.4 การโจมตีแบบปลอมเป็นคนกลาง (Man In The Middle: MITM) นั้นผู้โจมตีต้องสามารถเข้าถึงแพ็กเก็ตที่ส่งระหว่างเครือข่ายได้เช่น ผู้โจมตีอาจอยู่ที่ ISP ซึ่งสามารถตรวจจับแพ็กเก็ตที่รับส่งระหว่างเครือข่ายภายในและเครือข่ายอื่นโดยผ่าน ISP การโจมตีนี้จะใช้แพ็กเก็ตสนิฟเฟอร์เป็น

เครื่องมือเพื่อขโมยข้อมูลหรือใช้เซสชั่นเพื่อแอ็กเซสเครือข่ายภายในหรือวิเคราะห์การจราจรของเครือข่ายหรือผู้ใช้

2.3.5 การโจมตีแบบการปฏิเสธการให้บริการการโจมตีเซิร์ฟเวอร์โดยการทำให้เซิร์ฟเวอร์นั้นไม่สามารถให้บริการได้ ซึ่งปกติจะทำโดยการใช้ทรัพยากรของเครื่องจนถึงขีดจำกัด ตัวอย่างเช่น การโจมตีจะทำได้โดยการเปิดการเชื่อมต่อ (Connection) กับเครื่องแม่ข่ายคอมพิวเตอร์จนถึงขีดจำกัดทำให้ผู้ใช้อื่นไม่สามารถเข้ามาใช้บริการได้ สามารถแบ่งได้ 2 กลุ่มคือ

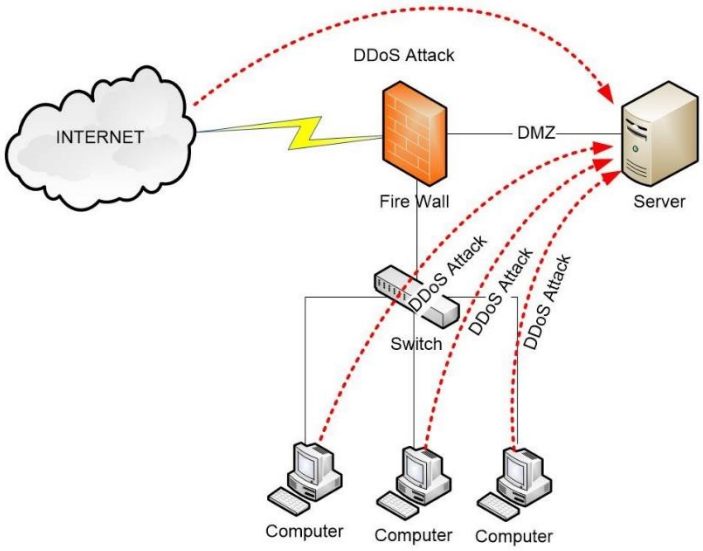
2.3.5.1 การโจมตีด้วยโครงข่าย (Network Base Attack) หรือการโจมตีด้วยขนาด (Volumetric Base Attack) ผู้โจมตีจะส่งข้อมูลที่มีปริมาณที่สูงมากเข้าไปที่เป้าหมายเพื่อทำให้การรับและส่งข้อมูลเต็มระบบ (Congestion) จนไม่สามารถติดต่อสื่อสารกับผู้ใช้งานทั่วไปได้โดยเทคนิคที่อยู่ในประเภทนี้คือ SYN Flood, UDP Flood, Ping of Death, Reflection และ Amplification เป็นต้น ซึ่งการโจมตีประเภทนี้จะพบบ่อยมากและป้องกันได้ยาก

2.3.5.2 การโจมตีด้วยแอพพลิเคชั่น (Application Base Attack) จะส่งข้อมูลที่อยู่ในชั้น Application Layer ของมาตรฐานการสื่อสารคอมพิวเตอร์ระบบเปิด (Open System Interconnection: OSI) เพื่อมุ่งเน้นไปให้แอพพลิเคชั่นหยุดทำงาน ซึ่งการโจมตีชนิดนี้จะอยู่ในระดับที่สูงกว่าการโจมตีด้วยโครงข่าย จะต้องอาศัยความเชี่ยวชาญของผู้ดูแลระบบเพื่อคัดกรอง แยกแยะข้อมูลที่ส่งเข้ามายังระบบ โดยเทคนิคที่อยู่ในประเภทนี้เช่น HTTP Flood, SSL Floodและ Slowloris และยังสามารถโจมตีผ่านทางช่องโหว่ของระบบได้ด้วย

2.3.6 โทรจันฮอร์ส (Trojan Horse) โปรแกรมคอมพิวเตอร์ที่แฝงโค้ดสำหรับการใช้ประโยชน์หรือทำลายระบบที่รันโดย โปรแกรมนี้ส่วนใหญ่จะถูกแนบมากับ E-mail ทั่วไปที่แฝงส่วนที่เป็นอันตรายต่อระบบเมื่อรันโปรแกรมนี้

2.4 รูปแบบการโจมตีแบบ DDoS

การโจมตีโดยปฏิเสธการให้บริการแบบกระจาย DDoS เป็นการโจมตีแบบหลายเส้นทางดังภาพที่ 2.4



ภาพประกอบที่ 2.4 การโจมตีแบบ DDoS จากระบบเครือข่ายภายในและภายนอกองค์กร

จึงมีผลที่รวดเร็วและอันตรามากกว่าใช้เส้นทางเดียวในการโจมตีแบบ DoS โดยใช้ช่องโหว่ของอุปกรณ์เพื่อทำให้อุปกรณ์หยุดทำงานหรือใช้การส่งข้อมูลจำนวนมากๆส่งผลโดยตรงต่อประสิทธิภาพในการให้บริการ ทำให้ประสิทธิภาพลดลง หรือทำให้ระบบคอมพิวเตอร์หยุดทำงาน เนื่องจากการโจมตีจะทำการเพิ่มปริมาณการเชื่อมต่อจำนวนมากทำให้ปริมาณการใช้แบนด์วิดธ์และการประมวลผลมีปริมานสูงขึ้นมากผิดปกติ ผู้ที่โจมตีแบบ DDoS มักจะนำเครื่องมือที่จะใช้ในการโจมตีไปติดตั้งบนคอมพิวเตอร์ที่ถูกเจาะไว้แล้ว คอมพิวเตอร์ที่ได้รับเครื่องมือนี้เข้าไปจะเรียกว่าซอมบี้ซึ่งเมื่อมีจำนวนพอสมควรก็จะระดมส่งข้อมูลในรูปแบบที่ควบคุมได้โดยผู้ควบคุมการโจมตีไปยังเหยื่อหรือเป้าหมายที่ต้องการซึ่งการโจมตีรูปแบบนี้มักจะก่อให้เกิดการใช้แบนด์วิดธ์อย่างเต็มที่จนผู้อื่นไม่สามารถใช้งานได้ตามปกติหรือทำให้ระบบที่ถูกโจมตีไม่มีทรัพยากรเหลือพอที่จะให้บริการผู้ใช้ธรรมดาได้รูปแบบการโจมตีที่นิยมใช้กันก็มีอย่าง SYN flood, UDP flood, ICMP flood, Smur และ Fraggle เป็นต้น ซึ่งมีรายละเอียดดังนี้

2.4.1 การโจมตีแบบ SYN Flood เป็นการโจมตีโดยการส่ง แพ็คเก็ต TCP ที่ตั้งค่า SYN บิตไว้ไปยังเป้าหมาย เสมือนกับการเริ่มต้นร้องขอการติดต่อแบบ TCP ตามปกติ ผู้โจมตีสามารถปลอมไอพีของ Source Address ได้เครื่องที่เป็นเป้าหมายก็จะตอบสนองโดยการส่ง SYN-ACK กลับมายัง Source IP Address ที่ระบุไว้ ซึ่งผู้โจมตีจะควบคุมเครื่องที่ถูกระบุใน Source IP Address ไม่ให้ส่งข้อมูลตอบกลับ ทำให้เกิดการโจมตีไม่ได้ปิดการเชื่อมต่อที่ได้เปิดไว้ (Half Open) ขึ้นที่เครื่องเป้าหมายหากมีการส่ง SYN flood จำนวนมาก ก็จะทำให้คิวของการให้บริการของเครื่องเป้าหมายเต็มทำให้ไม่สามารถให้บริการตามปกติได้นอกจากนี้ SYN flood ที่ส่งไปจำนวนมากยังอาจจะทำให้เกิดการใช้แบนด์วิดธ์อย่างเต็มที่อีกด้วย

2.4.2 การโจมตีแบบ ICMP Flood เป็นการส่งแพ็คเก็ต ICMP ขนาดใหญ่จำนวนมากไปยังเป้าหมายทำให้เกิดการใช้งานแบนด์วิดธ์เต็มที่

2.4.3 การโจมตีแบบ UDP Flood เป็นการส่งแพ็คเก็ต UDP จำนวนมากไปยังเป้าหมายซึ่งทำให้เกิดการใช้แบนด์วิดธ์อย่างเต็มที่และหรือทำให้ทรัพยากรของเป้าหมายถูกใช้ไปจนหมดโดยจะส่ง UDP packet ไปยัง port ที่กำหนดไว้ เช่น 53 (DNS)

2.4.4 การโจมตีแบบ Teardrop โดยปกติเราเตอร์จะไม่ยอมให้แพ็กเก็ตขนาดใหญ่ผ่านได้จะต้องทำ Fragment เสียก่อนจึงจะยอมให้ผ่านได้และเมื่อผ่านไปแล้วเครื่องของผู้รับปลายทางจะนำแพ็กเก็ตที่ถูกแบ่งออกเป็นชิ้นส่วนด้วยวิธีการ Fragment มารวมเข้าด้วยกันเป็นแพ็กเก็ตที่สมบูรณ์ การที่สามารถนำมารวมกันได้นี้จะต้องอาศัยค่า Offset ที่ปรากฏอยู่ในแพ็กเก็ตแรกและแพ็กเก็ตต่อไป สำหรับการโจมตีแบบ Teardrop ผู้โจมตีจะส่งค่า Offset ในแพ็กเก็ตที่สองและต่อไปที่จะทำให้เครื่องรับปลายทางเกิดความสับสนหากระบบปฏิบัติการไม่สามารถรับมือกับปัญหานี้ก็จะทำให้ระบบหยุดการทำงานในทันที

2.4.5 การโจมตีแบบ Land Attack ลักษณะการโจมตีประเภทนี้เป็นการส่ง SYN ไปที่เครื่องเป้าหมายเพื่อขอการเชื่อมต่อซึ่งเครื่องที่เป็นเป้าหมายจะต้องตอบรับคำขอการเชื่อมต่อด้วย SYN ACK ไปที่เครื่องคอมพิวเตอร์ต้นทางเสมอแต่เนื่องจากว่า IP Address ของเครื่องต้นทางกับเครื่องที่เป็นเป้าหมายนี้มี IP Address เดียวกันโดยการใช้วิธีการสร้าง IP Address ซึ่งโปรโตคอลของเครื่องเป้าหมายไม่สามารถแยกแยะได้ว่า IP Address ที่เข้ามาเป็นเครื่องปัจจุบันหรือไม่ ก็จะทำการตอบสนองด้วย SYN ACK ออกไปหากแอดเดรสที่ขอเชื่อมต่อเข้ามาเป็นแอดเดรสเดียวกับเครื่องเป้าหมายผลก็คือ SYN ACK นี้จะย้อนเข้าหาตนเองและเช่นกันที่การปล่อย SYN ACK แต่ละครั้งจะต้องมีการปันส่วนของหน่วยความจำเพื่อการนี้จำนวนหนึ่งซึ่งหากผู้โจมตีส่งคำขอเชื่อมต่อออกมาอย่างต่อเนื่องก็จะเกิดปัญหาการจัดสรรหน่วยความจำ

2.4.6 การโจมตีแบบ Smurf ผู้โจมตีจะส่ง ICMP Echo Request ไปยัง Broadcast Address ในเครือข่ายที่เป็นตัวกลาง โดยปลอม Source IP Address เป็น IP Address ของระบบที่ต้องการโจมตีซึ่งจะทำให้เครือข่ายที่เป็นตัวกลางส่ง ICMP Echo Reply กลับไปยัง IP Address ของเป้าหมายทันทีซึ่งทำให้มีการใช้งานแบนด์วิดธ์อย่างเต็มที่

2.5 การแปลงชุดข้อมูล (Data Transformation)

ก่อนที่จะนำข้อมูลส่งเข้าสู่การสร้างแบบจำลองต้องดำเนินการแปลงข้อมูลให้อยู่ในรูปแบบที่ตรงตามข้อจำกัดของแบบจำลองที่เลือกใช้ การแปลงคาขอมูลดวยวิธีการนอมัลไลซ (Normalization) เปนการลดขอบเขตของขอมูลให้น้อยลงโดยจะอยู่ในชวง 0 ถึง 1 ซึ่งทำให้แบบจำลองสามารถนำไปใช้ประมวลผลได้โดยมีวิธีการดังนี้

2.5.1 การแปลงคาข่อมูลในลักษณะเปนเชิงเส่นให้อยู่ในช่วงระยะสูงสุดต่ำสุด (Min-Max Normalization) ดังสมการ

$$\mathbf{V} = \frac{x\text{-}min_{(x)}}{max_{(x)} - min_{(x)}} \qquad (2\text{-}1)$$

2.5.2 *การปรับการกระจายของข้อมูล (Z-Score) โดยปรับข้อมูลให้มีค่าเท่ากับ 0 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 1 ดังสมการ*

$$\mathbf{V} = \frac{x\text{-}mean_{X}}{stand\_dev_{X}} \qquad (2\text{-}2)$$

2.5.3 การแปลงข้อมูลเดิมให้เป็นเลขทศนิยม (Decimal Scaling) ดังสมการ

$$\mathbf{V} = \frac{V}{10^{j}} \qquad (2\text{-}3)$$

2.5.4 วิธีการเปลี่ยนค่าข้อมูลที่นำเข้าโดย (Sigmoidal Normalization) มีช่วงระยะข้อมูล คือ -1 ถึง +1 ในการใช้ฟังชั่น Sigmoi มีสูตรดังนี้

$$\mathbf{V} = \frac{1 - e^{-a}}{1 + e^{-a}} \qquad (2\text{-}4)$$

2.6 การคัดเลือกคุณลักษณะพิเศษ (Feature selection)

การจำแนกประเภทข้อมูลพบว่าคุณลักษณะพิเศษมีจำนวนมากและมีหลายคุณลักษณะพิเศษที่มีความสำคัญในการแบ่งแยกคลาส (class) ออกเป็นเชิงบวกหรือเชิงลบได้ ดังนั้นจึงจำเป็นต้องทำการคัดเลือกคุณลักษณะพิเศษที่สำคัญมาใช้งาน ขั้นตอนนี้เรียกว่าการคัดเลือกคุณลักษณะพิเศษ (feature selection) ซึ่งสามารถแบ่งได้เป็น 2 กลุ่มใหญ่ดังนี้

2.6.1 การคัดเลือกคุณลักษณะพิเศษโดยใช้การคำนวณหาค่าน้ำหนัก (Filter approach) เป็นวิธีการหาค่าความสัมพันธ์ระหว่างแต่ละคุณลักษณะพิเศษและคลาส ทำการเลือกคุณลักษณะพิเศษโดยเรียงลำดับตามค่าน้ำหนักที่คำนวณได้ เลือกคุณลักษณะพิเศษที่มีค่าน้ำหนักมากกว่าที่ต้องการมาใช้วิธีการนี้ต่างจากวิธีการ Wrapper จะไม่มีการสร้างแบบจำลองเพื่อคัดเลือกคุณลักษณะ

พิเศษ เทคนิคในการคำนวณค่าน้ำหนักของคุณลักษณะพิเศษ เช่น Information Gain, Chi-Square และCorrelation

2.6.2 การคัดเลือกคุณลักษณะพิเศษด้วยการสร้างแบบจำลอง (Wrapper approach) เป็นวิธีการสร้างแบบจำลองขึ้นมาจากกลุ่มของคุณลักษณะพิเศษที่กำหนดไว้และวัดประสิทธิภาพการทำงานของแบบจำลอง และเลือกกลุ่มของคุณลักษณะพิเศษที่ทำให้แบบจำลองมีประสิทธิภาพมากที่สุดมาใช้งาน เช่น แบบจำลองที่ให้ค่าความถูกต้องมากที่สุด การคัดเลือกคุณลักษณะพิเศษด้วยวิธีการนี้แบ่งย่อยได้เป็น 2 แบบคือ

2.6.2.1 Forward Selection เป็นการสร้างแบบจำลองโดยการเพิ่มคุณลักษณะพิเศษทีละ 1 คุณลักษณะพิเศษ ถ้าคุณลักษณะพิเศษที่ใส่เพิ่มให้ประสิทธิภาพที่ดีจะเก็บไว้และเลือกคุณลักษณะพิเศษอื่นๆ มาเพิ่มต่อไปจนประสิทธิภาพของแบบจำลองไม่ได้ดีขึ้นก็จะหยุดทำงาน

2.6.2.2 Backward Elimination เป็นการสร้างแบบจำลองที่เริ่มจากการใช้คุณลักษณะพิเศษทั้งหมดก่อนและตัดคุณลักษณะพิเศษที่ไม่สำคัญทิ้งไปทีละคุณลักษณะพิเศษถ้าประสิทธิภาพดีขึ้นก็ตัดคุณลักษณะพิเศษอื่นๆ ต่อไป

2.7 การค้นหาพารามิเตอร์

การค้นหาพารามิเตอร์ที่ดีที่สุดในการสร้าบแบบจำลองเพื่อให้ได้ประสิทธิภาพของแบบจำลองที่ดีที่สุดจากจำนวนพารามิเตอร์ที่สามารถกำหนดได้หลายค่าและแต่ละเทคนิคมีค่าพารามิเตอร์ที่แตกต่างกันทำให้ขั้นตอนการค้นหาค่าพารามิเตอร์ที่ดีที่สุดมีความสำคัญจำเป็นต้องมีการเลือกค่าพารามิเตอร์เพื่อทำให้แบบจำลองมีค่าความถูกต้อง (Accuracy) ที่ดีซึ่งตัวอย่างของพารามิเตอร์ได้แก่

- จำนวนชั้นของแบบจำลอง
- ชนิดของ Activation Function
- Learning Rate
- Loss Function

วิธีการในการค้นพารามิเตอร์หรือที่เราเรียกว่า Hyperparameter Optimization ยกตัวอย่างเช่น

2.7.1 Grid Search [8] เป็นวิธีการกำหนดค่าของพารามิเตอร์ที่ต้องการเป็นชุดใช้ในการสร้างแบบจำลองและดำเนินการสร้างแบบจำลองในทุกๆชุดของค่าพารามิเตอร์ที่เป็นไปได้ วิธีนี้จะใช้ได้ดีกับแบบจำลองขนาดเล็กและเมื่อนำมาใช้กับแบบจำลองขนาดใหญ่จะใช้เวลาในการค้นหาพารามิเตอร์มากขึ้นดังตัวอย่างเช่น

เทคนิคที่ 1 มีพารามิเตอร์จำนวน 3 พารามิเตอร์ กำหนดให้ค่าของ

A1 = 100, 300, 500, 800, 1000

A2 = gini, entropy

A3 = True, False

จะทำการสร้างแบบจำลองจากทุกๆค่าพารามิเตอร์ที่กำหนดไว้อย่างเช่น

รอบที่ 1 จะทำการสร้างแบบจำลอง โดยกำหนดให้ พารามิเตอร์ A1 = 100, A2 = gini และ A3 = True

รอบที่ 2 จะทำการสร้างแบบจำลอง โดยกำหนดให้ พารามิเตอร์ A1 = 300, A2 = gini และ A3 = True

รอบที่ 3 จะทำการสร้างแบบจำลอง โดยกำหนดให้ พารามิเตอร์ A1 = 500, A2 = gini และ A3 = True

จะทำการสร้างแบบจำลองจนครบทุกค่าพารามิเตอร์ที่กำหนดจากเทคนิคที่ 1 ค่าพารามิเตอร์ A1 มีจำนวน 5 ค่า A2 มีจำนวน 2 ค่าและ A3 มีจำนวน 2 ค่า ทำให้จำนวนในการสร้างแบบจำลองทั้งหมดจะเท่ากับ A1 x A2 x A3 หรือ 5 x 2 x 2 เท่ากับ 20 แบบจำลอง และนำแบบจำลองทั้งหมดมาทำการเรียงลำดับแบบจำลองที่ให้ค่าความถูกต้องสูงที่สุดและนำค่าพารามิเตอร์ดังกล่าวใช้ในการสร้างแบบจำลองต่อไป

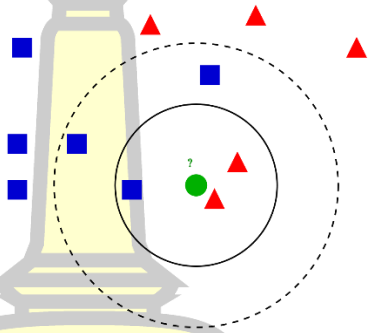
2.7.2 Random search [9] เป็นวิธีการสุ่มค่าพารามิเตอร์เฉพาะบางค่าจากพารามิเตอร์ที่กำหนดการสุ่มเพื่อที่จะให้ค่าพารามิเตอร์ที่สุ่มออกมาครอบคลุมพารามิเตอร์ทั้งหมด ซึ่งวิธีการสุ่มแบบนี้ยกตัวอย่างเช่น Sabol Sequence และ Hammersley Set แต่วิธีนี้ยังถือว่าไม่เหมาะกับการรันบนแบบจำลองขนาดใหญ่

2.8 เทคนิคการเรียนรู้ของเครื่องจักร (Machine Learning)

การเรียนรู้ของเครื่องจักร คือ การทำให้เครื่องคอมพิวเตอร์สามารถเรียนรู้งาน จากตัวอย่างหรือเรียนรู้จากประสบการณ์ที่เคยเกิดขึ้นในงานวิจัยนี้เลือกใช้วิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) เพื่อจำแนกข้อมูลโดยอาศัยคุณลักษณะของข้อมูลตัวอย่าง เพื่อที่จะสามารถจำแนกได้จะต้องทราบล่วงหน้าว่ามีกลุ่มอะไรบ้าง (Predefined Categories) และมีข้อมูลของหน่วยตัวอย่างในแต่ละกลุ่มก่อน ระบบจากข้อมูลนำเข้า (Input Data) เพื่อสร้างตัวแบบที่สอดคล้องกับข้อมูลที่สุด โดยแสดงความสัมพันธ์ระหว่างตัวแปรทำนายและกลุ่มของข้อมูลนำเข้า โดยแบ่งข้อมูลเป็นสองส่วน คือ

ข้อมูลเรียนรู้ (Training) และข้อมูลทดสอบ (Testing) ข้อมูลเรียนรู้จะถูกนำไปใช้ในการสร้างแบบจำลองเพื่อจำแนกข้อมูลและตัวแบบจำลองนี้จะถูกนำไปใช้จำแนกประเภทข้อมูลของหน่วยตัวอย่างในข้อมูลทดสอบ แบบจำลองที่ได้รับการทดสอบว่ามีความถูกต้องสูงก็จะถูกนำไปใช้ทำนายข้อมูลใหม่ โดยมี เทคนิดดังนี้

2.8.1 วิธีการเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors Algorithm: KNN) วิธีสำหรับค้นหาเพื่อนบ้านที่ใกล้ที่สุด [10] เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องจักรที่ไม่ต้องสร้างแบบจำลองเพื่อนำมาใช้สำหรับจำแนกข้อมูลโดยข้อมูลทั้งหมดจะถูกนำมาคำนวนหาค่าระยะทาง (Distance) ดังภาะประกอบที่ 2.5 แสดงที่ข้อมูลใหม่จะถูกมาคำนวนหาระยะทาง



ภาพประกอบที่ 2.5 แสดงข้อมูลของเทคนิค KNN [11]

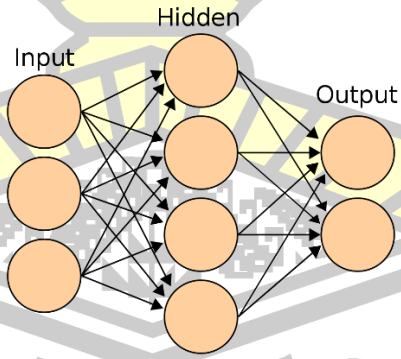
เพื่อนำมาเปรียบเทียบระยะทางระหว่างข้อมูลที่ต้องการนำมาจัดหมวดหมู่ y และข้อมูลทั้งหมด Xi ดังนั้นข้อมูลที่มีระยะทางน้องที่สุดจำนวน k ข้อมูลจึงถูกนำมาพิจารณาและในข้อมูลทั้งสิ้นจำนวน k ข้อมูลนั้นหากมีสมาชิกของกลุ่ม Ci ใดมากที่สุด ข้อมูลที่ต้องการนำมาจัดหมวกหมู่ y จะถูกกำหนดให้อยู่ในกลุ่มนั้น หากกำหนดให้ k = 3 ดังนั้นค่าของระยะทางที่น้อยที่สุด 3 ค่าจะถูกนำมาพิจารณาหากข้อมูลที่มีค่าระยะทางทั้งสิ้น 3 จำนวน อยู่ในกลุ่มดังต่อไปนี้ d = (C1,C2,C3) ข้อมูลที่ต้องการนำมาจัดหมวดหมู่จะถูกกำหนดให้เป็น C1 เนื่องจากมีจำนวนที่ปรากกฏมากที่สุด การคำนวนหาค่าระยะทางด้วยการหาระยะห่างแบบยุคลิต (Euclidean Distance) สามารถคำนวนได้ดังสมการที่ 2-5

$$d(x,y)\sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (2\text{-}5)$$

โดยที่ N คือ จำนวนของคุณลักษณะพิเศษ (Dimensions) ของข้อมูล
x,y คือข้อมูลที่อยู่ในชุดข้อมูลเรียนรู้ และ y คือข้อมูลที่ต้องการนำมาจำแนก

จากนั้นนำค่าระยะทาง d(x,y) ที่ได้ทั้งหมดเพื่อ Majority Vote ดั้งนั้นกลุ่ม ck ของข้อมูลที่ปรากฏบ่อยที่สุดจะถูกกำหนดให้เป็นผลลัพธ์ของ KNN

2.8.2 โครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron: MLP) โครงข่ายประสาทเทียม [12,13] แบบจำลองทางคณิตศาสตร์หรือแบบจำลองทางคอมพิวเตอร์สำหรับประมวลผลด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (Bioelectric network) ในสมองซึ่งประกอบด้วยเซลล์ประสาท (Neurons) และจุดประสานประสาท (Synapses) เกิดจากการเชื่อมต่อระหว่างเซลล์ประสาทเป็นเครือข่ายที่ทำงานร่วมกัน ซึ่งมีความสามารถในการเรียนรู้คล้ายคลึงกับสมองของมนุษย์เป็นที่นิยมใช้ในสาขาปัญญาประดิษฐ์ (Artificial Intelligent) โดยอาศัยข้อมูลเพื่อสร้างแบบจำลองเพื่อการพยากรณ์เหตุการในอนาคต โครงข่ายประสาทเทียมจะลดจำนวนของการทำนายที่ผิดพลาดให้ต่ำที่สุด พื้นฐานโครงข่ายประสาทเทียมประกอบด้วย 3 ส่วนหลักดังภาพที่ 2.6 แสดงถึงชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อนตัว (Hidden Layer) และชั้นข้อมูลออก (Output Layer)

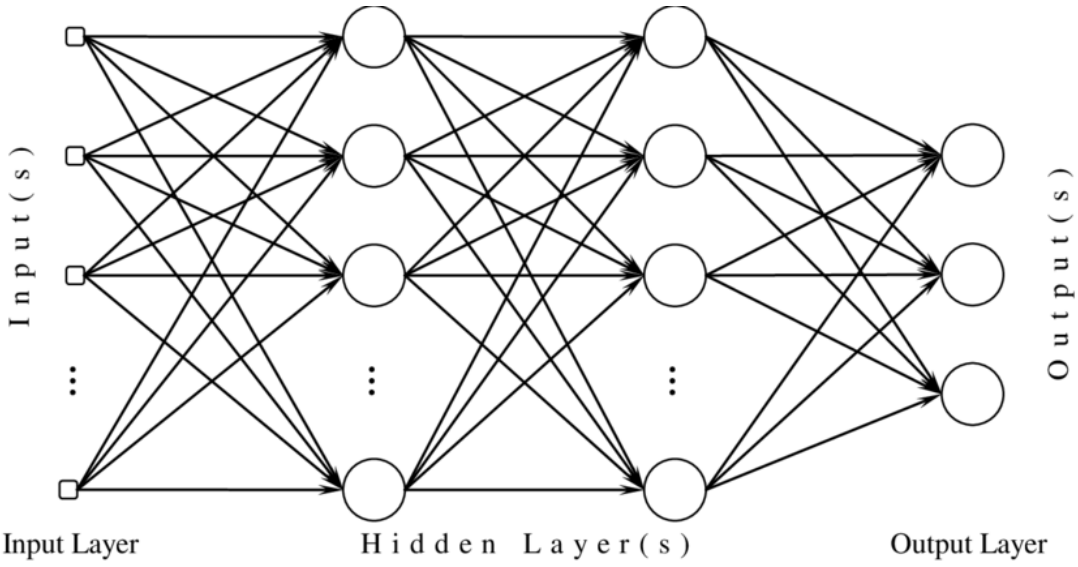


ภาพประกอบที่ 2.6 พื้นฐานโครงข่ายประสาทเทียม [14]

ชั้นข้อมูลเข้า (Input layer) จะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งในชั้นซ่อน (Hidden) จะมีฟังก์ชันสำหรับคำนวณเมื่อได้รับสัญญาณ (Output) จากโหนดในชั้นก่อนหน้านี้เรียกว่า Activation Function โดยในแต่ละชั้นไม่จำเป็นต้องเป็นฟังก์ชันเดียวกันก็ได้ชั้น โดยแปลงข้อมูลที่เข้ามาให้

สามารถแยกแยะความแตกต่างโดยใช้เส้นตรงเส้นเดียว (Linearly Separable) และก่อนที่ข้อมูลจะถูกส่งไปถึงชั้นข้อมูลออกในบางครั้งอาจจำเป็นต้องใช้ชั้นซ่อนตัวมากกว่า 1 ชั้นในการแปลงข้อมูลให้อยู่ในรูปเส้นตรงเส้นเดียว จนกระทั่งถึงชั้นข้อมูลออก ชั้นซ่อนตัวมีนิวรอนตั้งแต่หนึ่งตัวขึ้นไปโดยนิวรอนจะทำหน้าที่ในการประมวลผลลัพธ์ที่และส่งออกทางข้อมูลออก โดยที่นิวรอน 1 ตัว สามารถอ่านข้อมูลได้มากกว่า 1 ค่า หากมีข้อมูลเพียงค่าเดียว เรียกว่า เพอร์เซ็ปตรอนแบบ Single Input และข้อมูลที่มีค่ามากกว่า 1 ค่าเรียก เพอร์เซ็ปตรอนแบบ Miltiple Input ภายในนิวรอนแต่ละตัวประกอบไปด้วย Summation Function และ Activation Function โครงสร้างพื้นฐานของโครงข่ายประสาทเทียม เพอร์เซ็ปตรอนหลายชั้น Multilayer Perceptron (MLP) [13] เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้นใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการเรียนรู้แบบมีผู้สอนและใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) สำหรับการเรียนรู้กระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass)

สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้าและจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อและค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย สัญญาณที่มีโครงข่ายประสาทเทียมแบบ MLP มี 2 ประเภท คือ Function Signal เป็นสัญญาณเข้าที่มาจากโหนดในชั้นก่อนหน้าและจะส่งผ่านไปข้างหน้าจากโหนดหนึ่งไปสู่อีกโหนดหนึ่งส่วนสัญญาณผิดพลาดเป็นสัญญาณย้อนกลับที่เกิดขึ้นที่โหนดในชั้นข้อมูลออกของโครงข่ายประสาทเทียมและถูกส่งผ่านย้อนกลับจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง

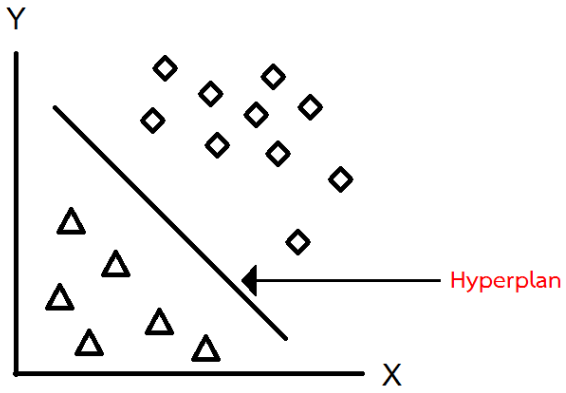


ภาพประกอบที่ 2.7 แสดงชั้นเพอร์เซ็ปตรอนหลายชั้น [15]

หลักการทำงานของ MLP คือในแต่ละชั้นของชั้นซ่อนตัวจะมีฟังก์ชันสำหรับคำนวณเมื่อได้รับสัญญาณข้อมูลออกจากโหนดในชั้นก่อนหน้านี้เรียกว่า Activation Function โดยในแต่ละชั้นไม่จำเป็นต้องเป็นฟังก์ชันเดียวกันก็ได้ชั้นซ่อนตัวจะแปลงข้อมูลที่เข้ามาในชั้นให้สามารถแยกแยะความแตกต่างโดยใช้เส้นตรงเส้นเดียวและก่อนที่ข้อมูลจะถูกส่งไปถึงชั้นข้อมูลออกในบางครั้งอาจจำเป็นต้องใช้ชั้นซ่อนตัวมากกว่า 1 ชั้นในการแปลงข้อมูลให้อยู่ในรูปเส้นตรงเส้นเดียวในการคำนวณหาข้อมูลออกในปัญหาการจำแนกทำได้โดยการใส่ข้อมูลเข้าไปในโครงข่ายประสาทเทียมที่เราได้ทำการหาไว้แล้วจากนั้นให้ทำการเปรียบเทียบค่าของข้อมูลขาออกในชั้นข้อมูลออกและให้ทำการเลือกค่าของข้อมูลที่มีค่าสูงกว่าและทำการรับค่าของพยากรณ์ที่ตรงกับ Neuron ที่เลือกและให้นำค่าของมาเปรียบเทียบกับค่าที่ยอมรับได้หากค่าอยู่ในช่วงที่รับได้ก็ให้ทำการรับข้อมูลชุดถัดไปแต่หากค่าของมากกว่าค่าที่ยอมรับได้ให้ทำการปรับค่าน้ำหนักตามขั้นตอนที่ได้กล่าวไว้ข้างต้นเมื่อทำการปรับน้ำหนักเรียบร้อยแล้วให้ทำการรับข้อมูลชุดถัดไปและทำตามขั้นตอนซ้ำอีกรอบจนกระทั่งถึงข้อมูลชุดสุดท้ายจะนับเป็น 1 รอบของการคำนวณ (Epoch) จากนั้นจะทำการหาค่าผิดพลาดรวมเฉลี่ยจากค่าเฉลี่ยของที่ได้เก็บค่าเอาไว้เพื่อใช้ในการตรวจสอบว่าค่าโดยเฉลี่ยในการจำแนกนั้นมีค่าน้อยกว่าค่าผิดพลาดที่ยอมรับได้หรือไม่ถ้าใช่แสดงว่าโครงข่ายประสาทเทียมที่สร้างขึ้นนั้นสามารถให้ผลลัพธ์ที่ถูกต้องของทุกๆข้อมูลแล้วจึงทำการจบการเรียนรู้ได้แต่ถ้าไม่ใช่ให้กลับไปทำตามขั้นตอนแรกโดยเริ่มรับข้อมูลชุดที่ 1 ใหม่ สามารถนำมาใช้ในการจำแนกข้อมูลได้เช่น การตรวจจับการบุกรุก [16]

2.8.3 วิธีการของซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) วิธีการจำแนกข้อมูลที่มีประสิทธิภาพและมีความถูกต้องแม่นยำ นักวิจัยจึงนำไปช่วยแก้ไขปัญหาทางด้าน

Classifcation อัลกอริทึ่มของ SVM [17] ทำหน้าที่ในการหาเส้นแบ่ง (Hyperplane) ที่เหมาะสมที่สุด (Optimal) ดังภาพประกอบที่ 2.8



ภาพประกอบที่ 2.8 แสดงตำแหน่งข้อมูลสองกลุ่มที่อยู่ในฟีเจอร์สเปซ (Feature Space)

ที่มีระยะห่าง (Margin) ระหว่างข้อมูล (Training Points) กับเส้นแบ่งมากที่สุดโดยที่ Training Points ที่เข้าใกล้เส้นเส้นแบ่งจะถูกเรียกว่า Support Vectors แรกเริ่ม SVM ถูกออกแบบมาเพื่อใช้จัดหมวดหมู่ข้อมูลเฉพาะที่เป็น 2 กลุ่มโดยใช้สมการเส้นตรง (Linear Model) ในการแบ่งกลุ่มข้อมูลโดยฟังก์ชันที่ใช้สำหรับตัดสินใจในการแบ่งข้อมูลคือ

$$f(x) = sign(w^T x + b) \tag{2-6}$$

โดย w คือ ค่าเฉลี่ยน้ำหนัก (Weight Vector)
b คือ ค่าไบแอส (Bias)

โดยจะสร้างเส้นแบ่งที่เป็นเส้นตรงเพื่อให้ทราบว่าเส้นตรงที่แบ่งสองกลุ่มออกจากกันนั้นเส้นตรงใดเป็นเส้นที่ดีที่สุดสามารถเขียนเป็นสมการได้ดั่งนี้

$$X = ((X1, Y1)), \ldots, ((Xi, Yi)) \tag{2-7}$$

โดยให้ X คือ ลักษณะเด่น

จากสมการ 2-7 เปนสมการเสนตรง เพื่อวิเคราะห์หาเสนตรงบนไฮเปอรเพลน ซึ่งแบงกลุ่มข้อมูลที่มีลักษณะเชิงเสนสองกลุมออกจากกันโดยมีการกำหนดกลุมของขอมูลทั้งสองฝงเปนเพียงสองคาซึ่งแทนดวยคา Y ขอมูลที่เปนตัวกำหนดความชันและระนาบที่เกิดขึ้นบนไฮเปอรเพลนเกิดจากคูขอมูล (w,b) และกำหนดสมการที่เปนตัวบงบอกขอมูลแตละกลุมวาอยูสวนไหนของเสนแบงไฮเปอร์เพลนแสดงได้ดังสมการที่ (2-8) และสมการที่ (2-9) เมื่อนำสมการเงื่อนไขทั้งหมดมาวิเคราะหเชิงเรขาคณิต โดยพิจารณาในกรณีที่ขอมูลถูกแบงกลุมได้สมบูรณตามเงื่อนไขซัพพอรตเวกเตอรแมชชีนในสมการที่ (2-10) และขอมูลใช้สอนให้ระบบเรียนรูต้องอยูในรูปแบบของเชิงเสนสามารถแสดงลักษณะการวางตัวของกลุมขอมูลได้ดังภาพประกอบที่ 2.9 เวกเตอรของขอมูลที่ถูกปอนเขาเพื่อให้การเรียนรูแทนด้วยสมการและขอมูลทั้งสองดานแบงเปนบวกและลบสถานะของขอมูลจึงแทนด้วย y ซึ่งมีสองคา คือ y=1 และ y= -1 แตทั้งนี้ก็ยังตัดสินไมได้วาเสนแบงนั้นควรจะเปนเสนใดจึงจะดีที่สุด วิธีการที่ใชในการหาเส้นแบงที่ดีที่สุดคือการเพิ่มเสนขอบให้กับเสนแบงทั้งสองข้างทำให้ได้เส้นใหมที่จะถือเปนเสนขอบของขอมูลแตละฝงอีกด้วยเสนขอบของเสนแบงนั้นจะเปนเสนที่สัมผัสกับคาขอมูลในฟเจอรสเปซที่ใกลที่สุดเสนขอบของทั้งสองเสนที่เพิ่มขึ้นมานี้ถูกแทนดวยสมการ "w" ^ "T" "x+b ≥ y ≥1" ถาอยูดาน y = 1 และ "w" ^"T" "x+b ≤ y ≤-1" ถาเส้นขอบของเสนแบงใดๆ ที่มีความกวางมากที่สุด แสดงให้เห็นวาขอมูลสองชุดมีการแยกกันชัดเจนมากที่สุดดังนั้นเสนแบงที่มีเสนขอบกวางที่สุดจึงเปนเสนแบงที่ดีที่สุดโดยเรียกเส้นประที่แบ่งข้อมูลทั้งสองซึ่งสามารถเขียนเปนสมการการคำนวณความกวางของเส้น โดยคำนวณจากสมการที่ (2-8) และ (2-9) เมื่อแทนค่าลงไปแลว

ให้ y =1 $w^T+b \geq y$ (2-8)

ให้ y =-1 $w^T+b \leq y$ (2-9)

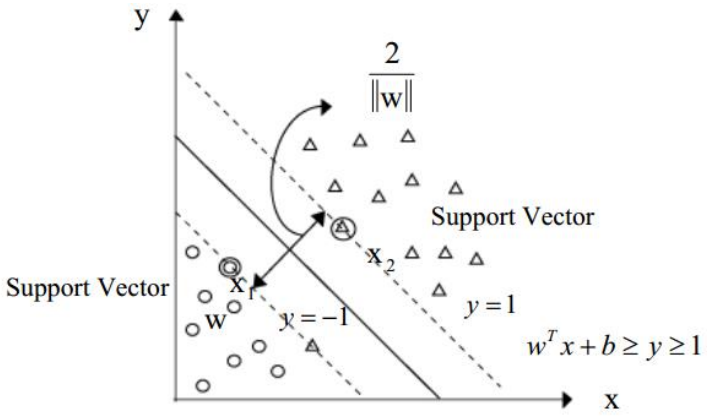
$y(w^T x+b)-1 \geq 0$ (2-10)

โดยให้ y คือ ค่ากลุ่มข้อมูล (1,-1)

w คือ ค่าความชัน

x คือ ค่าลักษณะเด่น

b คือ ค่าคงที่



ภาพประกอบที่ 2.9 การวางตัวของข้อมูลในลักษณะเชิงเส้น [18]

$$w^T x^+ + b = 1$$
$$w^T x^- + b = -1$$
$$w^T (x^+ + x^-) = 2$$
$$M = \left(\frac{w}{\|w\|} \right)^T (x^+ - x^-)$$
$$M = \frac{2}{\|w\|} \qquad (2\text{-}11)$$

ให้ M คือ ความกว้างของเส้นขอบ หลักจากที่ได้สมการที่ (2-10) และ สมการที่ (2-11) ของการหาเส้นแบ่งและค่าความกว้างตามลำดับ

2.9 การทดสอบประสิทธิภาพ

เทคนิค Cross-validation เป็นการทดสอบประสิทธิภาพของแบบจำลองเนื่องจากผลที่ได้มีความน่าเชื่อถือการวัดประสิทธิภาพด้วยวิธีนี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน [19,20] จะแสดงด้วยค่า k เช่น 5-fold คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากันหลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของแบบจำลองจนครบจำนวนที่แบ่งไว้

ตัวอย่าง การทดสอบด้วยวิธี 5-fold Cross-validation

รอบที่ 1 ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยข้อมูลส่วนที่ 1 ถึง 4 เป็นข้อมูลสำหรับเรียนรู้และข้อมูลส่วนที่ 5 เป็นข้อมูลทดสอบประสิทธิภาพ

1	2	3	4	5
Training	Training	Training	Training	Testing

รอบที่ 2 ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยข้อมูลส่วนที่ 1 ถึง3 และ 5 เป็นข้อมูลสำหรับเรียนรู้และข้อมูลส่วนที่ 4 เป็นข้อมูลทดสอบประสิทธิภาพ

1	2	3	4	5
Training	Training	Training	Testing	Training

รอบที่ 3 ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยข้อมูลที่ 1 ถึง 2 และ 4 ถึง 5 เป็นข้อมูลสำหรับเรียนรู้และข้อมูลส่วนที่ 3 เป็นข้อมูลทดสอบประสิทธิภาพ

1	2	3	4	5
Training	Training	Testing	Training	Training

รอบที่ 4 ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยข้อมูลที่ 1 และ 3 ถึง 5 เป็นข้อมูลสำหรับเรียนรู้และข้อมูลส่วนที่ 2 เป็นข้อมูลทดสอบประสิทธิภาพ

1	2	3	4	5
Training	Testing	Training	Training	Training

รอบที่ 5 ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยข้อมูลที่ 2 ถึง 5 เป็นข้อมูลสำหรับเรียนรู้และข้อมูลส่วนที่ 1 เป็นข้อมูลทดสอบประสิทธิภาพ

1	2	3	4	5
Testing	Training	Training	Training	Training

2.10 การประเมินผล

การประเมินผลลัพธ์การทำนาย (Confusion Matrix) เปรียบเทียบกับผลลัพธ์จริงที่กำหนดไว้ และ

True Positive (TP) คือ สิ่งที่ทำนายว่าจริงและบอกว่ามันจริง

True Negative (TN) คือ สิ่งที่ทำนายว่าไม่จริงและบอกว่ามันไม่จริง

False Positive (FP) คือ สิ่งที่ทำนายว่าจริงแต่บอกว่าไม่จริง

False Negative (FN) คือ สิ่งที่ทำนายว่าไม่จริงแต่บอกว่าจริง

Accuracy คือ ค่าความถูกต้องของทำนายหาได้จากสูตร

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (2\text{-}12)$$

Recall (True Positive Rate) คือ ค่าที่บอกว่าทำนายได้ว่าจริงเป็นอัตราส่วนเท่าไรของจริงทั้งหมด หาได้จากสูตร

$$Recall = TP/(TP+FN) \quad (2\text{-}13)$$

True Negative Rate (TNR) คือ ค่าที่บอกว่าโปรแกรมทำนายได้ว่าไม่จริง เป็นอัตราส่วนเท่าไรของจริงทั้งหมดหาได้จากสูตร

$$TNR = TN/(TN+FP) \quad (2\text{-}14)$$

False Positive Rate (TPR) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าจริง เป็นอันตราส่วนเท่าไรของไม่จริงทั้งหมดหาได้จากสูตร

$$TPR = FP/(TN+FP) \quad (2\text{-}15)$$

False Negative Rate (FNR) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าไม่จริง เป็นอันตราส่วนเท่าไรของจริงทั้งหมดหาได้จากสูตร

$$FNR = FN/(TP+FN) \quad (2\text{-}16)$$

Precision คือ ค่าที่บอกว่าโปรแกรมทำนายว่าจริงถูกต้องเท่าไรหาได้จากสูตร

$$\text{Precision} = TP/(TP+FP) \quad (2\text{-}17)$$

F-measure คือ การวัดค่า Precision และ Recall พร้อมกันของแบบจำลองโดยพิจารณาแยกทีละคลาส

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall}/\text{Precision} + \text{Recall} \quad (2\text{-}18)$$

2.11 งานวิจัยที่เกี่ยวข้อง (Related Work)

โดยมีเนื้อหาที่เกียวกับการตรวจสอบการโจมตีแบบ DDoS เพื่อจัดหมวดหมู่การใช้งานแบบปกติและถูกโจม ดังนี้

Perakvić และคณะ [21] นำเสนอวิธีการประยุกต์ใช้โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) เพื่อตรวจจับ (Detection) และการจำแนกการโจมตีแบบ DDoS โดยใช้ข้อมูลที่เผยแพร่ออนไลน์ทั้งสิ้นจำนวน 4 ชุด มีข้อมูลรวมกันทั้งสิ้น 4,986 ชุด โดยข้อมูลทั้งหมดแบ่งวิธีการบุกรุกออกเป็น 4 ประเภท คือ DNS DDoS attack, CharGen DDoS attack, UDP DDoS attack และ normal traffic โดยได้ทดสอบกับชั้นซ่อน (Hidden Layer) ตั้งแต่ 30, 35, 40, 45, 50 และ 55 จากการทดลองปรากฏว่า Hidden Layer จำนวน 50 Neurons ให้ผลการทดลอง 95.6% ซึ่งสูงที่สุด

ในงานวิจัยนี้ Hsieh and Chan [22] นำเสนอวิธีการตรวจจับการโจมตีแบบ DDoS โดยใช้โครงข่ายประสาทเทียม และใช้เฟรมเวิร์กของ Apache Spark ซึ่งใช้สำหรับจัดการข้อมูลขนาดใหญ่ (Big Data) และทำงานแบบคลัสเตอร์ (Cluster) โดยทำงานได้รวดเร็วกว่า Hadoop MapReduce ถึง 100 เท่า ในงานวิจัยนี้ได้ใช้ข้อมูลชุด ARPA 2000 LLDOS 1.0 โดยมีคุณลักษณะพิเศษ (Feature) ทั้งสิ้นจำนวน 7 คุณลักษณะ ได้แก่ จำนวนของแพ็กเกจ (Number of Packets) ขนาดโดยเฉลี่ยของแพ็กเกจ (Average of Packet Size) ค่าเฉลี่ยเวลา (Time Interval Variance) ความแตกต่างของขนาดของแพ็กเกจ (Packet Size Variance) จำนวนไบต์ (Number of Bytes) ความถี่ของข้อมูล (Packet Rate) และ ขนาดข้อมูล (Bit Rate) ซึ่งข้อมูลถูกแบ่งประเภทของการบุกรุกออกเป็น 2 ประเภท คือ Normal และ Attack โดยมีข้อมูล Normal ทั้งสิ้น 51,040 ชุด และข้อมูล Attack จำนวนทั้งสิ้น 74,480 ชุด ข้อมูลทั้งหมดถูกแบ่งออกเป็น 30 % สำหรับข้อมูลชุดเรียนรู้ และ 70% สำหรับข้อมูลชุดทดสอบ จากการทดลองพบว่าให้ความถูกต้อง 94%

Singh และ Tiwari [23] นำเสนอวิธีการตรวจจับการบุกรุก (Intrusion Detection) โดยใช้เทคนิค ID3 เพื่อลดจำนวนของคุณลักษณะพิเศษ (Reduced Features) จากข้อมูลชุด KDD ให้เหลือเพียง 18 Attribute โดยข้อมูลที่ใช้ในการทดสอบนี้แบ่งประเภทของการบุกรุกออกเป็น 4 ประเภท คือ Denial of Service (DoS), Remote to Local (R2l), User to Root (U2R) และ Probe โดยข้อมูลจำนวนทั้งสิ้น 26,167 ชุดถูกแบ่งออกเป็นสองส่วนเท่ากัน เพื่อใช้เป็นข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบโดยข้อมูลคุณลักษณะจะถูกนำไปเรียนรู้ด้วยเทคนิค K-Nearest Neighbor and Genetic Algorithm (KNNGA) เพื่อทำการจำแนกข้อมูลและได้นำไปเปรียบเทียบกับวิธี KNN และ SVM จากการทดลองพบว่า วิธีการที่นำเสนอในงานวิจัยให้ความถูกต้องสูงกว่า 98% ซึ่งสูงกว่าทั้งวิธี KNN และ SVM

Devaraju และ Ramakrishnan [24] นำเสนอวิธีโครงข่ายประสาทเทียมสำหรับจัดหมวดหมู่ (Neural Network Classifiers) ซึ่งประกอบด้วย 3 เทคนิคดังนี้ Feed Forward Neural Network (FFNN), Probabilistic Neural Network (PNN) และ Radial Basis Neural Network (RBNN) เพื่อทดสอบประสิทธิภาพของการจำแนกข้อมูลชนิดการโจมตีของระบบตรวจจับการบุกรุก (Intrusion Detection System) โดยทดสอบกับข้อมูลข้อมูล KDD ที่ประกอบด้วยคุณลักษณะจำนวน 41 คุณลักษณะ โดยมีชนิดของการโจมตีที่ต่างกัน 4 กลุ่ม คือ กลุ่มที่ 1 การโจมตีแบบ Denial of Service (DoS) ประกอบด้วย back, land, neptune, pod, smurf และ teardrop กลุ่มที่ 2 Remote-to-Local (R2L) ประกอบด้วย ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient และ warezmaster กลุ่มที่ 3 User-to-Root (U2R) ประกอบด้วย buffer_overflow, loadmodule, perl และ rootkit กลุ่มที่ 4 Probing ประกอบด้วย ipsweep, nmap, portsweep และ satan ข้อมูลแบ่งออกเป็น 7 กลุ่ม (Class) คือ normal class, Smurf class, Neptune class, Saint Class, Mail bomb class, Apache class และ Satan class ในการทดลองได้แบ่งข้อมูลออกเป็นข้อมูลชุดเรียนรู้ (Training Set) และข้อมูลชุดทดสอบ (Test Set) โดยข้อมูลชุดเรียนรู้และทดสอบจำนวนชุดละ 700 ข้อมูล ได้ถูกเลือกมาจากข้อมูลทั้งสิ้นจำนวน 7 Class โดยแต่ละ Class จะถูกเลือกมาจำนวน 100 ชุด จากการทดลองพบว่า PNN มีประสิทธิภาพดีที่สุด โดย PNN, FFNN และ RBNN มีประสิทธิภาพที่ 97.5%, 94.3% และ 65% ตามลำดับ

งานวิจัยของ Ingre B. and Yadav A. [25] ได้นำเสนอการวิเคราะห์ประสิทธิภาพชุดข้อมูล NSL-KDD โดยใช้เทคนิค ANN เพื่อประเมินประสิทธิภาพความแม่นยำในการจำแนกข้อมูล ทดลองกับชุดข้อมูล NSL-KDD มีข้อมูลสำหรับเรียนรู้ 125,973 ชุด และข้อมูลสำหรับประเมินประสิทธิภาพ 22,544 ชุด โดยมีคุณลักษณะพิเศษ 41 features และทำการลดขนาดคุณลักษณะพิเศษให้เหลือ 29 features ด้วยวิธีการ information gain, gain ratio และ correlation attribute algorithm มี 5 class ประกอบไปด้วย DOS, Probe, R2L, U2R และ Normal ส่งข้อมูลเข้าไปทำการจำแนกข้อมูล

ด้วยวิธีการ Levenberg-Marquardt (LM) และ BFGS quasi-Newton backpropagation จากผลการทดลองพบว่าเทคนิค LM มีชั้นซ่อน 21 ชั้น จำนวน 117 Epoch มีค่าความแม่นยำ 81.2% มีค่าความแม่นยำสูงกว่า BFGS ที่มีชั้นซ่อน 23 ชั้น 771 Epoch มีค่าความแม่นยำ 79.9%

งานวิจัยของ Pervez MS. and Farid DM. [26] ได้นำเสนอวิธีการเลือกคุณลักษณะและการจำแนกการบุกรุก โดยทดสอบกับชุดข้อมูล NSL-KDD ด้วยเทคนิค SVM ที่มีข้อมูลจำนวน 125,973 ชุด และมี 41 คุณลักษณะพิเศษ ส่งข้อมูลไปประเมินประสิทธิภาพด้วยวิธีการ cross-validation โดยกำหนดให้ K = 10 fold และทำการลดขนาดของคุณลักษณะพิเศษดังนี้ 41, 36, 29, 17, 14, 9, 6, 5, 4 และ 3 จากผลการทดลองพบว่าจำนวนคุณลักษณะพิเศษที่ให้ความแม่นยำสูงที่สุดคือ 41 คุณลักษณะพิเศษมีความแม่นยำ 99.01%

งานวิจัยของ Yusof ARA, Udzir NI, Selamat A, Hamdan H and Abdullah MT. [27] นำเสนอวิธีการเลือกคุณลักษณะพิเศษที่ปรับเปลี่ยนได้สำหรับการโจมตีแบบ DoS โดยใช้เทคนิคในการเลือกคือ Information Gain, Gain Ratio, Chi-squared และ Correlated features selection (CFS) นำไปทดสอบกับชุดข้อมูล NSL-KDD ที่มีข้อมูลสำหรับเรียนรู้จำนวน 125,973 แถว และข้อมูลสำหรับทดสอบจำนวน 22,544 แถว ส่งไปจำแนกข้อมูลด้วยการเรียนรู้ของเครื่องจักรด้วยเทคนิค Extreme Learning Machine จากผลการทดลองพบว่ามีจำนวนของคุณลักษณะพิเศษมีจำนวนแตกต่างกันตามลำดับดังนี้ 16, 15, 17 และ 14 และมีค่าความแม่นยำตามลำดับดังนี้ 83.9%, 92.1%, 96.2% และ 88.1% วิธีการ Chi-squared ที่มี 17 คุณลักษณะพิเศษ มีประสิทธิภาพค่าความแม่นยำที่ 96.2% สูงที่สุด

งานวิจัยของ Meena G และ Choudhary RR [28] นำเสนอวิธีการตรวจจับการบุกรุก โดยใช้เทคนิค J48 Graft และ NAIVE BAYES และนำไปทดสอบกับชุดข้อมูล 2 ชุด คือ KDD ที่มีข้อมูล 4,900,000 แถว และ NSL KDD จากผลการทดลองโดยเปรียบเทียบประสิทธิภาพค่าความแม่นยำในการจำแนกข้อมูลพบว่าเทคนิค J48 Graft ให้ค่าความแม่นยำสูงที่สุดมีความแม่ยำ 99.435% และเทคนิค NAIVE BAYES น้อยกว่าโดยมีความแม่นยำ 92.715%

บทที่ 3
วิธีการดำเนินการวิจัย

ในบทนี้กล่าวถึงขั้นตอนการดำเนินการิจัยเพื่อให้ได้ตามวัตถุประสงค์ที่ตั้งไว้ ประกอบด้วย ศึกษาและวิเคราะห์ข้อมูลชุดข้อมูล (Dataset) ของ KDD และ NSL KDD การเรียนรู้ด้วยเทคนิค MLP, SVM และ KNN และวิธีการหาค่าค่าพารามิเตอร์ที่ดีที่สุดและขั้นตอนในการทดสอบประสิทธิภาพด้วยวิธีการ Cross Validation โดยมีรายละเอียดดังนี้

3.1 ศึกษาและวิเคราะห์ข้อมูล

3.2 กระบวนการก่อนการสร้างแบบจำลอง

3.3 ขั้นตอนการทดลอง

3.4 การหาค่าพารามิเตอร์

3.5 การทดสอบประสิทธิภาพ

3.1 ศึกษาและวิเคราะห์ข้อมูล

งานวิจัยนี้เป็นงานวิจัยทางด้านความปลอดภัยของระบบเครือข่ายคอมพิวเตอร์เพื่อจำแนกการโจมตีแบบ DDoS โดยใช้ชุดข้อมูล KDD [29,30] มีข้อมูลจำนวน 4,898,431 ชุด และมีคุณลักษณะพิเศษจำนวน 41 คุณลักษณะพิเศษ และชุดข้อมูล NSL KDD [31] มีข้อมูลจำนวน 125,373 ชุด มีคุณลักษณะพิเศษจำนวน 41 คุณลักษณะพิเศษ ข้อมูลดังกล่าวเป็นข้อมูลปกติและการบุกรุกระบบเครือข่ายคอมพิวเตอร์ จากตารางที่ 3.1 แสดง คุณลักษณะพิเศษของชุดข้อมูลดั่งกล่าว

ตารางที่ 3.1 แสดงคุณลักษณะพิเศษของชุดข่อมูล KDD และ NSL KDD

ลำดับที่	คุณลักษณะพิเศษ	ชนิด
1	duration	continuous
2	protocol_type	Nominal
3	service	Nominal
4	flag	Nominal
5	src_bytes	continuous
6	dst_bytes	continuous
7	land	Nominal

ตารางที่ 3.1 แสดงคุณลักษณะพิเศษของชุดข้อมูล KDD และ NSL KDD

ลำดับที่	คุณลักษณะพิเศษ	ชนิด
8	wrong_fragment	continuous
9	urgent	continuous
10	hot	continuous
11	num_failed_logins	continuous
12	logged_in	Nominal
13	num_compromised	continuous
14	root_shell	continuous
15	su_attempted	continuous
16	num_root	continuous
17	num_file_creations	continuous
18	num_shells	continuous
19	num_access_files	continuous
20	num_outbound_cmds	continuous
21	is_host_login	Nominal
22	is_guest_login	Nominal
23	count	continuous
24	srv_count	continuous
25	serror_rate	continuous
26	srv_serror_rate	continuous
27	rerror_rate	continuous
28	srv_rerror_rate	continuous
29	same_srv_rate	continuous
30	diff_srv_rate:	continuous
31	srv_diff_host_rate	continuous
32	dst_host_count	continuous
33	dst_host_srv_count	continuous
34	dst_host_same_srv_rate	continuous
35	dst_host_diff_srv_rate	continuous
36	dst_host_same_src_port_rate	continuous

ตารางที่ 3.1 แสดงคุณลักษณะพิเศษของชุดข้อมูล KDD และ NSL KDD

ลำดับที่	คุณลักษณะพิเศษ	ชนิด
37	dst_host_srv_diff_host_rate	continuous
38	dst_host_serror_rate:	continuous
39	dst_host_srv_serror_rate	continuous
40	dst_host_rerror_rate	continuous
41	dst_host_srv_rerror_rate	continuous

ชุดข้อมูลจะมีลักษณะที่เปนแบบค่าต่อเนื่องเชิงปริมาณ (Continuous) และค่าเดียวข้อมูลเชิงคุณภาพ (Nominal) โดยคุณลักษณะพิเศษจะมีความหมายที่ตางกันเชน คุณลักษณะพิเศษ

- ระยะเวลาในการเชื่อมตอ (Duration)
- ชนิดโปรโตคอล (Protocol_type) เช่น TCP, UDP และ ICMP
- ประเภทของบริการ (Service) เชน http, ftp หรือ Telnet

ภายในชุดข้อ KDD และ NSL KDD มูลจะประกอบด้วย Normal Class และลักษณะการโจมตีแบ่งได้ 4 Class คือ

1) การโจมตีแบบ Denial of Service (DOS) การโจมตีที่มีการส่งแพ็กเก็ตจำนวนมากไปยังเป้าหมายทำให้ไม่สามารถให้บริการได้
2) การโจมตีแบบ Remote to Local (R2L) การพยายามเข้าถึงระบบของเป้าหมายโดยไม่ได้รับอนุญาตในการเข้าถึง
3) การโจมตีแบบ User to Root (U2R) การพยายามใช้งานสิ่งที่ไม่ได้รับอนุญาตในการเข้าถึงยังสิทธ์ Supper-user (root)
4) การโจมตีแบบ Probing เป็นลักษณะการตรวจสอบข้อมูลบนเครือข่าย โดยหาช่องโหว่ของเป้าหมายเพื่อใช้เป็นข้อมูลในการโจมตีรูปแบบการโจมตีที่นิยมใช้กัน เช่น Nmap หรือ Port Scanning เป็นต้น

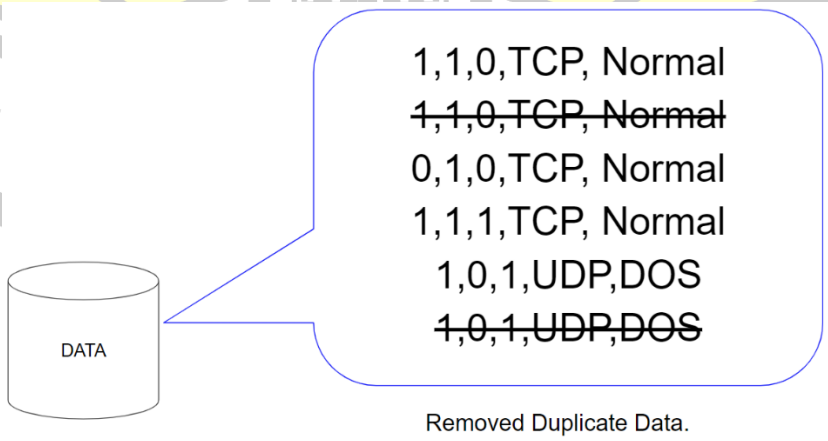
ลักษณะของการจัดเรียงรูปแบบภายในชุดข้อมูลจะแสดงดังตารางที่ 3.2

ตารางที่ 3.2 TABLE 3 ONE ROW OF A DATA

Data Sample	Lable
0,tcp,http,SF,219,1098,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00, 1.00,0.00,0.00,7,255,1.00,0.00,0.14,0.05,0.00,0.01,0.00,0.00	Normal
0,tcp,telnet,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,1,0.50,1.00,0.00,0.00,0.50, 1.00,0.00,1,2,1.00,0.00,1.00,1.00,1.00,0.50,0.00,0.00	Neptune
0,icmp,ecr_i,SF,1032,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,263,263,0.00,0.00,0.00, 0.00,1.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00	Smurf
0,tcp,http,SF,54540,8314,0,0,0,2,0,1,1,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.0 0,1.00,0.00,0.00,251,251,1.00,0.00,0.00,0.00,0.00,0.00,0.02,0.02	Back
0,udp,private,SF,28,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,21,21,0.00,0.00,0.00,0.00 ,1.00,0.00,0.00,255,21,0.08,0.02,0.08,0.00,0.00,0.00,0.00,0.00	Teardrop
0,icmp,tim_i,SF,564,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1. 00,0.00,0.00,1,1,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00	Pod
0,tcp,finger,S0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00, 0.00,0.00,1,8,1.00,0.00,1.00,0.38,1.00,0.12,0.00,0.00	Land

3.2 กระบวนการก่อนการสร้างแบบจำลอง

3.2.1 วิธีกำจัดข้อมูลซ้ำ (Data Deduplication) ในกระบวนการก่อนการสร้างแบบจำลอง ผู้วิจัยได้ดำเนินการตรวจสอบข้อมูลและนำข้อมูลที่ซ้ำออกจากชุดข้อมูลดังภาพประกอบที่ 3.1 โดยพิจารณาจากแถวข้อมูลที่เหมือนกันทั้ง 41 คุณลักษณะพิเศษ และ Class ให้เหลือข้อมูลที่ไม่ซ้ำซ้อน



ภาพประกอบที่ 3.1 แสดงลักษณะข้อมูลซ้ำและถูกกำจัด

3.2.2 การแปลงชุดข้อมูล (Data Transformation) ให้อยู่ในรูปแบบที่สามารถส่งข้อมูลเพื่อสร้างแบบจำลองได้ ภายในชุดข้อมูล KDD และ NSL KDD มีชนิดที่เป็นค่าเดียวและค่าต่อเนื่อง ดำเนินการแปลงคาตัวอักษรให้อยู่ในรูปแบบของตัวเลข (Numeric) โดยจะแทนคาด้วยตัวเลข (Label Encoding) ตั้งแต 1 ถึง N ให้ N คือจำนวนของรูปแบบตัวอักษรทั้งหมดในคุณลักษณะ

ตารางที่ 3.3 คุณลักษณะที่ชื่อว่า Protocol_type จะถูกแทนค่าด้วย

Protocol_type	New Value
icmp	1
tcp	2
udp	3

ตารางที่ 3.4 คุณลักษณะที่ชื่อว่า flag จะถูกแทนค่าด้วย

Flag	New Value	Flag	New Value
REJ	1	S	7
S0	2	RSTR	8
S1	3	RSTO	9
S2	4	RSTOS0	10
S3	5	OTH	11
SF	6		

ตารางที่ 3.5 คุณลักษณะที่ชื่อว่า Service จะถูกแทนค่าด้วย

Service	New Value	Service	New Value	Service	New Value	Service	New Value
time	1	rje	19	harvest	37	telnet	55
echo	2	ssh	20	discard	38	shell	56
ldap	3	efs	21	netstat	39	imap4	57
link	4	ftp	22	courier	40	eco_i	58
http	5	netbios_dgm	23	pm_dump	41	ecr_i	59
smtp	6	netbios_ssn	24	printer	42	red_i	60
uucp	7	netbios_ns	25	private	43	pop_2	61

ตารางที่ 3.5 คุณลักษณะที่ชื่อว่า Service จะถูกแทนค่าด้วย

Service	New Value	Service	New Value	Service	New Value	Service	New Value
auth	8	remote_job	26	sql_net	44	pop_3	62
nnsp	9	http_8001	27	tftp_u	45	login	63
nntp	10	hostnames	28	sunrpc	46	tim_i	64
name	11	uucp_path	29	Z39_50	47	urh_i	65
exec	12	http_2784	30	gopher	48	urp_i	66
aol	13	iso_tsap	31	domain	49	ntp_u	67
IRC	14	csnet_ns	32	finger	50	vmnet	68
X11	15	domain_u	33	klogin	51	other	69
bgp	16	ftp_data	34	kshell	52	whois	70
ctf	17	http_443	35	supdup	53		
mtp	18	daytime	36	systat	54		

การสร้างแบบจำลองเพื่อเปรียบเทียบเทคนิคที่มีประสิทธิภาพในการจำแนกข้อมูลโดยใช้ชุดข้อมูล KDD และ NSL KDD โดยในการทดลองได้เลือกใช้ข้อมูลชุด Normal และข้อมูลการโจมตีแบบ DDoS ที่มีจำนวน 6 คลาส เท่านั้น ดังแสดงในตารางที่ 3.2 และนำชุดข้อมูลที่ได้ดำเนินการกำจัดข้อมูลซ้ำและเปลี่ยนแปลงข้อมูล ทำการแปลงชุดข้อมูล KDD และ NSL KDD ออกเป็น 3 ชุดข้อมูล ดังนี้ ชุดที่ 1 จำนวน 2 คลาส คือ ข้อมูล Normal และข้อมูลการโจมตีแบบ DDoS ที่เปลี่ยนแปลงจากทั้ง 6 คลาส ให้เป็น 1 คลาส คือ Attack จำนวนข้อมูลทั้งหมดใน ตารางที่ 3.6

ตารางที่ 3.6 Number of 2 Class DATA SET

Classes	KDD	NSL KDD
Normal	279,388	19,420
Attack	247,267	12,354
Total	526,655	31,774

ชุดที่ 2 จำนวน 6 class คือ ชุดข้อมูลที่ได้นำ Normal ออกไปเหลือไว้เพียงข้อมูลการโจมตีแบบ DDoS คือ Neptune, Pod, Smurf, Teardrop, Land และ Back จำนวนข้อมูลทั้งหมดใน ตารางที่ 3.7

ตารางที่ 3.7 Number of 6 Class DATA SET

Classes	KDD 1999	NSL KDD
Neptune	242,149	9,314
Smurf	3,007	1,330
Back	968	718
Teardrop	918	892
Pod	206	82
Land	19	18
Total	247,267	12,354

ชุดที่ 3 มี 7 class เป็นชุดข้อมูลที่ประกอบได้ด้วย Neptune, Pod, Smurf, Teardrop, Land, Back และ Normal จำนวนข้อมูลทั้งหมด 526,655 แถว ดังตารางที่ 3.8

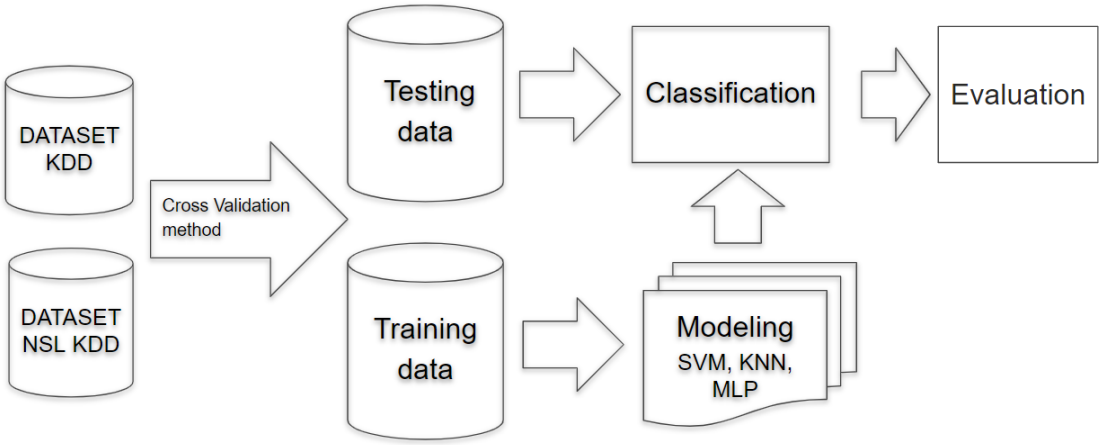
ตารางที่ 3.8 Number of 7 Class DATA SET

Classes	KDD 1999	NSL KDD
Neptune	242,149	9,314
Smurf	3,007	1,330
Back	968	718
Teardrop	918	892
Pod	206	82
Land	19	18
Normal	279,388	19,420
Total	526,655	31,774

และแปลงชุดข้อมูลให้อยู่ในรูปแบบที่สามารถส่งข้อมูลเพื่อสร้างแบบจำลอง

3.3 ขั้นตอนการทดลอง

วิธีการสร้างแบบจำลองเลือกใช้เทคนิควิธีการเรียนรู้แบบมีผู้สอนโดยเลือกใช้ 3 เทคนิคและนำมาเปรียบเทียบประสิทธิภาพของแต่ละเทคเนคโดยเลือกใช้เทคนิค MLP, KNN และ SVM ขั้นตอนก่อนสร้างแบบจำลองดำเนินการหาค่าพารามิเตอร์ที่ดีที่สุดของทุกเทคนิคโดยใช้วิธีการ Grid Search จากค่าพารามิเตอร์ที่กำหนดไว้โดยทำการทดลองทุกค่าพารามิเตอร์และทำการจัดลำดับค่าพารามิเตอร์ที่ดีที่สุดและนำค่าพารามิเตอร์ดังกล่าวมาใช้ในการสร้างแบบจำลอง เพื่อให้ได้ค่าค่าพารามิเตอร์ที่ให้ผลการประเมินผลลัพธ์การทำนายสูงที่สุดและนำไปทดลองตามขั้นตอนที่แสดงดังภาพประกอบที่ 3.2



ภาพประกอบที่ 3.2 *ขั้นตอนการสร้างแบบจำลอง*

3.4 การหาค่าพารามิเตอร์

วิธีการหาค่าพารามิเตอร์ของแบบจำลองของเทคนิค MLP, KNN และ SVM โดยใช้เทคนิคใช้วิธีการโดยเลือกใช้วิธีการ Grid Search โดยกำหนดชุดของพารามิเตอร์ดังนี้

เทคนิค MLP กำหนดค่าพารามิเตอร์ดังนี้

hidden_layer_sizes = [5, 10, 20, 30, 40, 50, 100, 150, 200, 300, 400, 500, 1000]

เทคนิค KNN กำหนดค่าพารามิเตอร์ดังนี้

n_neighbors = [1, 3, 5, 10]

weights = [uniform, distance]

algorithm = [ball_tree, kd_tree, brute]

leaf_size = [10, 50, 100, 200, 300]

เทคนิค SVM กำหนดค่าพารามิเตอร์ดังนี้

ชุดที่ 1 kernel = [rbf]
gamma = [1e-3, 1e-4]
C = [1, 10, 100, 1000]

ชุดที่ 2 kernel = [linear]
C = [1, 10, 100, 1000]

โดยทำการหาค่าพารามิเตอร์จากชุดข้อมูล KDD และ NSL KDD จำนวนข้อมูล 10% ถูกนำมาใช้ในการหาค่าพารามิเตอร์และดำเนินการสร้างแบบจำลองทุกๆค่าพารามิเตอร์และทำการวัดประสิทธิภาพของแบบจำลองด้วยวิธีการ Cross Validation และทำการเรียงลำดับค่าพารามิเตอร์ที่ให้ผลความถูกต้องมากที่สุดถึงน้อยที่สุด

3.5 การทดสอบประสิทธิภาพ

ในการทดสอบประสิทธิภาพของวิธีการที่นำเสนอนั้นจะใช้วิธีการทดสอบโดยใช้เทคนิค K-Fold Cross Validation ชุดข้อมูลจะถูกแบ่งออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบโดยกำหนดให้ K = 2, 5 และ 10 ทำการทดลองทั้งหมด 10 รอบและนำผลการทดลองนำมาหาค่าเฉลี่ยและประเมินผลลัพธ์การทำนายเพื่อเปรียบเทียบประสิทธิภาพแต่ละเทคนิคในในการจำแนกการโจมตีแบบ DDoS โดยทำการทดสอบเปรียบเทียบประสิทธิภาพดังนี้

1) Accuracy คือ ค่าความถูกต้องของทำนายหาได้จากสูตร
2) Recall (True Positive Rate) คือ ค่าที่บอกว่าทำนายได้ว่าจริงเป็นอัตราส่วนเท่าไรของจริงทั้งหมด
3) True Negative Rate (TNR) คือ ค่าที่บอกว่าโปรแกรมทำนายได้ว่าไม่จริง เป็นอัตราส่วนเท่าไรของจริงทั้งหมด
4) False Positive Rate (TPR) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าจริง เป็นอันตราส่วนเท่าไรของไม่จริงทั้งหมด
5) False Negative Rate (FNR) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าไม่จริง เป็นอันตราส่วนเท่าไรของจริงทั้งหมด

บทที่ 4
ผลการศึกษา

ผลการดำเนินการวิจัยในงานนี้ผู้วิจัยได้ใช้เทคนิค SVM, KNN และ MLP ใช้วิธีการ Grid Search ในการหาค่าพารามิเตอร์ในการสร้างแบบจำลองเพื่อจำแนกข้อมูล การวัดประสิทธิภาพแบบจำลองได้ทดลองแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบด้วย Cross validation และวัดประสิทธิภาพการจำแนกข้อมูลของแบบจำลองที่สร้างจากแต่ละเทคนิคโดยวัดค่า Accuracy และแสดงตัวอย่างการแปลผลการจำแนกข้อมูล ซึ่งในบทนี้จะแสดงรายละเอียดของแต่ละส่วนในงานวิจัย ดังนี้

4.1 เครื่องมือและข้อมูลที่ใช้ในการทดลอง

4.2 ผลการวิเคราะห์ชุดข้อมูล

4.3 ผลการหาค่าพารามิเตอร์

4.4 ผลการทดลองประสิทธิภาพของอักอริทึม

4.1 เครื่องมือและข้อมูลที่ใช้ในการทดลอง

เครื่องมือที่ใช้ในการทดลองได้แก่ เครื่องคอมพิวเตอร์ที่ทำงานบนระบบปฏิบัติการ Virtual Machine จำลองการทำงานของคอมพิวเตอร์เสมือนมีคอมพิวเตอร์ 2 เครื่องซ้อนกันอยู่ในคอมพิวเตอร์เพียงเครื่องเดียว โดยกำหนดให้มีหน่วยประมวลผลกลาง Intel Xeon 12 Core หน่วยความจำ 16 กิกะไบต์ ระบบปฏิบัติการ Windows 10 และภาษา Python ในการทดลอง

4.2 ผลการวิเคราะห์ชุดข้อมูล

4.2.1 ผลการกำจัดข้อมูลซ้ำจาก Benchmark Dataset จำนวน 2 ชุดได้แก่ KDD ที่มีจำนวนข้อมูลทั้งสิ้น 4,898,431 แถว และเมื่อทำการกำจัดข้อมูลซ้ำทำให้ขนาดของข้อมูลลดลงเหลือ 529,655 แถว และชุดข้อมูล NSL KDD ที่มีจำนวน 125,373 และเมื่อทำการกำจัดข้อมูลซ้ำทำให้ขนาดของข้อมูลลดลงเหลือ 12,354 แถว

4.2.2 ผลการการแปลงข้อมูลเพื่อให้สามารถนำไปใช้งานกับการเรียนรู้ของเครื่องจักรได้ทำการแทนค่าของคุณลักษณะพิเศษ Protocol_type, flag และ Service ดังภาพประกอบที่ 4.1 แสดงตัวอย่างของข้อมูล 20 ที่ยังไม่ถูกแทนค่า

0	tcp	uucp_path	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	262	11	0	0	1	1	0.04	0.06	0	255	11	0.04	0.06	0	0	0	0	1	1	neptune
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	129	10	1	1	0	0	0.08	0.06	0	255	10	0.04	0.05	0	0	0.79	1	0	0	neptune
0	tcp	uucp	RSTO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	115	2	0	0	1	1	0.02	0.07	0	255	2	0.01	0.06	0	0	0	0	1	1	neptune
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	112	9	0	0	1	1	0.08	0.06	0	255	9	0.04	0.05	0	0	0	0	1	1	neptune
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	190	15	1	1	0	0	0.08	0.08	0	255	15	0.06	0.08	0	0	1	1	0	0	neptune
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	20	0	0	1	1	0.13	0.05	0	255	20	0.08	0.05	0	0	0	0	1	1	neptune
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	220	5	0	0	1	1	0.02	0.07	0	255	5	0.02	0.07	0	0	0	0	1	1	neptune
0	tcp	private	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	210	2	1	1	0	0	0.01	0.07	0	255	2	0.01	0.07	0	0	1	1	0	0	neptune
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	249	8	0	0	1	1	0.03	0.07	0	255	8	0.03	0.07	0	0	0	0	1	1	neptune
0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	291	17	0	0	1	1	0.06	0.05	0	255	17	0.07	0.06	0	0	0	0	1	1	neptune
0	tcp	ftp_data	SF	203	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	125	0.49	0.02	0.49	0	0	0	0	0	normal
0	tcp	http	SF	238	17015	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4	5	0	0	0	0	1	0	0.4	4	255	1	0	0.25	0.02	0	0	0	0	normal
0	ud	other	SF	146	105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	4	0.02	0.06	0.09	0	0	0	0.01	0	normal
0	tcp	http	SF	293	682	0	0	0	0	0	1	0	0	0	0	0	0	0	0	25	26	0	0	0	0	1	0	0.0	255	255	1	0	0	0	0	0	0	0	normal
0	tcp	http	SF	258	31760	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	49	255	1	0	0.02	0.02	0	0	0	0	normal
0	ud	ntp_u	SF	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	5	0.02	0.01	0.02	0	0	0	0	0	normal
4	tcp	http	SF	218	11135	0	0	0	0	0	1	0	0	0	0	0	0	0	0	6	9	0	0	0	0	1	0	0.3	255	255	1	0	0	0	0	0	0	0	normal
0	tcp	smtp	SF	285	334	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	3	0	0	0	0	1	0	1	39	130	0.72	0.13	0.03	0.02	0	0	0	0.01	normal
0	tcp	bgp	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	213	19	0	0	1	1	0.09	0.07	0	255	19	0.07	0.07	0	0	0	0	1	1	neptune
0	tcp	smtp	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	207	12	1	1	0	0	0.06	0.06	0	255	12	0.05	0.06	0	0	1	1	0	0	neptune

ภาพประกอบที่ 4.1 แสดงตัวอย่างข้อมูลที่ไม่ถูกแทนค่า

และแทนค่าของคุณลักษณะพิเศษ Protocol_type, flag และ Service ข้อมูลจะถูกแทนค่าดังภาพประกอบที่ 4.2

0	2	29	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	262	11	0	0	1	1	0.04	0.06	0	255	11	0.04	0.06	0	0	0	0	1	1	neptune
0	2	43	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	129	10	1	1	0	0	0.08	0.06	0	255	10	0.04	0.05	0	0	0.79	1	0	0	neptune
0	2	7	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	115	2	0	0	1	1	0.02	0.07	0	255	2	0.01	0.06	0	0	0	0	1	1	neptune
0	2	43	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	112	9	0	0	1	1	0.08	0.06	0	255	9	0.04	0.05	0	0	0	0	1	1	neptune
0	2	43	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	190	15	1	1	0	0	0.08	0.08	0	255	15	0.06	0.08	0	0	1	1	0	0	neptune
0	2	43	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	20	0	0	1	1	0.13	0.05	0	255	20	0.08	0.05	0	0	0	0	1	1	neptune
0	2	43	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	220	5	0	0	1	1	0.02	0.07	0	255	5	0.02	0.07	0	0	0	0	1	1	neptune
0	2	43	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	210	2	1	1	0	0	0.01	0.07	0	255	2	0.01	0.07	0	0	1	1	0	0	neptune
0	2	43	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	249	8	0	0	1	1	0.03	0.07	0	255	8	0.03	0.07	0	0	0	0	1	1	neptune
0	2	43	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	291	17	0	0	1	1	0.06	0.05	0	255	17	0.07	0.06	0	0	0	0	1	1	neptune
0	2	34	6	203	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	125	0.49	0.02	0.49	0	0	0	0	0	normal
0	2	5	6	238	17015	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4	5	0	0	0	0	1	0	0.4	4	255	1	0	0.25	0.02	0	0	0	0	normal
0	3	69	6	146	105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	4	0.02	0.06	0.09	0	0	0	0.01	0	normal
0	2	5	6	293	682	0	0	0	0	0	1	0	0	0	0	0	0	0	0	25	26	0	0	0	0	1	0	0.08	255	255	1	0	0	0	0	0	0	0	normal
0	2	5	6	258	31760	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	49	255	1	0	0.02	0.02	0	0	0	0	normal
0	3	67	6	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	5	0.02	0.01	0.02	0	0	0	0	0	normal
4	3	5	6	218	11135	0	0	0	0	0	1	0	0	0	0	0	0	0	0	6	9	0	0	0	0	1	0	0.33	255	255	1	0	0	0	0	0	0	0	normal
0	2	6	6	2851	334	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	3	0	0	0	0	1	0	1	39	130	0.72	0.13	0.03	0.02	0	0	0	0.01	normal
0	2	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	213	19	0	0	1	1	0.09	0.07	0	255	19	0.07	0.07	0	0	0	0	1	1	neptune
0	2	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	207	12	1	1	0	0	0.06	0.06	0	255	12	0.05	0.06	0	0	1	1	0	0	neptune

ภาพประกอบที่ 4.2 แสดงตัวอย่างข้อมูลที่ถูกแทนค่า

4.3 ผลการหาค่าพารามิเตอร์

นำข้อมูลจำนวน 10% ของชุดข้อมูล KDD 2 Class ดำเนินการค่าพารามิเตอร์ที่ดีที่สุดเมื่อนำไปหาค่าพารามิเตอร์จากเทคนิค SVM ได้ค่าพารามิเตอร์ดังตารางที่ 4.1

นำข้อมูลจำนวน 10% ของชุดข้อมูล KDD 6 Class ดำเนินการค่าพารามิเตอร์ที่ดีที่สุดเมื่อนำไปหาค่าพารามิเตอร์จากเทคนิค SVM ได้ค่าพารามิเตอร์ดังตารางที่ 4.2

ตารางที่ 4.1 ผลการหาค่าพารามิเตอร์จากชุดข้อมูล KDD 6 คราส

ลำดับที่	ค่าพารามิเตอร์	ความถูกต้อง
1	{'kernel': 'linear', 'C': 1}	0.967
2	{'kernel': 'linear', 'C': 10}	0.967
3	{'kernel': 'linear', 'C': 100}	0.967
4	{'kernel': 'linear', 'C': 1000}	0.967
5	{'kernel': 'rbf', 'C': 10, 'gamma': 0.001}	0.958
6	{'kernel': 'rbf', 'C': 100, 'gamma': 0.001}	0.958
7	{'kernel': 'rbf', 'C': 100, 'gamma': 0.0001}	0.958
8	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.001}	0.958
9	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.0001}	0.958
10	{'kernel': 'rbf', 'C': 1, 'gamma': 0.001}	0.918
11	{'kernel': 'rbf', 'C': 10, 'gamma': 0.0001}	0.918
12	{'kernel': 'rbf', 'C': 1, 'gamma': 0.0001}	0.866

นำข้อมูลจำนวน 10% ของชุดข้อมูล KDD 7 Class ดำเนินการค่าพารามิเตอร์ที่ดีที่สุดเมื่อนำไปหาค่าพารามิเตอร์จากเทคนิค SVM ได้ค่าพารามิเตอร์ดังตารางที่ 4.3

นำข้อมูลจำนวน 10% ของชุดข้อมูล NSL KDD 2 Class ดำเนินการค่าพารามิเตอร์ที่ดีที่สุดเมื่อนำไปหาค่าพารามิเตอร์จากเทคนิค SVM ได้ค่าพารามิเตอร์ดังตารางที่ 4.4

นำข้อมูลจำนวน 10% ของชุดข้อมูล NSL KDD 6 Class ดำเนินการค่าพารามิเตอร์ที่ดีที่สุดเมื่อนำไปหาค่าพารามิเตอร์จากเทคนิค SVM ได้ค่าพารามิเตอร์ดังตารางที่ 4.5

นำข้อมูลจำนวน 10% ของชุดข้อมูล NSL KDD 7 Class ดำเนินการค่าพารามิเตอร์ที่ดีที่สุดเมื่อนำไปหาค่าพารามิเตอร์จากเทคนิค SVM ได้ค่าพารามิเตอร์ดังตารางที่ 4.6

ตารางที่ 4.1 ผลการหาค่าพารามิเตอร์จากชุดข้อมูล KDD 7 คราส

ลำดับที่	ค่าพารามิเตอร์	ความถูกต้อง
6	{'kernel': 'rbf', 'C': 100, 'gamma': 0.0001}	0.971
8	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.0001}	0.971
9	{'kernel': 'linear', 'C': 1}	0.968
10	{'kernel': 'linear', 'C': 10}	0.967
11	{'kernel': 'linear', 'C': 100}	0.967
12	{'kernel': 'linear', 'C': 1000}	0.967
7	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.001}	0.966
3	{'kernel': 'rbf', 'C': 10, 'gamma': 0.001}	0.965
5	{'kernel': 'rbf', 'C': 100, 'gamma': 0.001}	0.965
4	{'kernel': 'rbf', 'C': 10, 'gamma': 0.0001}	0.964
1	{'kernel': 'rbf', 'C': 1, 'gamma': 0.001}	0.959
2	{'kernel': 'rbf', 'C': 1, 'gamma': 0.0001}	0.646

ตารางที่ 4.1 ผลการหาค่าพารามิเตอร์จากชุดข้อมูล KDD 2 คราส

ลำดับที่	ค่าพารามิเตอร์	ความถูกต้อง
1	{'kernel': 'rbf', 'C': 100, 'gamma': 0.001}	1
2	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.001}	1
3	{'kernel': 'rbf', 'C': 10, 'gamma': 0.001}	0.999
4	{'kernel': 'rbf', 'C': 100, 'gamma': **0.0001}**	0.999
5	{'kernel': 'rbf', 'C': 1000, 'gamma': **0.0001}**	0.999
6	{'kernel': 'linear', 'C': 1}	0.999
7	{'kernel': 'linear', 'C': 10}	0.999
8	{'kernel': 'linear', 'C': 100}	0.999
9	{'kernel': 'linear', 'C': 1000}	0.999
10	{'kernel': 'rbf', 'C': 1, 'gamma': 0.001}	0.998
11	{'kernel': 'rbf', 'C': 10, 'gamma': 0.0001}	0.998
12	{'kernel': 'rbf', 'C': 1, 'gamma': 0.0001}	0.996

ตารางที่ 4.1 ผลการหาค่าพารามิเตอร์จากชุดข้อมูล NSL KDD 6 คราส

ลำดับที่	ค่าพารามิเตอร์	ความถูกต้อง
1	{'kernel': 'rbf', 'C': 100, 'gamma': 0.001}	1
2	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.001}	1
3	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.0001}	1
4	{'kernel': 'linear', 'C': 1}	1
5	{'kernel': 'linear', 'C': 10}	1
6	{'kernel': 'linear', 'C': 100}	1
7	{'kernel': 'linear', 'C': 1000}	1
8	{'kernel': 'rbf', 'C': 10, 'gamma': 0.001}	0.955
9	{'kernel': 'rbf', 'C': 100, 'gamma': 0.0001}	0.955
10	{'kernel': 'rbf', 'C': 10, 'gamma': 0.0001}	0.95
11	{'kernel': 'rbf', 'C': 1, 'gamma': 0.001}	0.907
12	{'kernel': 'rbf', 'C': 1, 'gamma': 0.0001}	0.155

ตารางที่ 4.1 ผลการหาค่าพารามิเตอร์จากชุดข้อมูล NSL KDD 7 คราส

ลำดับที่	ค่าพารามิเตอร์	ความถูกต้อง
1	{'kernel': 'linear', 'C': 100}	0.882
2	{'kernel': 'linear', 'C': 1000}	0.878
3	{'kernel': 'linear', 'C': 10}	0.873
4	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.001}	0.856
5	{'kernel': 'rbf', 'C': 100, 'gamma': 0.001}	0.809
6	{'kernel': 'linear', 'C': 1}	0.798
7	{'kernel': 'rbf', 'C': 1000, 'gamma': 0.0001}	0.797
8	{'kernel': 'rbf', 'C': 10, 'gamma': 0.001}	0.652
9	{'kernel': 'rbf', 'C': 100, 'gamma': 0.0001}	0.643
10	{'kernel': 'rbf', 'C': 1, 'gamma': 0.001}	0.641
11	{'kernel': 'rbf', 'C': 10, 'gamma': 0.0001}	0.641
12	{'kernel': 'rbf', 'C': 1, 'gamma': 0.0001}	0.318

นำข้อมูลจำนวน 10% ของชุดข้อมูล KDD 2 Class, KDD 6 Class, KDD 7 Class, NSL KDD 2 Class, NSL KDD 6 Class และ NSL KDD 7 Class ดำเนินการค่าพารามิเตอร์ Hidden Layer ที่ดีที่สุดเมื่อของเทคนิค MLP ได้ค่าพารามิเตอร์ที่ดีที่สุดดังตารางที่ 4.7 ในขั้นตอนของการหาค่า Hidden Layer Sizes จะทำซ้ำจนกว่าค่าความผิดพลาดที่เปลี่ยนแปลงน้อยกว่า 0.0001 จำนวน 2 epochs

ตารางที่ 4.1 ผลการหาค่าพารามิเตอร์จากชุดข้อมูล NSL KDD 6 คราส

ชุดข้อมูล	ค่าพารามิเตอร์	ความถูกต้อง
KDD 2 Class	hidden_layer_sizes 150	0.999
KDD 6 Class	hidden_layer_sizes 20	1.000
KDD 7 Class	hidden_layer_sizes 500	0.999
NSL KDD 2 Class	hidden_layer_sizes 200	0.979
NSL KDD 6 Class	hidden_layer_sizes 30	0.998
NSL KDD 7 Class	hidden_layer_sizes 200	0.988

นำข้อมูลจำนวน 10% ของชุดข้อมูล KDD 2 Class, KDD 6 Class, KDD 7 Class, NSL KDD 2 Class, NSL KDD 6 Class และ NSL KDD 7 Class ดำเนินการค่าพารามิเตอร์ n_neighbors, weights, leaf_size, algorithm ที่ดีที่สุดเมื่อของเทคนิค KNN ได้ค่าพารามิเตอร์ที่ดีที่สุดดังตารางที่ 4.8 ในขั้นตอนของการค้นหาจะทำการสร้างแบบจำลอง 3 folds ในแต่ละ folds มีทั้งสิ้น 120 แบบจำลองจำนวนทั้งหมด 360 แบบจำลอง และแสดงค่าพารามิเตอร์ที่ให้ค่าความถูกต้องสูงที่สุด

นำค่าพารามิเตอร์ที่ให้ค่าความถูกต้องสูงที่สุดดำเนินการทดสอบความถูกต้องในการจำแนกประเภทของการโจมตีแบบ DDoS

ตารางที่ 4.8 ผลการหาค่าพารามิเตอร์จากทุกชุดข้อมูลด้วยเทคนิค KNN

ชุดข้อมูล	ค่าพารามิเตอร์	ความถูกต้อง
KDD 2 Class	n_neighbors': 1, 'weights': 'uniform', 'leaf_size': 10, 'algorithm': 'ball_tree'	1.000
KDD 6 Class	n_neighbors': 1, 'weights': 'uniform', 'leaf_size': 10, 'algorithm': 'ball_tree'	1.000
KDD 7 Class	n_neighbors': 1, 'weights': 'uniform', 'leaf_size': 10, 'algorithm': 'ball_tree'	1.000
NSL KDD 2 Class	n_neighbors': 1, 'weights': 'uniform', 'leaf_size': 10, 'algorithm': 'ball_tree'	0.990
NSL KDD 6 Class	n_neighbors': 1, 'weights': 'uniform', 'leaf_size': 10, 'algorithm': 'ball_tree'	1.000
NSL KDD 7 Class	n_neighbors': 1, 'weights': 'uniform', 'leaf_size': 10, 'algorithm': 'ball_tree'	0.990

4.4 ผลการทดลองประสิทธิภาพของเทคนิค

จากการทดลองสร้างแบบจำลองการเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยเทคนิค SVM, KNN และ MLP ในงานวิจัยนี้ได้นำค่า Accuracy มาทำการเปรียบเทียบโดยทดสอบกับ Benchmark Dataset จำนวน 2 ชุดได้แก่ KDD ที่มีจำนวนข้อมูลทั้งสิ้น 529,655 แถว และชุดข้อมูล NSL KDD ที่มีจำนวน 12,354 แถว ชุดข้อมูลทั้ง 2 ชุดจะถูกแบ่งออกออกเป็น 3 กลุ่มย่อย ได้แก่ 2 Class, 6 Class และ 7 Class วิธีการทดสอบประสิทธิภาพของแบบจำลองนั้นใช้วิธีการทดสอบโดยใช้เทคนิค K-Fold Cross Validation ชุดข้อมูลจะถูกแบ่งออกเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ โดยกำหนดให้ K = 2, 5 และ 10 และทำการทดลองทั้งหมด 10 รอบ

เมื่อ K = 2 ข้อมูลจะถูกแบ่งออกเป็น 2 ส่วน 1 ส่วนเป็นข้อมูลสำหรับเรียนรู้และ 1 ส่วนจะถูกนำไปเพื่อทดสอบประสิทธิภาพของแบบจำลองและทำการสลับวนสลับกันเป็นชุดทดสอบจนครบทั้ง 2 ส่วน การทดสอบทั้งสิ้น 20 การทดลอง (10 round x 2 Folds = 20 การทดลอง)

เมื่อ K = 5 ข้อมูลจะถูกแบ่งออกเป็น 5 ส่วน ข้อมูล 4 ส่วนจะใช้ในการเรียนรู้และ 1 ส่วนจะถูกนำไปเพื่อทดสอบประสิทธิภาพของแบบจำลองและทำการสลับวนสลับกันเป็นชุดทดสอบจนครบทั้ง 5 ส่วนการทดสอบทั้งสิ้น 50 การทดลอง (10 round x 5 Folds = 50 การทดลอง)

เมื่อ K = 10 ข้อมูลจะถูกแบ่งออกเป็น 10 ส่วน ข้อมูล 9 ส่วนจะใช้ในการเรียนรู้และ 1 ส่วนจะถูกนำไปเพื่อทดสอบประสิทธิภาพของแบบจำลองและทำการสลับวนสลับกันเป็นชุดทดสอบจนครบ ทั้ง 10 ส่วน เมื่อทดสอบจำนวนทั้งสิ้น 10 รอบ หมายถึง การทดสอบทั้งสิ้น 100 การทดลอง

(10 round x 10 Folds = 100 การทดลอง) คำนวนหาค่าเฉลี่ยเพื่อให้เกิดความเชื่อมั่นตามหลักสถิติได้ผลการทดลองของชุดข้อมูล KDD ดังตารางที่ 4.9 และชุดข้อมูล NSL-KDD ดังตารางที่ 4.10

ตารางที่ 4.9 ผลการทดลองแสดงค่า Accuracy Values ของชุดข้อมูล KDD

จำนวนของชุดข้อมูล	ข้อมูลเรียนรู้/ข้อมูลทดสอบ (%)	วิธีการและความถูกต้อง(%)		
		SVM	KNN	MLP
KDD 2 Class	90/10	99.085 ±0.037	99.989 ±0.003	99.919 ±0.075
	80/20	99.055 ±0.031	99.989 ±0.003	99.924 ±0.052
	50/50	98.946 ±0.0 22	99.983 ±0.003	99.833 ±0.131
KDD 6 Class	90/10	98.995 ±0.071	99.998 ±0.003	99.975 ±0.021
	80/20	98.944 ±0.048	99.998 ±0.001	99.981 ±0.016
	50/50	98.781 ±0.020	99.996 ±0.002	99.975 ±0.018
KDD 7 Class	90/10	99.096 ±0.027	99.988 ±0.003	99.925 ±0.028
	80/20	99.053 ±0.035	99.989 ±0.004	99.813 ±0.237
	50/50	98.935 ±0.028	99.984 ±0.002	99.944 ±0.019

จากตารางที่ 4.9 แสดงค่า ACCURACY ของแบบจำลองในการจำแนกข้อมูลการโจมตีแบบ DDoS เมื่อทดสอบกับชุดข้อมูล KDD 2 Class

แบ่งข้อมูลออกเป็น 10 ส่วนด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.989%, MLP 99.919% และ SVM 99.085%

แบ่งข้อมูลออกเป็น 5 ส่วน ด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.989%, MLP 99.924% และ SVM 99.055%

แบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.983%, MLP 99.833% และ SVM 98.946%

ทดสอบกับชุดข้อมูล KDD 6 Class เมื่อแบ่งข้อมูลออกเป็น 10 ส่วนด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.998, MLP 99.975% และ SVM 98.995%

แบ่งข้อมูลออกเป็น 5 ส่วน ด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.998%, MLP 99.981% และ SVM 98.944%

แบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.996%, MLP 99.975% และ SVM 98.781%

ทดสอบกับชุดข้อมูล KDD 7 Class แบ่งข้อมูลออกเป็น 10 ส่วนด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.988, MLP 99.925% และ SVM 99.096%

แบ่งข้อมูลออกเป็น 5 ส่วน ด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.989%, MLP 99.813% และ SVM 99.053%

แบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธีการ Cross validation นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.984%, MLP 99.944% และ SVM 98.935%

ตารางที่ 4.10 ผลการทดลองแสดงค่า Accuracy Values ของชุดข้อมูล NSL-KDD

จำนวนของชุดข้อมูล	ข้อมูลเรียนรู้/ข้อมูลทดสอบ (%)	วิธีการและความถูกต้อง(%)		
		SVM	kNN	MLP
NSL-KDD 2 Class	90/10	92.322 ±0.484	99.113 ±0.135	98.084 ±0.276
	80/20	91.940 ±0.288	99.175 ±0.088	98.036 ±0.748
	50/50	91.171 ±0.194	99.191 ±0.044	98.091 ±0.265
NSL-KDD 6 Class	90/10	95.364 ±0.603	99.951 ±0.057	98.730 ±1.200
	80/20	92.728 ±0.587	99.951 ±0.026	96.807 ±4.627
	50/50	84.981 ±0.539	99.867 ±0.057	98.345 ±1.358
NSL-KDD 7 Class	90/10	92.464 ±0.357	99.072 ±0.228	98.181 ±0.204
	80/20	91.802 ±0.272	99.155 ±0.051	98.165 ±0.155
	50/50	91.182 ±0.183	99.087 ±0.076	98.066 ±0.137

จากตารางที่ 4.10 แสดงค่า ACCURACY ของแบบจำลองในการจำแนกข้อมูลการโจมตีแบบ DDoS เมื่อทดสอบกับชุดข้อมูล NSL-KDD 2 Class และทำการทดสอบประสิทธิภาพโดยแบ่งข้อมูลออกดังนี้

แบ่งข้อมูลออกเป็น 10 ส่วนด้วยวิธีการ Cross validation กำหนดให้ K = 10 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.113%, MLP 98.084% และ SVM 92.322%

แบ่งข้อมูลออกเป็น 5 ส่วน ด้วยวิธีการ Cross validation กำหนดให้ K = 5 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.175%, MLP 98.036% และ SVM 91.940%

แบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธีการ Cross validation กำหนดให้ K = 2 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.191%, MLP 98.091% และ SVM 91.171% เมื่อทดสอบกับชุดข้อมูล KDD 6 Class

แบ่งข้อมูลออกเป็น 10 ส่วนด้วยวิธีการ Cross validation กำหนดให้ K = 10 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.951, MLP 98.730% และ SVM 95.364%

แบ่งข้อมูลออกเป็น 5 ส่วน ด้วยวิธีการ Cross validation กำหนดให้ K = 5 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.951%, MLP 96.807% และ SVM 92.728%

แบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธีการ Cross validation กำหนดให้ K = 2 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.867%, MLP 98.345% และ SVM 84.981%

ทดสอบกับชุดข้อมูล KDD 7 Class แบ่งข้อมูลออกเป็น 10 ส่วนด้วยวิธีการ Cross validation กำหนดให้ K = 10 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.072, MLP 98.181% และ SVM 92.464%

แบ่งข้อมูลออกเป็น 5 ส่วน ด้วยวิธีการ Cross validation กำหนดให้ K = 5 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.155%, MLP 98.165% และ SVM 91.802% และ

แบ่งข้อมูลออกเป็น 2 ส่วน ด้วยวิธีการ Cross validation กำหนดให้ K = 2 นำมาทดสอบประสิทธิภาพของแบบจำลองพบว่าเทคนิค KNN มีค่ามากที่สุด 99.087%, MLP 98.066% และ SVM 91.182%

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อนำเสนอกระบวนการวิเคราะห์การโจมตีแบบ DDoS ด้วยการเรียนรู้ของเครื่องจักรภายใต้แนวคิดของการจำแนกข้อมูล สามารถสรุปผลการศึกษา อภิปรายผล รวมทั้งข้อเสนอแนะแนวทางในการวิจัยได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผล

5.3 ข้อเสนอแนะและงานวิจัยในอนาคต

5.1 สรุปผลการวิจัย

จากผลการทดลองได้นำเสนอวิธีการทางด้านการเรียนรู้เครื่องจักร (Machine Learning) ซึ่งประกอบด้วยเทคนิค KNN, MLP และ SVM เพื่อใช้สำหรับจำแนกการโจมตีแบบ DDoS โดยได้ทดสอบกับ Benchmark Dataset จำนวน 2 ชุด ได้แก่ KDD ที่ถูกกำจัดข้อมูลซ้ำทำให้มีจำนวนข้อมูลทั้งสิ้น 529,655 แถว และชุดข้อมูล NSL KDDที่ ถูกกำจัดข้อมูลซ้ำทำให้มีจำนวน 12,354 แถว โดยชุดข้อมูลทั้ง 2 จะถูกแบ่งออกออกเป็น 3 กลุ่มย่อย ได้แก่ 2, 6 และ 7 Class เพื่อทดสอบความถูกต้องในการจำแนกประเภทของการโจมตีแบบ DDoS และทดสอบประสิทธิภาพด้วยวิธีการ Cross validation ที่กำหนดให้ K = 2, 5 และ 10 เมื่อทดสอบจำนวนทั้งสิ้น 10 รอบพบว่าวิธี KNN เป็นวิธีที่ดีที่สุดเมื่อเทียบกับวิธี MLP และ SVM ซึ่งมีอัตราความถูกต้องสูงที่สุดถึง 99.99%

5.2 อภิปรายผล

จากการสร้างแบบจำลองจำแนกวิธีการโจมตีแบบ DDoS ข้างต้นแสดงให้เห็นว่าการจำแนกการโจมตีแบบ DDoS ด้วยวิธี KNN มีประสิทธิภาพสูงที่สุดในการจำแนกข้อมูลสาเหตุเนื่องจาก ข้อมูลการโจมตีแบบ ในกระบวนการจำแนกการข้อมูลด้วยวิธีการ KNN สามารถจำแนกกลุ่มข้อมูลที่มีจำนวนน้อยและสามารถจำแนกได้อย่างถูกต้องทำให้ได้ค่าความถูกต้องในการจำแนกสูงที่สุดเมื่อเปรียบเทียบกับวิธี SVM และ MLP ที่จำแนกข้อมูลผิดพลาดเมื่อกลุ่มของข้อมูลมีจำนวนน้อย เมื่อนำวิธีการดังกล่าวเปรียบเทียบกับวิธีของ Perakovic และคณะ นำเสนอวิธีการประยุกต์ใช้โครงข่ายประสาทเทียมการจำแนกการโจมตีแบบ DDoS โดยใช้ข้อมูลที่เผยแพร่ออนไลน์ทั้งสิ้นจำนวน 4 ชุด มีข้อมูลรวมกันทั้งสิ้น 4,986 ชุด โดยข้อมูลทั้งหมดแบ่งวิธีการบุกรุกออกเป็น 4 ประเภท คือ DNS DDoS attack, CharGen DDoS attack, UDP DDoS attack และ Normal จากการทดลองปรากฏ

ว่า Hidden Layer จำนวน 50 ชั้น ให้ผลการทดลอง 95.6% ซึ่งสูงที่สุด เมื่อนำมาเปรียบเทียบวิธีการในการทดลองมีความแต่ต่างของชุดข้อมูลที่ใช้แต่ได้นำวิธีการ MLP มาใช้ในการทดลองผลพบว่าให้ผลความถูกต้อง 98.73%

เมื่อนำไปเปรียบเทียบงานวิจัยของ Hsieh and Chan ที่นำเสนอวิธีการตรวจจับการโจมตีแบบ DDoS โดยใช้โครงข่ายประสาทเทียมโดยใช้เฟรมเวิร์กของ Apache Spark ซึ่งใช้ข้อมูลชุด ARPA 2000 LLDOS 1.0 ซึ่งข้อมูลถูกแบ่งประเภทของการบุกรุกออกเป็น 2 ประเภท คือ Normal และ Attack โดยมีข้อมูล Normal ข้อมูลทั้งหมดถูกแบ่งออกเป็น 30 % สำหรับข้อมูลชุดเรียนรู้ และ 70% สำหรับข้อมูลชุดทดสอบ จากการทดลองพบว่าให้ความถูกต้อง 94% เมื่อนำมาเปรียบเทียบวิธีการในการทดลองมีความแต่ต่างของชุดข้อมูลที่ใช้แต่ได้นำวิธีการ MLP มาใช้ในการทดลองผลและได้ทำการแบ่งข้อมูล 90% สำหรับข้อมูลชุดเรียนรู้ และ 10% สำหรับข้อมูลชุดทดสอบ จากผลการทดลองพบว่าให้ผลความถูกต้อง 98.73%

เมื่อนำไปเปรียบเทียบงานวิจัยของ Singh และ Tiwari นำเสนอวิธีการตรวจจับการบุกรุกโดยเทคนิค ID3 เพื่อลดจำนวนของคุณลักษณะพิเศษ จากข้อมูลชุด KDD ให้เหลือเพียง 18 คุณลักษณะพิเศษ ข้อมูลชุดเรียนรู้และข้อมูลชุดทดสอบถูกแบ่งออกเป็นสองส่วนเท่ากันและถูกนำไปเรียนรู้ด้วยเทคนิค KNNGA เพื่อทำการจำแนกข้อมูลและได้นำไปเปรียบเทียบกับวิธี KNN และ SVM จากการทดลองพบว่า วิธีการที่นำเสนอในงานวิจัยให้ความถูกต้องสูงกว่า 98% ซึ่งสูงกว่าทั้งวิธี KNN และ SVM เมื่อนำมาเปรียบเทียบวิธีการในการทดลองมีความแต่ต่างของชุดข้อมูล KDD ที่ถูกลบข้อมูลซ้ำและไม่ได้ทำการลดขนาดของคุณลักษณะพิเศษ โดยใช้คุณลักษณะพิเศษทั้งหมด 41 คุณลักษณะพิเศษ จากผลการทดลองพบว่าเทคนิค KNN ให้ผลความถูกต้อง 99.99% และเทคนิค SVM ให้ผลความถูกต้อง 99.09%

เมื่อนำไปเปรียบเทียบงานวิจัยของ Devaraju และ Ramakrishnan นำเสนอวิธีโครงข่ายประสาทเทียม 3 เทคนิค FFNN, PNN และ RBNN เพื่อทดสอบประสิทธิภาพของการจำแนกข้อมูลชนิดการโจมตีทดสอบกับข้อมูลข้อมูล KDD โดยข้อมูลชุดเรียนรู้และทดสอบจำนวนชุดละ 700 ข้อมูลได้ถูกเลือกมาจากข้อมูลทั้งสิ้นจำนวน 7 Class โดยแต่ละ Class จะถูกเลือกมาจำนวน 100 ชุด จากการทดลองพบว่า PNN มีประสิทธิภาพดีที่สุด โดย PNN, FFNN และ RBNN มีประสิทธิภาพที่ 97.5%, 94.3% และ 65% ตามลำดับ เมื่อเปรียบเทียบวิธีการดังกล่าวในการจำแนกข้อมูลการโจมตีเทคนิค MLP ให้ค่าความถูกต้องที่สูงกว่าเมื่อใช้จำนวนข้อมูลที่มากกว่าในการเรียนรู้แล้วทดสอบถึง 99.98%

เมื่อนำไปเปรียบเทียบงานวิจัยของ Ingre B. and Yadav A. ได้นำเสนอการวิเคราะห์ประสิทธิภาพชุดข้อมูล NSL KDD โดยใช้เทคนิค ANN เพื่อประเมินประสิทธิภาพความแม่นยำในการจำแนกข้อมูลทำการลดขนาดคุณลักษณะพิเศษให้เหลือ 29 คุณลักษณะพิเศษ ด้วยวิธีการ Information Gain, Gain Ratio และ Correlation Attribute Algorithm ทำการจำแนกข้อมูลด้วย

วิธีการ LM และ BFGS จากผลการทดลองพบว่าเทคนิค LM มีชั้นซ่อน 21 ชั้น จำนวน 117 Epoch มีค่าความแม่นยำ 81.2% มีค่าความแม่นยำสูงกว่า BFGS ที่มีชั้นซ่อน 23 ชั้น 771 Epoch มีค่าความแม่นยำ 79.9% เมื่อนำชุดข้อมูล NSL KDD มาใช้เทคนิค MLP ในการจำแนกข้อมูลที่มี 41 คุณลักษณะพิเศษ ให้ค่าความถูกต้อง 98.7%

เมื่อนำไปเปรียบเทียบงานวิจัยของ Pervez MS. and Farid DM. นำเสนอวิธีการเลือกคุณลักษณะทดสอบกับชุดข้อมูล NSL KDD ด้วยเทคนิค SVM ที่มีข้อมูลจำนวน 125,973 แถว จากผลการทดลองพบว่ามีความแม่นยำ 99.01% เมื่อนำชุดข้อมูลที่ลบข้อมูลซ้ำเหลือข้อมูลทั้งสิ้น 31,774 แถว และนำไปทดสอบด้วยเทคนิค SVM ให้ค่าความถูกต้อง 95.36% จากผลการทดลองแสดงให้เห็นว่าเมื่อไม่นำข้อมูลที่ซ้ำออก จะทำให้ผลการจำแนกข้อมูลที่สูงกว่า เนื่องจากข้อมูลที่ใช้ในการทดสอบอาจจะเป็นข้อมูลที่ซ้ำทำให้ได้ผลการจำแนกข้อมูลที่สูงขึ้นได้

เมื่อนำไปเปรียบเทียบงานวิจัยของ Yusof ARA, Udzir NI, Selamat A, Hamdan H and Abdullah MT. นำเสนอวิธีการเลือกคุณลักษณะพิเศษที่ปรับเปลี่ยนได้สำหรับการโจมตีแบบ DoS โดยใช้เทคนิคในการเลือกคือ Chi-squared นำไปทดสอบกับชุดข้อมูล NSL-KDD จำแนกข้อมูลด้วยเทคนิค ELM จากผลการทดลองพบว่ามีจำนวนของคุณลักษณะพิเศษ 17 คุณลักษณะพิเศษ มีค่าความแม่นยำที่ 96.2% สูงที่สุด เนื่อนำมาเปรียบเทียบด้วยเทคนิค SVM, MLP และ KNN ซึ่งเทคนิค KNN ให้ค่าความถูกต้องสูงที่สุด 99.95% ที่มีจำนวน 41 คุณลักษณะพิเศษ

เมื่อนำไปเปรียบเทียบงานวิจัยของ Meena G และ Choudhary RR นำเสนอวิธีการจำแนกด้วยเทคนิค J48 Graft และ NAIVE BAYES ทดสอบกับชุดข้อมูล 2 ชุด คือ KDD และ NSL KDD จากผลการทดลองพบว่าเทคนิค J48 Graft ให้ค่าความแม่นยำสูงที่สุดมีความแม่ยำ 99.435% และเทคนิค NAIVE BAYES น้อยกว่าโดยมีความแม่นยำ 92.715% เมื่อนำชุดข้อมูลดังกล่าวมาจำแนกข้อมูลด้วยเทคนิค KNN ด้วยชุดข้อมูล KDD ให้ความถูกต้อง 99.99% และชุดข้อมูล NSL KDD ให้ความถูกต้อง 99.95%

เมื่อนำวิธีการ SVM, MLP และ KNN ทดสอบกับชุดข้อมูลที่ได้จากระบบเครือข่ายจริงของมหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์ที่มีทั้งหมด 19,506 แถว มีคุณลักษณะพิเศษ 6 คุณลักษณะพิเศษ คือ Source_IP, Destination_IP, Protocol, Length, Source Port, Destination Port และมี 2 คลาสคือ Normal และ Attack จากผลการทดลองเทคนิค KNN ค่าความถูกต้องสูงสุด 99.74% เทคนิค SVM และ MLP ให้ค่าความถูกต้องตามลำดับดังนี้ 98.92% 98.84%

5.3 ข้อเสนอแนะ และงานวิจัยในอนาคต

จากงานวิจัยฉบับนี้ ผู้วิจัยได้เสนอแนะประเด็นสำคัญในการจำแนกการโจมตีแบบ DDoS ขั้นตอนในการจัดการข้อมูล วิธีการหาค่าพารามิเตอร์ที่ดีที่สามารถนำไปใช้ในวิธีการหาค่าพารามิเตอร์ของเทคนิคหรือวิธีการอื่นๆได้ เพื่อให้ได้ค่าความถูกต้องที่สูงขึ้น

สามารถนำงานวิจัยฉบับนี้ไปพัฒนาการจำแนกข้อมูลการโจมตีแบบ DDoS ด้วยวิธีการอื่นๆ เพื่อลดจำนวนของคุณลักษณะพิเศษที่ใช้ในการจำแนกข้อมูลหรือลดเวลาในการจำแนกข้อมูลโดยไม่ทำให้อัตราความถูกต้องลดลง

นำวิธีการดังกล่าวทำการทดลองกับการโจมตีในลักษณะอื่นหรือสามารถประยุคใช้วิธีการในงานวิจัยนี้เพื่อเป็นแนวทางในการพัฒนาระบบการจำแนกการโจมตีแบบ DDoS หรือการโจมตีทางระบบเครือข่ายคอมพิวเตอร์ได้

บรรณานุกรม

[1] Dulaney EA. CompTIA Security Study Guide Exam SY0-201. 4th ed. Canada: Sybex; 2008.

[2] จตุชัย แพงจันทร์. เจาะระบบ Network 2nd Edition. 1st ed. นนทบุรี: ไอดีซี อินโฟ ดิสทริบิวเตอร์ เซ็นเตอร์; 2551. 600 p.

[3] Peng T, Leckie C, RM Rao K. Detecting distributed denial of service attacks using source IP address. Monit Proc 3rd Int IFIP-TC6 Netw Conf. 2004;771–82.

[4] Zuckerman E, Roberts H, Mcgrady R, York J, Palfrey J. Distributed Denial of Service Attacks Against Independent Media and Human Rights Sites. 2010.

[5] Lau F, Rubin SH, Smith MH, Trajkovic L. Distributed Denial of Service Attacks. In: Systems Man and Cybernetics (SMC), IEEE International Conference on. 2000. p. 2275–80.

[6] NETWORK ADMINISTRATION: TCP/IP PROTOCOL FRAMEWORK [Internet]. 2019. Available from: https://www.dummies.com/programming/networking/network-administration-tcpip-protocol-framework/

[7] Perez H. IP header [Internet]. 2019. Available from: http://www.tutorialspoint.com/ipv4/images/ip_header.jpg

[8] Grid Search [Internet]. 2019. Available from: https://scikit-learn.org/stable/modules/grid_search.html

[9] Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res. 2012;13:281–305.

[10] Hu L-L, He Z-S, Shi X-H, Kong X-Y, Li H-P, Lu W-C. A nearest neighbor algorithm based predictor for the prediction of enzyme-small molecule interaction. Protein Pept Lett [Internet]. 2012 Jan;19(1):91–8. Available from: http://ieeexplore.ieee.org/document/1053964/

[11] Antti Ajanki. Example of k-nearest neighbour classificationnb [Internet]. 2019. Available from: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#/media/File:KnnClassification.svg

[12] Simon Haykin (McMaster University, Hamilton, Ontario C. Neural Networks A Comprehensive Foundation. 2005. p. 823.

[13] Marius-Constantin P, Balas VE, Perescu-Popescu L, Mastorakis N. Multilayer perceptron and neural networks. WSEAS Trans Circuits Syst. 2009;8(7):579–88.

[14] Cburnett. An example artificial neural network with a hidden laye [Internet]. 2019. Available from: https://upload.wikimedia.org/wikipedia/commons/thumb/e/e4/Artificial_neural_network.svg/1024px-Artificial_neural_network.svg.png

[15] Omer Galip Saracoglu. General structure of a multilayer perceptron (MLP) [Internet]. 2019. Available from: https://www.researchgate.net/figure/General-structure-of-a-multilayer-perceptron-MLP_fig2_51873136

[16] Veselý A, Brechlerova D. Neural networks in intrusion detection systems. In: Agrarian Perspectives, International conference on. 2003. p. 35–9.

[17] CORTES C, VAPNIK V. Support Vector Networks. Chem Biol Drug Des. 2009 Aug;74(2):142–7.

[18] Gosavi A, Khot S. Facial Expression Recognition Using Principal Component

Analysis. Int J soft Comput Eng. 2013;3(4):258–62.

[19] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In 1995.

[20] PACHARAWONGSAKDA E. Cross-validation Test [Internet]. 2014 [cited 2017 Jun 8]. Available from: http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

[21] Perakovic D, Perisa M, Cvitic I, Husnjak S. Artificial Neuron Network Implementation in Detection and Classification of DDoS Traffic. In: 24th Telecommunications Forum (TELFOR). 2016. p. 1–4.

[22] Hsieh CJ, Chan TY. Detection DDoS Attacks Based on Neural-Network using Apache Spark. In: Applied System Innovation (ICASI), International Conference on. 2016. p. 1–4.

[23] Singh P, Tiwari A. An Efficient Approach for Intrusion Detection in Reduced Features of KDD99 Using ID3 and Classification with KNNGA. In: 2015 Second Advances in Computing and Communication Engineering, IEEE International Conference on. 2015. p. 445–52.

[24] Devaraju S, Ramakrishnan S. Performance Analysis of Intrusion Detection System using Various Neural Network Classifiers. In: Recent Trends in Information Technology, IEEE International Conference on. 2011. p. 1033–8.

[25] Ingre B, Yadav A. Performance analysis of NSL-KDD dataset using ANN. In: Signal Processing and Communication Engineering Systems - Proceedings of SPACES 2015, in Association with IEEE International Conference on. 2015. p. 92–6.

[26] Pervez MS, Farid DM. Feature selection and intrusion classification in NSL-KDD

cup 99 dataset employing SVMs. In: SKIMA 2014 - 8th Software, Knowledge, Information Management and Applications, International Conference on. 2014.

[27] Yusof ARA, Udzir NI, Selamat A, Hamdan H, Abdullah MT. Adaptive feature selection for denial of services (DoS) attack. In: Applications Information and Network Security, AINS 2017, IEEE Conference on. 2018. p. 1–4.

[28] Meena G, Choudhary RR. A review paper on IDS classification using KDD 99 and NSL KDD dataset in WEKA. In: Computer Communications and Electronics (COMPTELIX), International Conference on. 2017. p. 553–8.

[29] KDD Cup 1999 Data [Internet]. 2017 [cited 2017 Nov 17]. Available from: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[30] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. IEEE Symp Comput Intell Secur Def Appl CISDA 2009. 2009;(Cisda):1–6.

[31] NSL-KDD [Internet]. 2019. Available from: https://iscxdownloads.cs.unb.ca/iscxdownloads/NSL-KDD/#NSL-KDD

บรรณานุกรม

ประวัติผู้เขียน

ชื่อ	นายธนพล เริ่มปลูก
วันเกิด	วันที่ 12 เมษายน พ.ศ. 2530
สถานที่เกิด	อำเภอเมือง จังหวัดปราจีนบุรี
สถานที่อยู่ปัจจุบัน	269 ถนนจิตรบำรุง ตำบลในเมือง อำเภอเมือง จังหวัดสุรินทร์
ตำแหน่งหน้าที่การงาน	นักวิชาการคอมพิวเตอร์
สถานที่ทำงานปัจจุบัน	มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์ 145 หมู่15 ตำบลนอกเมือง อำเภอเมืองสุรินทร์ จังหวัดสุรินทร์ 32000
ประวัติการศึกษา	พ.ศ. 2550 ปริญญาครุศาสตรอุตสาหกรรมบัณฑิต (ค.อ.บ.) สาขาวิชาเทคโนโลยีคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์ พ.ศ. 2562 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม
ทุนวิจัย	ทุนสนับสนุนการศึกษาบุคลากรสายสนับสนุน “มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน วิทยาเขตสุรินทร์”