บทที่ 3

วิธีการดำเนินการวิจัย

การดำเนินการวิจัยการสร้างชุดข้อมูลในการฝึกสอนไฟร์วอลล์ปัญญาประดิษฐ์ด้วยเทคโนโลยีโครงข่ายประสาทเทียมจากกฎของไฟร์วอลล์ มีจุดประสงค์เพื่อพัฒนาชุดข้อมูลฝึกสอนที่สร้างจากกฎของไฟร์วอลล์เพื่อให้ชุดข้อมูลฝึกสอนสามารถสอนโมเดลได้ถูกต้องและแม่นยำอย่างมีประสิทธิภาพ

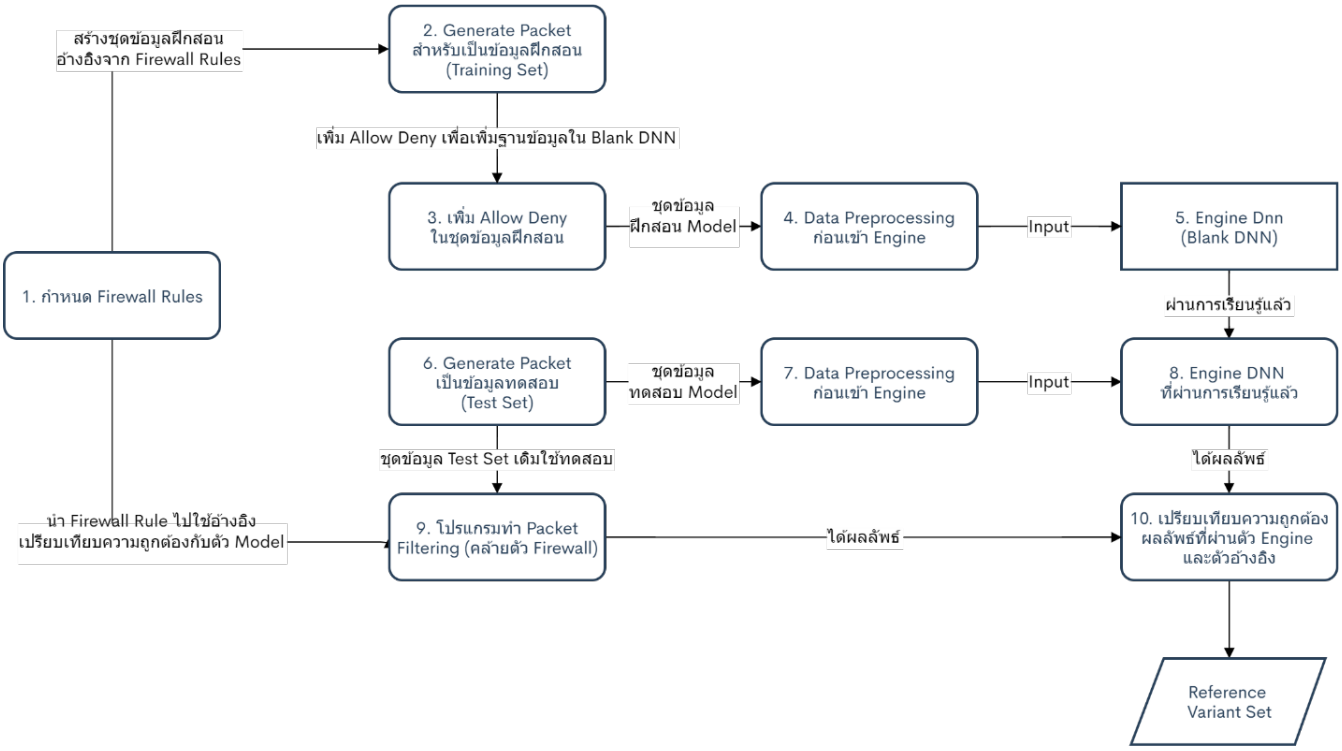
3.1 การเลือกใช้เทคโนโลยีและไลบราลีเพื่อพัฒนาโมเดล

ในการพัฒนาโปรแกรมที่เกี่ยวข้องและโมเดลฝึกสอนถูกเขียนขึ้นโดยภาษา python ทำให้ต้องศึกษาการทำงานและการใช้งานเพื่อให้พัฒนาได้สอดคล้องกับความต้องการของงานวิจัย

- TensorFlow ไลบราลีพื้นฐานในการพัฒนา Neural Network Model
  Version - TensorFlow 2.3.0 สามารถใช้ได้กับ python 64bit เท่านั้น
- Sklearn เป็นเครื่องมือในการทำงานของ Machine learning ทำงานโดยพื้นฐานของ Numpy
  Version – scikit-learn 0.23.2
- Keras เป็น Deep learning framework ที่สามารถประมวลผลได้ทั้ง CPU และ GPU
  Version – keras 2.4.3
- Pandas เป็นไลบราลีช่วยในการจัดกลุ่ม แยกประเภทข้อมูลกลุ่มโครงสร้างเช่น ไฟล์นามสกุล csv
  Version – pandas 1.1.2
- Pip เครื่องมือที่ช่วยในการติดตั้งแพ็คเกจในภาษา python
  Version – pip 20.2.3
- NVIDIA CUDA เครื่องมือช่วยให้คอมพิวเตอร์สามารถประมวลผลผ่าน GPU ได้
  Version – CUDA 11.1.0
- NVIDIA cuDNN เครื่องมือช่วยในการประมวลผล DNN ผ่าน GPU
  Version – cuDNN 8.0

3.2 กระบวนการพัฒนาชุดข้อมูลฝึกสอน Training model

ในการวิจัยจะมุ่งเน้นไปที่การพัฒนาชุดข้อมูลฝึกสอน โดยการเปรียบเทียบหาผลลัพธ์จากการนำชุดฝึกสอนไปผ่านโมเดล DNN และได้ผลลัพธ์ออกมาที่มีความแม่นยำมากที่สุด ซึ่งการทดลองดังกล่าวจำเป็นต้องทำด้วยกันหลายครั้ง ซึ่งในแต่ละครั้งการทดลองจะมีกระบวนการดำเนินงาน ดังนี้



จากรูปภาพ ทำให้แบ่งขั้นตอนการทดลองหลักๆได้เป็น 3 ส่วนใหญ่ตามการทำงานของโปรแกรม ได้แก่

- การสร้างชุดข้อมูลฝึกสอน หรือ training set ก่อนเข้าโมเดล
- การสร้างชุดข้อมูลทดสอบ หรือ test set หลังโมเดลผ่านการเรียนรู้
- การเปรียบเทียบผลลัพธ์ค่าความถูกต้องของโมเดลที่ทดสอบกับ test set

ส่วนที่ 1 การสร้างชุดข้อมูลฝึกสอน หรือ train set ก่อนเข้าโมเดล

การออกแบบกฎของไฟร์วอลล์

เงื่อนไขหลักของการวิจัยนี้คือการสร้างชุดข้อมูลฝึกสอนจากกฎของไฟร์วอลล์เพื่อให้ได้ระบบการป้องกันที่ได้มาตรฐาน ในส่วนแรกจึงจำเป็นต้องสร้างกฎของไฟร์วอลล์ที่จะนำมาใช้อ้างอิงทั้งการสร้างชุดข้อมูลทดสอบและการเปรียบเทียบความถูกต้องของโมเดลในช่วงสุดท้ายของการดำเนินการ

หลังจากสร้างกฎของไฟร์วอลล์แล้ว จะต้องดำเนินการสร้างชุดข้อมูลฝึกสอนที่มีการอ้างอิง ซึ่งจะต้องคำนวณความเป็นไปได้ทั้งหมดของข้อมูลภายใน packet ที่จะเกิดขึ้นในแต่ละ Field ที่ต้องมีการทำ Packet Filtering รวมกับพารามิเตอร์การตัดสินใจของ Firewall ที่จะตัดสินว่าจะทำการ Allow หรือ Deny ข้อมูลชุดนั้น โดยทั้งหมดจะถูกทำภายใต้โปรแกรม generate packet

- Action
  ความเป็นไปได้คือ Allow หรือ Deny ที่กำหนดไว้แล้ว ความเป็นไปได้จึงเป็น 1
- Source Address
  ความเป็นไปได้ขึ้นอยู่กับ subnet mask เช่น /24 จะเป็นไปได้ทั้งหมด $2^{(32-24)}$ ความเป็นไปได้
- Source Port
  ความเป็นได้ขึ้นอยู่กับจำนวน port ใน pull ที่กำหนดไว้
- Destination Address
  ความเป็นไปได้ขึ้นอยู่กับ subnet mask เช่น /24 จะเป็นไปได้ทั้งหมด $2^{(32-24)}$ ความเป็นไปได้
- Destination Port
  ความเป็นได้ขึ้นอยู่กับจำนวน port ใน pull ที่กำหนดไว้
- Protocol
  ประกอบไปด้วย TCP และ UDP

เมื่อได้ข้อมูลที่ถูก generate โดยอ้างอิงจากกฎของไฟร์วอลล์แล้วจะยังไม่สามารถเข้าโมเดลได้ จะต้องมีการเปลี่ยนแปลงรูปแบบของชุดข้อมูลให้โมเดลสามารถอ่านได้ เพื่อให้ง่ายต่อการเรียกใช้และบันทึกในครั้งถัดไปจึงได้มีการตั้งค่าให้แปลงข้อมูลให้อยู่ในรูปเลขฐานสองที่ถูกบันทึกอยู่ในไฟล์นามสกุล csv

ส่วนที่ 2 การสร้างชุดข้อมูลทดสอบ หรือ test set หลังโมเดลผ่านการเรียนรู้

ในส่วนนี้จะคล้ายคลึงกับส่วนแรก แต่การสร้างชุดข้อมูลทดสอบเพื่อเป็นแบบทดสอบสำหรับโมเดลที่ผ่านการเรียนรู้ว่ามีการ Filtering ที่ถูกต้องแม่นยำหรือไม่ ทำให้ชุดข้อมูล test set จะไม่มีการกำหนดพารามิเตอร์

Allow หรือ Deny ในข้อมูลชุดนั้น โดยชุดข้อมูลทดสอบทั้งหมดจะถูกสร้างและแปลงข้อมูลผ่านโปรแกรม generate packet เช่นกัน

ส่วนที่ 3 การเปรียบเทียบผลลัพธ์ค่าความถูกต้องของโมเดลที่ทดสอบกับ test set

เมื่อการทดสอบเสร็จสิ้น ในส่วนที่ 3 จะเป็นการนำชุดข้อมูลฝึกสอนผ่านโปรแกรมตรวจสอบความถูกต้องที่อ้างอิงจากกฎของไฟร์วอลล์ที่ออกแบบโดยตรง ซึ่งทำให้ข้อมูลที่ได้นั้นจะถูกต้องทั้งหมด และนำมาเปรียบเทียบกับผลลัพธ์ที่ได้จากโมเดลโดยผลลัพธ์ที่ได้จากการเปรียบเทียบจะอยู่ในรูปของ Reference Variant Set ดังภาพ

		Positive	Negative
Positive		True Positive (TP) Correct variant allele or position call	False Positive (FP) Incorrect variant allele or position call.
Negative		False Negative (FN) Incorrect reference genotype or no call.	True Negative (TN) Correct reference genotype or no call.

ผลลัพธ์ที่ได้จะประกอบไปทั้งหมด 4 ค่า ได้แก่

True Positive โมเดลอนุญาติให้ข้อมูลผ่านตรงตามกฎของไฟร์วอลล์ ให้ Allow ถูกต้อง

True Negative โมเดลไม่อนุญาติให้ข้อมูลผ่านตรงตามกฎของไฟร์วอลล์ ให้ Deny ถูกต้อง

False Positive โมเดลอนุญาติให้ข้อมูลผ่าน ไม่ตรงตามกฎของไฟร์วอลล์ ให้ Allow ผิดพลาด

False Negative โมเดลไม่อนุญาติให้ข้อมูลผ่าน ไม่ตรงตามกฎของไฟร์วอลล์ ให้ Deny ผิดพลาด

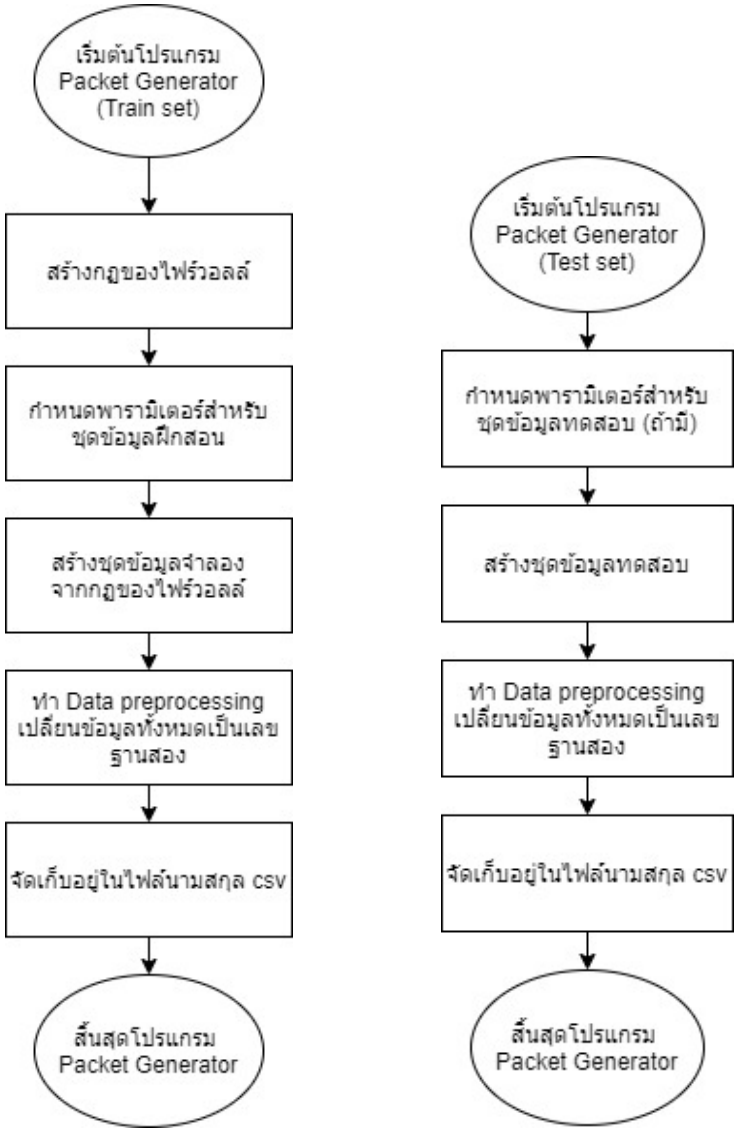
ผลลัพธ์ที่ได้จะเป็นไปตามสูตร

ความแม่นยำ (Accuracy) = SUM(TP, TN) / SUM(TP, TN, FP, FN)

3.3 กระบวนการสร้างโปรแกรมและเครื่องมือที่เกี่ยวข้อง

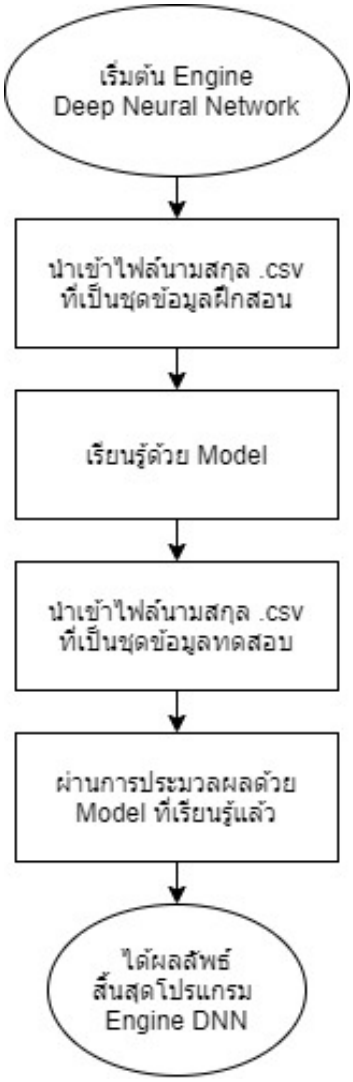
Packet Generator

เป็นโปรแกรมที่ใช้ในการสร้างชุดข้อมูล Packet โดยสุ่มจากพารามิเตอร์ที่กำหนดจากกฎของไฟร์วอลล์ โดยชุดข้อมูลที่ได้จากการสุ่มจะถูกนำไปแปลงค่าข้อมูลเป็นเลขฐานสอง บันทึกเก็บไว้ในไฟล์นามสกุล CSV ก่อนจะนำไปเรียกใช้ต่อในโมเดล Depp Neural Network โดยโปรแกรมนี้จะถูกแบ่งไปใช้ในการทำงาน 2 ส่วน ได้แก่ ส่วนที่ใช้ในการสร้างชุดข้อมูลฝึกสอน และ ส่วนที่ใช้ในการสร้างชุดข้อมูลทดสอบ



Deep Neural Network Model

เป็นเครื่องมือสร้าง Artificial Intelligent ที่พัฒนาขึ้นเอง โดยพัฒนาและประยุกต์โมเดลให้สามารถเรียนรู้กับชุดข้อมูลฝึกสอนที่ป้อนเข้าไป นำไปประมวลผล ตัดสินใจได้ว่าจะชุดข้อมูลที่ป้อนค่าเข้าไปนั้นเป็น Allow หรือ Deny



Compare Reference

เป็นโปรแกรมตรวจสอบความถูกต้องแม่นยำของโมเดล โดยชุดข้อมูลทดสอบจะถูกทำ Packet Filtering ที่โปรแกรมนี้ (ใส่ Allow และ Deny ให้) ผลลัพธ์จะออกมามีความแม่นยำสูง และเมื่อนำไปเปรียบเทียบกับโมเดล DNN แล้ว จะสามารถเปรียบเทียบและวิเคราะห์ความถูกต้องได้