การแก้ไขปัญหาในกระบวนการเรียนรู้แบบสหพันธ์สำหรับการจำแนกภาพที่มีการกระจายตัวของกลุ่มข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยแกนส์

โดย

ฐิติ ชื่นบุบผา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2565

SOLVING NON-IID IN FEDERATED LEARNING FOR IMAGE CLASSIFICATION USING GANS

BY

THITI CHUENBUBPHA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCE AND TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2022

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

ฐิติ ชื่นบุบผา

เรื่อง

การแก้ไขปัญหาในกระบวนการเรียนรู้แบบสหพันธ์สำหรับการจำแนกภาพที่มีการกระจายตัวของกลุ่มข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยแกนส์

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

เมื่อ วันที่ 13 มิถุนายน พ.ศ. 2566

ประธานกรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.เสาวลักษณ์ วรรธนาภา)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก _____
(ผู้ช่วยศาสตราจารย์ ดร.ประภาพร รัตนธำรง)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม _____
(ผู้ช่วยศาสตราจารย์ ดร.ฐาปนา บุญชู)

กรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.วนิดา พฤทธิวิทยา)

กรรมการสอบวิทยานิพนธ์ _____
(ดร.ชูชาติ หฤไชยะศักดิ์)

คณบดี _____
(รองศาสตราจารย์ ดร.สุเพชร จิรขจรกุล)

หัวข้อวิทยานิพนธ์	การแก้ไขปัญหาในกระบวนการเรียนรู้แบบสหพันธ์สำหรับการจำแนกภาพที่มีการกระจายตัวของกลุ่มข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยแกนส์
ชื่อผู้เขียน	ฐิติ ชื่นบุบผา
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.ประภาพร รัตนธำรง
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.ฐาปนา บุญชู
ปีการศึกษา	2565

บทคัดย่อ

ในปัจจุบันข้อมูลภาพสามารถนำมาประยุกต์ใช้ให้เกิดประโยชน์ในการทำงานได้หลากหลาย โดยในหลายหน่วยงานได้มีการเก็บข้อมูลภาพเพื่อนำมาใช้วิเคราะห์จำแนกประเภทด้วยโมเดลปัญญาประดิษฐ์และการเรียนรู้ของเครื่องกันมากขึ้น ซึ่งการจะได้มาซึ่งโมเดลที่มีความแม่นยำสูง จำเป็นจะต้องใช้ข้อมูลจำนวนมากในการเรียนรู้ และในหลายองค์กรอาจจะมีความต้องการในการพัฒนาโมเดลในลักษณะเดียวกัน ซึ่งหากสามารถแบ่งปันข้อมูลกันได้ก็จะสามารถช่วยพัฒนาโมเดลที่มีประสิทธิภาพที่ดียิ่งขึ้นได้ อย่างไรก็ตามข้อมูลเหล่านั้นอาจจะเป็นข้อมูลที่เป็นความลับของหน่วยงานที่ต้องถูกปกป้องด้วยความเป็นส่วนตัวของข้อมูล จึงไม่สามารถแบ่งปันข้อมูลเหล่านั้นซึ่งกันแหละกันในการที่จะนำไปใช้ในการสร้างระบบที่เป็นประโยชน์ต่อการทำงานได้ ระบบการเรียนรู้แบบสหพันธ์ได้เข้ามาแก้ไขปัญหาความเป็นส่วนตัวของข้อมูล โดยใช้เทคนิคการแลกเปลี่ยนพารามิเตอร์ของโมเดลโดยตรงแทนการแลกเปลี่ยนข้อมูล แต่ทว่าการเรียนรู้แบบสหพันธ์นั้นยังมีปัญหาเรื่องความถูกต้องเมื่อการกระจายตัวของข้อมูลระหว่างสมาชิกที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน ความถูกต้องของโมเดลหลักที่ถูกสร้างขึ้นด้วยการเรียนรู้แบบสหพันธ์ในกรณีดังกล่าวจะมีความถูกต้องที่ หนึ่งในวิธีแก้ปัญหาคือการใช้เทคนิคการเพิ่มข้อมูลเพื่อให้การกระจายตัวของข้อมูลกลับมาสมดุล แต่ยังมีประเด็นในการที่ต้องส่งข้อมูลออกไปยังแหล่งข้อมูลอื่น งานวิจัยนี้จึงสนใจการนำเทคนิคแกนส์มาใช้สำหรับสร้างข้อมูลสังเคราะห์ โดยคาดหวังให้การสร้างสังเคราะห์จากแกนส์จะสามารถแก้ไขปัญหา

การกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน ทำให้โมเดลหลักจากการเรียนรู้แบบสหพันธ์นั้นมีค่าความถูกต้องที่มากขึ้น

คำสำคัญ: การเรียนรู้แบบสหพันธ์, ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน, แกนส์, การเรียนรู้แบบกระจายศูนย์, โครงข่ายประสาทเทียม

Thesis Title	SOLVING NON-IID IN FEDERATED LEARNING FOR IMAGE CLASSIFICATION USING GANS
Author	Thiti Chuenbubpha
Degree	Master of Science (Computer Science)
Department/Faculty/University	Computer Science Faculty of Science and Technology Thammasat University
Thesis Advisor	Assistant Professor Prapaporn Rattanatamrong, Ph.D.
Thesis Co-Advisor	Assistant Professor Thapana Boonchoo, Ph.D.
Academic Year	2022

ABSTRACT

Nowadays image data can be applied in various ways to enhance work processes. Many organizations have been collecting image data to utilize it for classification and analysis using artificial intelligence and machine learning models. To achieve highly accurate models, a large amount of data is necessary for training. Moreover, many organizations may have a similar need to develop models with similar characteristics. If data can be shared, it can contribute to the improvement of more efficient models. However, these data might contain sensitive information that needs to be protected for privacy reasons, preventing their sharing for the development of beneficial systems.

Federated learning has emerged as a solution to address the privacy concerns associated with sharing sensitive data. It involves exchanging model parameters directly instead of sharing the actual data. Nevertheless, federated learning faces challenges in terms of accuracy when the distributed data among non-identical and non-independent members is imbalanced. The accuracy of the main model built

through federated learning may be affected. One approach to address this issue is to employ data augmentation techniques to rebalance the distribution of data. However, there is still an issue of sending data to external sources.

This research aims to explore the use of GANs (Generative Adversarial Networks) to generate synthetic data, with the expectation that synthesizing data from GANs can mitigate the problem of imbalanced and non-independent data distribution. Consequently, it can lead to improved accuracy of the main model trained using federated learning.

Keywords: Federated Learning, Non-IID, GANs, Distributed Learning, Neural Networks

กิตติกรรมประกาศ

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.ประภาพร รัตนธำรง และ ผู้ช่วยศาสตราจารย์ ดร.ฐาปนา บุญชู ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ที่สละเวลาให้ คำปรึกษา เสนอแนะแนวทาง สนับสนุน และช่วยเหลือ ใส่ใจและให้กำลังใจเสมอมาสำหรับ ความสำเร็จของวิทยานิพนธ์ฉบับนี้ต้องขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณคณาจารย์ประจำภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ทุกท่านที่คอยมอบความรู้ที่เป็นความรู้อันมีค่ายิ่ง ให้แก่ผู้วิจัย ช่วยเพิ่มคุณค่าให้กับวิทยานิพนธ์ฉบับนี้มากขึ้น

ขอบคุณทุนบัณฑิตเรียนดีจากคณะวิทยาศาสตร์และเทคโนโลยี ประจำปีการศึกษา 2562 เพื่อศึกษาต่อในระดับบัณฑิตศึกษา

ขอบคุณภาควิชาวิทยาการคอมพิวเตอร์และคณะวิทยาศาสตร์และเทคโนโลยีสำหรับทุนในเสนอผลงานภายในประเทศ ณ จังหวัดพิษณุโลกประจำปีการศึกษา 2565

ฐิติ ชื่นบุบผา

สารบัญ

สารบัญตาราง

สารบัญภาพ

รายการสัญลักษณ์และคำย่อ

สัญลักษณ์/คำย่อ	คำเต็ม/คำจำกัดความ
FL	Federated Learning
Non-iid	Non-Independent and Identically Distributed
GANs	Generative Adversarial Networks

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของโครงงาน

ในปัจจุบันเป็นยุคที่มีข้อมูลสารสนเทศจำนวนมาก หลากหลายองค์กรมีข้อมูลสารสนเทศเป็นของตัวเอง และด้วยเทคโนโลยีด้วยการเรียนรู้เชิงลึก (deep learning) จึงสามารถนำข้อมูลเหล่านั้นมาใช้พัฒนาระบบวิเคราะห์ข้อมูลเพื่อนำมาใช้ประโยชน์เพื่อเพิ่มประสิทธิภาพในการทำงานขององค์กรได้ แต่ในขณะเดียวกัน อาจจะมีบางองค์กรที่ไม่สามารถนำข้อมูลสารสนเทศที่มีมาใช้งานได้โดยตรง เนื่องจากข้อมูลที่มีไม่เพียงพอต่อการสร้างโมเดลสำหรับวิเคราะห์ข้อมูล ดังนั้นในการแบ่งปันข้อมูลระหว่างองค์กรที่มีข้อมูลชนิดเดียวกัน จึงเป็นสิ่งที่สามารถเพิ่มขีดจำกัดในการวิเคราะห์ข้อมูลของแต่ละองค์กรได้ และในยุคที่ข้อมูลเป็นสิ่งที่มีค่าเปรียบเสมือนน้ำมันนั้น การปกป้องข้อมูลส่วนบุลคล และความเป็นส่วนตัวของข้อมูล (data privacy) ก็เป็นเรื่องที่ได้รับความสนใจมากขึ้นในหลายองค์กร รวมถึงองค์กรที่เก็บข้อมูลสำคัญที่จำเป็นจะต้องรักษาความเป็นส่วนตัวเป็นพิเศษ เช่น โรงพยาบาลที่ต้องเก็บข้อมูลประวัติการรักษา เป็นต้น ทำให้มีข้อจำกัดในการแบ่งปันข้อมูลระหว่างองค์กรเพื่อนำไปใช้ในการสร้างโมเดลสำหรับการวิเคราะห์ข้อมูล กระทั่งในปี 2017 McMahan ได้นำเสนออัลกอริทึม FedAVG [26] ซึ่งถือเป็นอัลกอริทึมการเรียนรู้แบบสหพันธ์ (Federated Learning) เป็นตัวแรก โดยได้นำเสนอแนวคิดการใช้งานข้อมูลที่เก็บไว้ในแต่ละองค์กรที่แตกต่างกันในการเรียนรู้โมเดล โดยที่ไม่ต้องทำการเคลื่อนย้ายข้อมูลจริง (raw data) โดยการแลกเปลี่ยนพารามิเตอร์ (parameter) แทนในแต่ละรอบของการเรียนรู้ และนำพารามิเตอร์ที่ได้จากแต่ละแหล่งข้อมูลมารวมกันเพื่อสร้างเป็นโมเดลหลัก ด้วยวิธีการนี้ทำให้สามารถใช้งานข้อมูลในแต่ละองค์กรเพื่อสร้างองค์ความรู้ให้กับโมเดล และยังรักษาไว้ซึ่งความเป็นส่วนตัวของข้อมูล

การเรียนรู้แบบสหพันธ์นั้นมีข้อจำกัดโดยเฉพาะเมื่อข้อมูลที่อยู่ในแต่ละแห่งนั้นมีความกระจายตัวที่ไม่เหมือนกันหรือข้อมูลไม่เป็นอิสระต่อกัน (non-Independent and Identically Distributed : non-iid) ซึ่งส่งผลจุดสมดุลที่ดีที่สุดของโมเดลที่สร้างขึ้นจากข้อมูลของแต่ละองค์กร (local optima) นั้นอยู่ห่างจากจุดสมดุลที่ดีที่สุดที่เป็นไปได้ (global optima) เมื่อนำโมเดลของแต่ละแหล่งข้อมูล (local model) มารวมกันแล้วจะทำให้ได้จุดสมดุลที่ดีที่สุดที่อยู่ห่างจากจุดสมดุลที่ดีที่สุดที่เป็นไปได้ เช่น การที่โรงพยาบาลต้องการใช้ข้อมูลภาพเอ็กซเรย์ในการสร้างโมเดลทำนายโอกาสในการเกิดโรค โดยหากข้อมูลที่แต่ละโรงพยาบาลเก็บนั้นมีความต่างในด้านคุณลักษณะของข้อมูลมากเกินไป เช่น ภาพเอ็กซเรย์ที่ได้จาก โรงพยาบาลเด็ก และโรงพยาบาลผู้สูงอายุนั้นมีความแตกต่างกันใน

ด้านคุณลักษณะของข้อมูล แม้ว่าจะเป็นภาพเอ็กซเรย์ในตำแหน่งเดียวกัน แต่รูปร่างของเด็กและผู้ใหญ่ก็มีความแตกต่างกัน รวมถึงรอยโรคที่แสดงในเด็กและผู้ใหญ่ก็อาจจะมีความแตกต่างกัน ส่งผลให้คุณลักษณะของข้อมูลภาพเอ็กซเรย์ของทั้งสองโรงพยาบาลนี้มีความแตกต่างกัน

โดยในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระของข้อมูลนั้น Hangyu Zhu (2021) ได้นำเสนอบทความเกี่ยวกับปัญหานี้กับการเรียนรู้แบบสหพันธ์ [24] หนึ่งในวิธีการแก้ปัญหาที่น่าสนใจคือการแก้ไขจากแหล่งข้อมูลโดยตรง โดยมีสองวิธีที่เป็นไปได้คือ การแบ่งปันข้อมูล (sharing data) และการเพิ่มข้อมูล (data augmentation) โดยการแบ่งปันข้อมูลนั้น Yue Zhao (2018) ได้นำเสนอวิธีการ [43] ในการแก้ไขการแก้ไขปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระของข้อมูล ด้วยการแบ่งข้อมูลประมาณ 5% ของแต่ละแหล่งข้อมูล เพื่อสร้างข้อมูลศูนย์กลาง (shared data) สำหรับแบ่งปันเพื่อใช้ในการเรียนรู้ของแต่ละแหล่งข้อมูล ซึ่งการแบ่งปันข้อมูลนั้นจะทำให้ระบบการเรียนรู้แบบสหพันธ์นั้นสูญเสียความสามารถในการปกป้องความเป็นส่วนตัวของข้อมูล ซึ่งเป็นหัวใจสำคัญของการเรียนรู้แบบสหพันธ์ วิธีการเพิ่มข้อมูลจึงมีความน่าสนใจกว่า ในขั้นตอนการเพิ่มข้อมูลนั้น หากใช้เทคนิคการเพิ่มข้อมูลแบบปกติ ในแต่ละแหล่งข้อมูลจะสามารถแก้ปัญหาความไม่เท่ากันของข้อมูล (data imbalance) ในบางคลาส (class) ได้ด้วยการเพิ่มข้อมูลของคลาสนั้นๆจากข้อมูลที่มีอยู่ ซึ่งเป็นข้อมูลที่มีคุณลักษณะเหมือนเดิม เมื่อเข้าสู่กระบวนการเรียนรู้ อาจจะทำให้เกิดการโอเวอร์ฟิตติ้ง (over-fitting) ได้ เนื่องจากข้อมูลในบางคลาสมีความหลากหลายของคุณลักษณะที่ไม่มากพอ

ในปี 2017 Ian J. Goodfellow ได้นำเสนอเทคนิคแกนส์ (Generative Adversarial Nets : GANs) [27] โดยแกนส์จะสามารถสร้างข้อมูลใหม่ได้ จากคุณลักษณะของข้อมูลที่มี โดยเมื่อใช้เทคนิคแกนส์ในการเรียนรู้คุณลักษณะ (features) ของข้อมูลภาพ จะได้โมเดลผู้สร้าง (generator model) ที่สามารถสร้างข้อมูลภาพสังเคราะห์ (synthetic image data) ที่มีคุณลักษณะเหมือนกับข้อมูลต้นฉบับที่ใช้ในการเรียนรู้ได้ โดยเมื่อใช้เทคนิคแกนส์ในการเรียนรู้ข้อมูลภาพจากทุกๆแหล่งในกระบวนการเรียนรู้แบบสหพันธ์ จะได้โมเดลผู้สร้างที่มีความสามารถในการสร้างข้อมูลภาพสังเคราะห์ที่มีคุณลักษณะเหมือนกับข้อมูลต้นฉบับมากที่สุด และสามารถใช้ข้อมูลเหล่านี้ในการแก้ปัญหาความกระจายตัวที่ไม่เหมือนกันหรือข้อมูลไม่เป็นอิสระต่อกัน โดยยังคงรักษาไว้ซึ่งความเป็นส่วนตัวของข้อมูล

งานวิจัยนี้มุ่งเน้นไปที่การแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลด้วยวิธีการเพิ่มข้อมูล โดยใช้ข้อมูลที่ถูกสร้างขึ้นจากแกนส์ โดยจะทำการทดลองเปรียบเทียบความถูกต้องของโมเดลซึ่งได้จากการเรียนรู้แบบสหพันธ์ที่ผ่านการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลด้วยวิธีการเพิ่มข้อมูลสังเคราะห์ที่สร้างจากแกนส์เปรียบเทียบกับโมเดลที่ได้จากการเรียนรู้แบบสหพันธ์โดยไม่ใช้ข้อมูลสังเคราะห์ และการ

แก้ปัญหาโอเวอร์ฟิตติ้งของกระบวนการเรียนรู้โมเดลแกนส์บนระบบเก็บข้อมูลแบบกระจายศูนย์โดยหาขนาดชุดข้อมูลย่อยที่เหมาะสมเพื่อใช้ในกระบวนการเรียนรู้ของโมเดลแกนส์ และใช้ปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมด (communication overhead) ให้น้อยที่สุด

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 สร้างโมเดลผู้สร้างจากข้อมูลจริงบนทุกๆ แหล่งข้อมูล เพื่อสร้างข้อมูลใหม่ที่มีคุณลักษณะเหมือนกับข้อมูลจริง และสามารถใช้แทนข้อมูลจริงในการเรียนรู้แบบสหพันธ์เพื่อแก้ปัญหาความกระจายตัวที่ไม่เหมือนกันหรือข้อมูลไม่เป็นอิสระต่อกัน โดยยังคงไว้ซึ่งความเป็นส่วนตัวของข้อมูล

1.2.2 ออกแบบขั้นตอนวิธีในการเรียนรู้ของแกนส์บนระบบที่มีการเก็บข้อมูลแบบกระจายตัวให้สามารถเรียนรู้คุณลักษณะของข้อมูลได้อย่างถูกต้องจากทุกแหล่งข้อมูล โดยหลีกเลี่ยงการก่อให้เกิดโอเวอร์ฟิตติ้งกับแหล่งข้อมูลที่เรียนรู้เป็นแหล่งสุดท้าย

1.2.3 ทดสอบการแก้ปัญหาการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระของข้อมูลในการเรียนรู้แบบสหพันธ์ด้วยการเพิ่มข้อมูลที่ถูกสร้างจากโมเดลผู้สร้าง โดยวัดผลความถูกต้องเทียบกับการการเรียนรู้แบบสหพันธ์ที่ไม่ใช้การเพิ่มข้อมูลในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระของข้อมูล

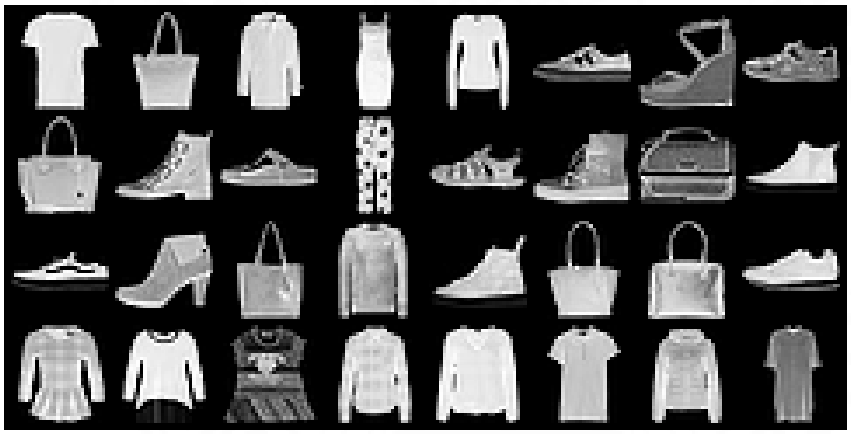
1.3 ขอบเขตงานวิจัย

1.3.1 ขอบเขตด้านข้อมูล

งานวิจัยนี้ออกแบบมาเพื่อทำงานกับข้อมูลภาพโดยเฉพาะ โดยใช้ชุดข้อมูล ดังนี้

- MNIST เป็นภาพลายมือของตัวเลข 0-9 โดยเป็นภาพขาวดำ ขนาด 28x28 พิกเซล ดังที่แสดงในรูปที่ 1.1 จำนวนทั้งหมด 60,000 ภาพ
- Fashion-MNIST เป็นภาพเครื่องแต่งกายจำนวน 10 ประเภท โดยเป็นภาพขาวดำ ดังที่แสดงในรูปที่ 1.2 ขนาด 28x28 พิกเซล จำนวนทั้งหมด 60,000 ภาพ

ภาพที่ 1.1 ภาพตัวอย่างชุดข้อมูล MNIST[50]



ภาพที่ 1.2 ภาพตัวอย่างชุดข้อมูล FMNIST[51]

1.3.2 ขอบเขตด้านโมเดล

งานวิจัยนี้จะใช้รูปแบบโมเดลสำหรับการเรียนรู้แบบสหพันธ์เป็นประเภทโครงข่ายประสาทเทียมแบบสังวัตนาการ (convolutional neural network) ซึ่งเป็นโมเดลสำหรับการทำการจำแนกประเภท (classification) ของภาพโดยเฉพาะเนื่องจากข้อมูลเป็นข้อมูลประเภทภาพ และโมเดลสำหรับการเรียนรู้ของแกนส์จะใช้คอนดิชันนอลแกนส์ (Conditional Generative Adversarial Networks : CGANs) เพื่อให้สามารถระบุประเภทของข้อมูลที่ต้องการสร้างได้

1.3.3 ขอบเขตด้านวิธีการ

งานวิจัยนี้จะวัดผลการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระของข้อมูลด้วยข้อมูลภาพที่ถูกสร้างขึ้นโดยแกนส์ โดยจะจำลองสถานการณ์ที่ระบบการเรียนรู้แบบสหพันธ์ต้องทำงานกับแหล่งข้อมูลหลายๆ แห่งที่มีการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระของข้อมูล โดยจะวัดผลจากความถูกต้องของโมเดลที่เรียนรู้กับข้อมูลที่ยังไม่ถูกแก้ปัญหา และโมเดลที่เรียนรู้กับข้อมูลที่แก้ปัญหาแล้วด้วยวิธีเพิ่มข้อมูลโดยใช้แกนส์

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 สามารถสร้างข้อมูลสังเคราะห์ที่มีคุณลักษณะใกล้เคียงกับข้อมูลต้นฉบับมากที่สุด ในกระบวนการเรียนรู้แบบสหพันธ์ โดยการเรียนรู้จากทุกแหล่งข้อมูล

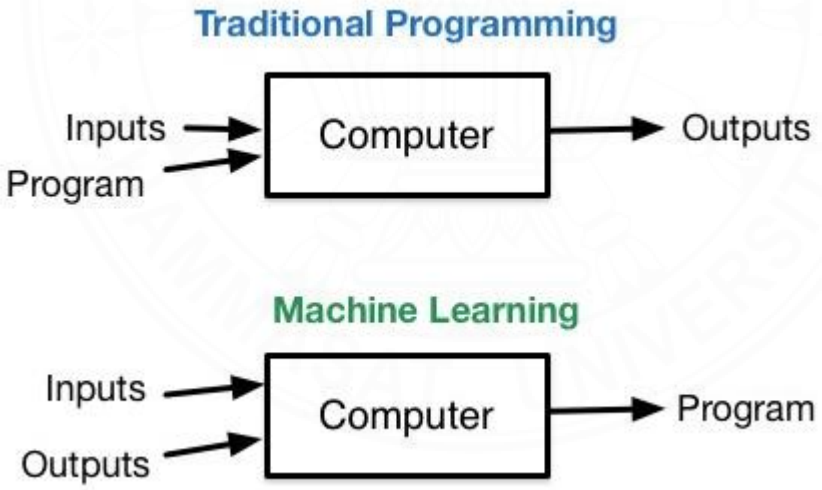
1.4.2 สามารถแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยเทคนิคการเพิ่มข้อมูล โดยยังคงไว้ซึ่งความเป็นส่วนตัวของข้อมูล

1.4.3 สามารถเพิ่มประสิทธิภาพการทำงานของระบบการเรียนรู้แบบสหพันธ์ได้ โดยใช้การเพิ่มข้อมูลที่ถูกสร้างขึ้นโดยแกนส์

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีเกี่ยวกับการเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง คือการทำให้คอมพิวเตอร์เรียนรู้ได้ด้วยตนเองโดยใช้ข้อมูล ซึ่งแตกต่างจากการเขียนโปรแกรมทั่วไปที่โปรแกรมเมอร์ (programmer) พยายามที่จะเขียนโปรแกรมเพื่อควบคุมให้ข้อมูลเข้า (input) มีค่าออกมาเป็นผลลัพธ์ (output) ในแบบที่ต้องการ [17] ในขณะที่การเรียนรู้ของเครื่องนั้น จะทำการป้อนข้อมูลเข้าและผลลัพธ์ที่ต้องการเข้าไปในระบบ เพื่อให้ระบบโมเดลสำหรับการจับคู่ข้อมูลเข้าไปยังผลลัพธ์ที่ต้องการ ดังที่แสดงในภาพที่ 2.1 โดยการเรียนรู้ของเครื่องนั้นจำเป็นที่จะต้องใช้ข้อมูลที่ทำการสกัดคุณสมบัติ (features extraction) โดยผู้เชี่ยวชาญแล้ว จึงจะสามารถเริ่มต้นกระบวนการเรียนรู้ของเครื่องได้



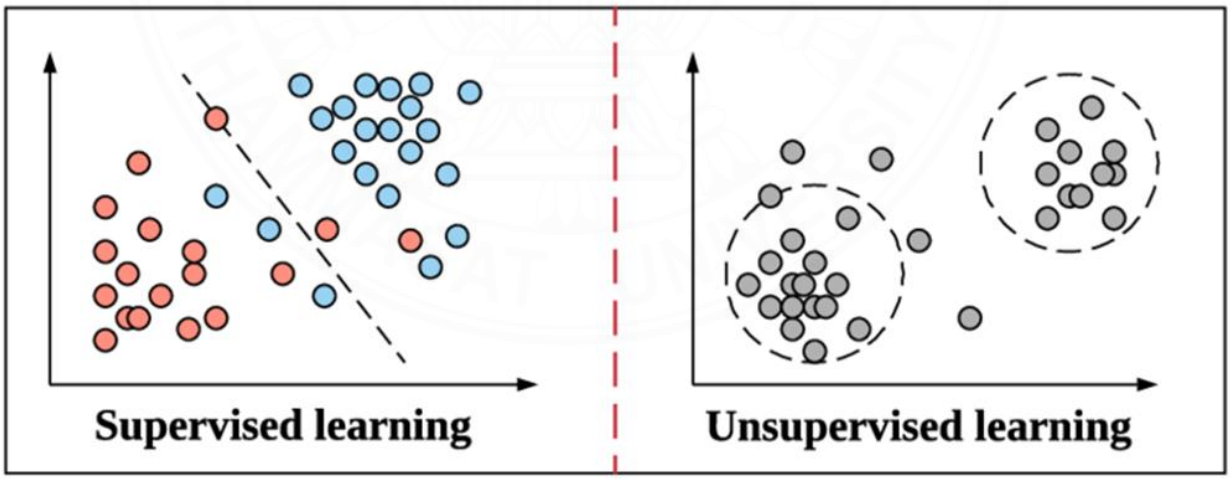
ภาพที่ 2.1 เปรียบเทียบการทำงานของการเขียนโปรแกรมทั่วไปเทียบกับการเรียนรู้ของเครื่อง[18]

2.1.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

การเรียนรู้แบบมีผู้สอนเป็นการเรียนรู้ที่ผู้ใช้งานจะทำการป้อนข้อมูลพร้อมกับข้อมูลผลลัพธ์ (target attribute) ที่ต้องการในขั้นตอนการเรียนรู้ โดยโมเดลจะทำการหาความสัมพันธ์ของข้อมูลเข้า (input) กับข้อมูลผลลัพธ์ตามแต่ละเทคนิคที่ใช้ เพื่อให้ได้โมเดลที่สามารถให้คำตอบจากข้อมูลเข้าใดๆได้ตามที่ต้องการ [17] โดยการเรียนรู้แบบมีผู้สอนนั้นจะเหมาะกับงานประเภทจำแนก (classification) และงานประเภทผลลัพธ์ต่อเนื่อง (regression)

2.1.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้แบบไม่มีผู้สอนนั้นจะแตกต่างกับการเรียนรู้แบบมีผู้สอนในแง่ของข้อมูลเข้า เนื่องจากในกระบวนการเรียนรู้แบบไม่มีผู้สอนนั้นไม่จำเป็นต้องมีข้อมูลผลลัพธ์ที่ต้องการ โดยในกระบวนการเรียนรู้จะทำการค้นหาความสัมพันธ์ของข้อมูลในหลายๆมิติ เพื่อจัดกลุ่มของข้อมูล (clustering) ที่มีคุณลักษณะเหมือนกันให้จำแนกเป็นกลุ่มเดียวกัน [18] ซึ่งจะแตกต่างจากการเรียนรู้แบบมีผู้สอนเนื่องจากไม่จำเป็นต้องมีข้อมูลผลลัพธ์ ดังที่แสดงในภาพที่ 2.2

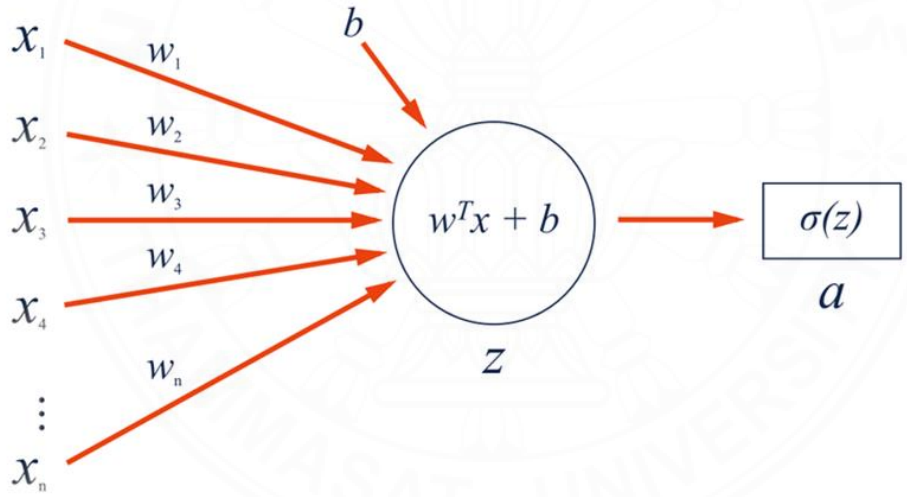


ภาพที่ 2.2 ตัวอย่างความแตกต่างของการเรียนรู้แบบมีผู้สอนและการเรียนรู้แบบไม่มีผู้สอน[38]

2.2 ทฤษฎีโครงข่ายประสาทเทียม (Artificial Neural Networks: ANNs)

2.2.1 หน่วยประสาทเทียม

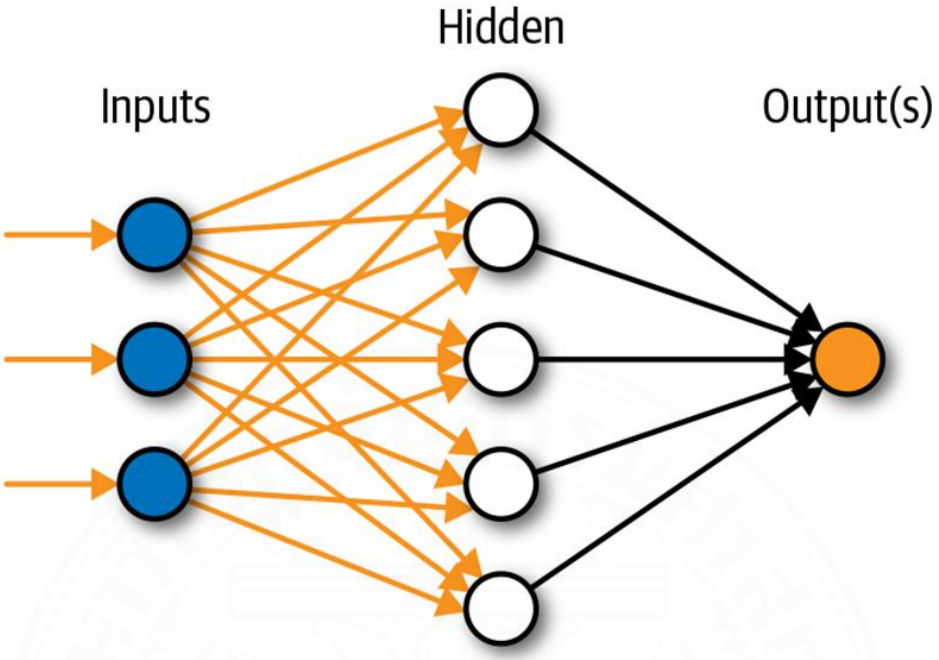
หน่วยประสาทเทียม (artificial neuron) คือ ระบบของคอมพิวเตอร์ที่ถูกสร้างขึ้นเพื่อเลียนแบบการทำงานระบบประสาทของมนุษย์ โดยในสมองของมนุษย์จะมีหน่วยประมวลผลเล็กๆ จำนวนมาก ที่เชื่อมโยงถึงกันด้วยโครงข่ายประสาท ซึ่งทำให้มนุษย์สามารถคิด วิเคราะห์ และจำแนกข้อมูลได้อย่างรวดเร็ว [18] โดยประสาทเทียมนั้นสร้างขึ้นโดยเลียนแบบการทำงานของหน่วยประมวลผลดังกล่าว โดยจะเรียกหน่วยประสาทเทียมนี้ว่าโหนด (node) ในโครงข่ายประสาทเทียม โดยจะมีข้อมูลเข้า (input) เป็น x โดยข้อมูลเข้าจะถูกนำมาคูณกับค่าถ่วงน้ำหนัก (weight) และบวกกับค่าเบี่ยงเบน (bias) เพื่อให้ได้ผลลัพธ์คือค่า z และนำผลลัพธ์ไปคำนวณผ่านฟังก์ชันกระตุ้น (activation function) เพื่อให้ได้ค่า a ซึ่งเป็นผลลัพธ์ของหน่วยประสาทเทียม ดังที่แสดงในภาพที่ 2.3



ภาพที่ 2.3 หน่วยประสาทเทียมที่จำลองจากระบบประสาทของมนุษย์[1]

2.2.2 โครงข่ายประสาทเทียม (Artificial Neural Networks)

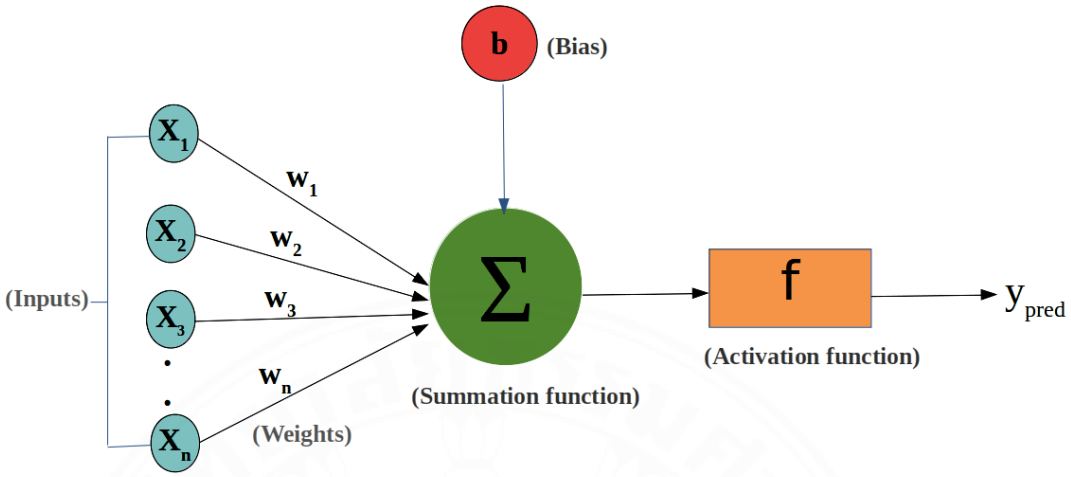
โครงข่ายประสาทเทียมคือการทำงานร่วมกันของหน่วยประสาทเทียมหลายๆหน่วยเชื่อมต่อกันในรูปแบบเดียวกันกับโครงข่ายประสาทของมนุษย์ ในการจำลองวิธีการเรียนรู้ของมนุษย์ เพื่อสร้างระบบคอมพิวเตอร์ที่มีความสามารถในเรียนรู้ และวิเคราะห์ได้เหมือนกับโครงข่ายประสาทของมนุษย์ [1] โครงข่ายประสาทเทียมมีโครงสร้างประกอบด้วย 3 ลำดับชั้น คือ ชั้นข้อมูลเข้า (input layer) ชั้นซ่อน (hidden layer) และชั้นผลลัพธ์ (output layer) ดังที่แสดงในภาพที่ 2.4



ภาพที่ 2.4 โครงสร้างของโครงข่ายประสาทเทียมโดยทั่วไป[1]

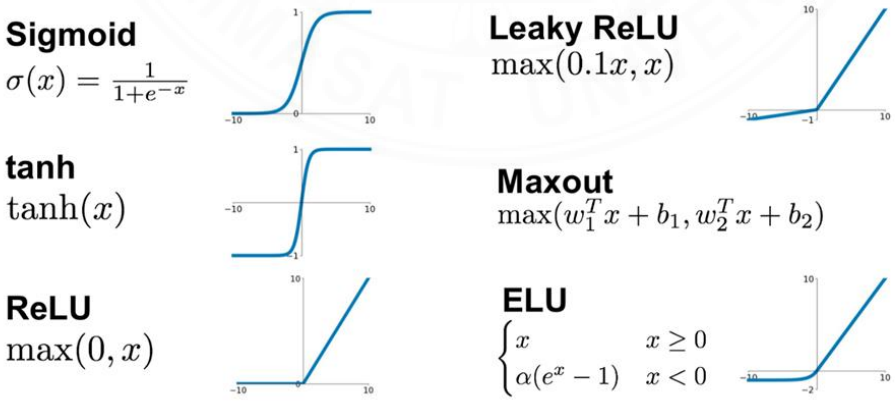
ชั้นข้อมูลเข้า จะรับข้อมูลเป็นคุณลักษณะสำคัญ (Feature) ซึ่งเป็นข้อมูลตัวเลขแบบ 1 มิติ โดยในชั้นข้อมูลเข้านั้น จะมีเพียงชั้นเดียว และมีหน้าที่ส่งข้อมูลไปยังชั้นถัดไป (ชั้นซ่อน)

ชั้นซ่อน ทำหน้าที่รับข้อมูลจากชั้นข้อมูลเข้า โดยชั้นซ่อนสามารถมีจำนวนชั้นได้มากกว่า 1 ชั้น โดยพื้นฐานแล้ว ประสิทธิภาพของโครงข่ายประสาทเทียมจะมากขึ้น หากมีจำนวนชั้นซ่อน และหน่วยประสาทเทียมจำนวนมาก [1] ภายในชั้นซ่อนจะมีตัวแปรที่ใช้ในการฝึกสอนโครงข่ายประสาทเทียม นั่นคือ ค่าถ่วงน้ำหนัก (weight) และ ค่าความเอนเอียง (bias) สองตัวแปรนี้จะสามารถเปลี่ยนแปลงค่าเองได้ในระหว่างการเรียนรู้ของโครงข่ายประสาทเทียม (trainable parameters) โดยในทุกๆหน่วยประสาทเทียมจะมีตัวแปรสองตัวนี้ไว้คอยปรับค่าของข้อมูลเข้า (Inputs) ซึ่งอาจเป็นชุดข้อมูลขนาดใดๆ X1, X2, ... Xn ที่ได้รับก่อนจะส่งไปยังฟังก์ชันการแปลง (activation function) เพื่อคำนวนออกมาเป็นผลลัพธ์ที่น่าจะเป็น (prediction) ดังที่แสดงในภาพที่ 2.5 โดยฟังก์ชันการแปลงนั้นจะเป็นตัวกำหนดผลลัพธ์ของหน่วยประสาทเทียมนั้นๆ ก่อนที่จะส่งไปยังหน่วยประสาทเทียมในชั้นถัดไป โดยการเลือกใช้ฟังก์ชันการแปลงจะส่งผลโดยตรงกับผลลัพธ์ของโครงข่ายประสาทเทียม



ภาพที่ 2.5 การทำงานของหน่วยประสาทเทียมในชั้นซ่อน[9]

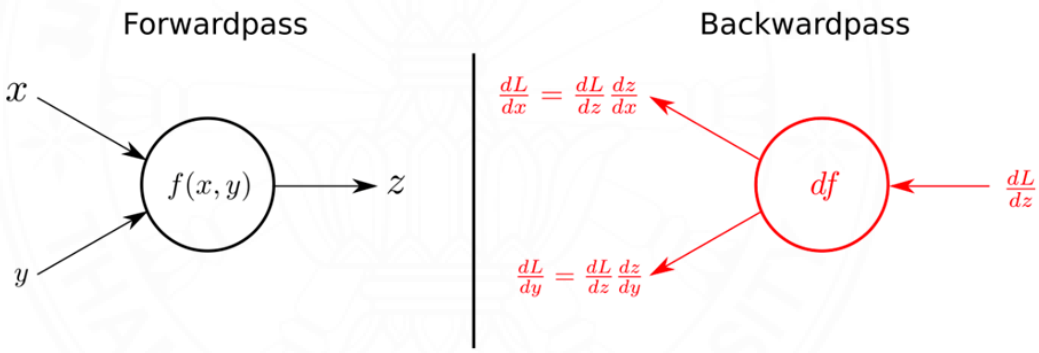
ชั้นผลลัพธ์ เป็นชั้นสุดท้ายในโครงข่ายประสาทเทียม โดยจะมีจำนวนหน่วยประสาทเทียมในชั้นนี้เท่ากับจำนวนประเภทของผลลัพธ์สำหรับข้อมูลชุดนั้นๆ โดยหน่วยประสาทเทียมในชั้นผลลัพธ์จะรับข้อมูลจากชั้นซ่อนชั้นสุดท้ายมาเป็นข้อมูลเข้าสำหรับชั้นผลลัพธ์ โดยเมื่อรับข้อมูลเข้าแล้ว หน่วยประสาทเทียมในชั้นผลลัพธ์แต่ละหน่วยจะได้ค่าที่ไม่เท่ากัน โดยหน่วยประสาทเทียมที่มีน้ำหนักมากกว่า จะถือว่าผลลัพธ์เป็นประเภทนั้น [10]



ภาพที่ 2.6 ฟังก์ชันการแปลงที่มีการใช้งานทั่วไป[11]

ในกระบวนการเรียนรู้ของโครงข่ายประสาทเทียม เริ่มต้นนั้นชั้นข้อมูลเข้าซึ่งจะมีขนาดเท่ากับจำนวนคุณลักษณะที่สนใจ จะทำการส่งข้อมูลไปยังหน่วยประสาทเทียมชั้นซ่อน โดยในแต่ละหน่วยประสาทเทียมชั้นซ่อนนั้น จะมีตัวแปรค่าถ่วงน้ำหนัก และค่าความเอนเอียง โดยจะทำหน้าที่ถ่วงน้ำหนักข้อมูลนำเข้า ก่อนจะส่งไปคำนวณฟังก์ชันการแปลงเพื่อให้ได้ผลลัพธ์ส่งไปยังหน่วยประสาทเทียมชั้นถัดไปจนกระทั่งไปถึงหน่วยประสาทเทียมชั้นผลลัพธ์ โดยมีฟังก์ชันการแปลงที่นิยมใช้ดังภาพที่ 2.6 โดยขั้นตอนทั้งหมดนี้เรียกว่าการประมวลผลไปข้างหน้า (forward pass)

David E. Rumelhart (1995) ได้นำเสนออัลกอริทึมที่มีชื่อว่า Backpropagation [12] โดยจะทำการคำนวณความผิดของผลลัพธ์ที่ได้จากโครงข่ายประสาทเทียม กับผลลัพธ์ที่แท้จริงออกมาเป็นค่าความผิดพลาด (error) ก่อนที่จะทำการคำนวณย้อนกลับ เพื่อปรับค่าถ่วงน้ำหนักของแต่ละหน่วยประสาทเทียมอีกครั้ง ให้มีความเหมาะสมกับข้อมูลมากขึ้น โดยขั้นตอนนี้เรียกว่า การประมวลผลย้อนกลับ (backward pass) โดยมีความแตกต่างกับการประมวลผลไปข้างหน้าดังภาพที่ 2.7

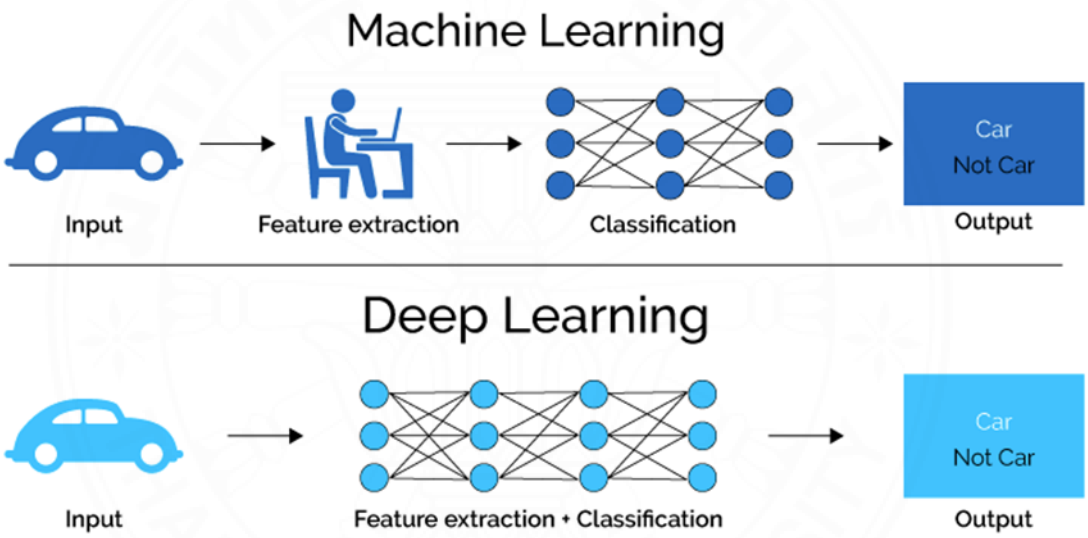


ภาพที่ 2.7 การทำงานของการประมวลผลไปข้างหน้า และการประมวลผลย้อนกลับ[13]

โดยเมื่อทำการประมวลผลไปข้างหน้าและประมวลผลแบบย้อนกลับไปเรื่อยๆ โครงข่ายประสาทเทียมจะมีประสิทธิภาพในการทำนายได้ดีขึ้น จากการปรับค่าถ่วงน้ำหนักของหน่วยประสาทเทียม โดยเมื่อได้ประสิทธิภาพในระดับที่ต้องการแล้วจะสามารถบันทึกค่าถ่วงน้ำหนักในแต่ละหน่วยประสาทเทียมเพื่อนำไปสร้างเป็นโมเดล (model) สำหรับทำนายข้อมูลต่อไปได้

2.3 ทฤษฎีเกี่ยวกับการเรียนรู้เชิงลึก (Deep Learning)

เป็นหนึ่งในประเภทของการเรียนรู้ของเครื่อง โดยวิธีการทำงานของการเรียนรู้เชิงลึกจะเป็นการลอกเลียนแบบโครงข่ายประสาทของสมองมนุษย์ [18] โดยสร้างโครงข่ายประสาทเทียมขึ้นมาโดยจะมีชั้นซ่อนจำนวนหลายชั้น และทำการเรียนรู้จากข้อมูลตัวอย่าง เพื่อสร้างเป็นโมเดลในการทำนายโดยจุดเด่นของการเรียนรู้เชิงลึกคือสามารถสกัดคุณลักษณะของข้อมูลได้เองโดยไม่ต้องพึ่งผู้เชี่ยวชาญ จึงทำให้สามารถประมวลผลภาษาธรรมชาติ การรู้จำ ภาพ เสียง หรือข้อมูลชีวสารสนเทศศาสตร์ได้ [17] โดยจะแตกต่างกับการเรียนรู้ของเครื่อง ดังที่แสดงในภาพที่ 2.8 ที่จะต้องมีผู้เชี่ยวชาญคอยสกัดคุณลักษณะของข้อมูลก่อนที่จะนำข้อมูลไปใช้ในการสร้างโมเดลได้



ภาพที่ 2.8 เปรียบเทียบการทำงานของการเรียนรู้ของเครื่องกับการเรียนรู้เชิงลึก[37]

2.4 โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Network: CNN)

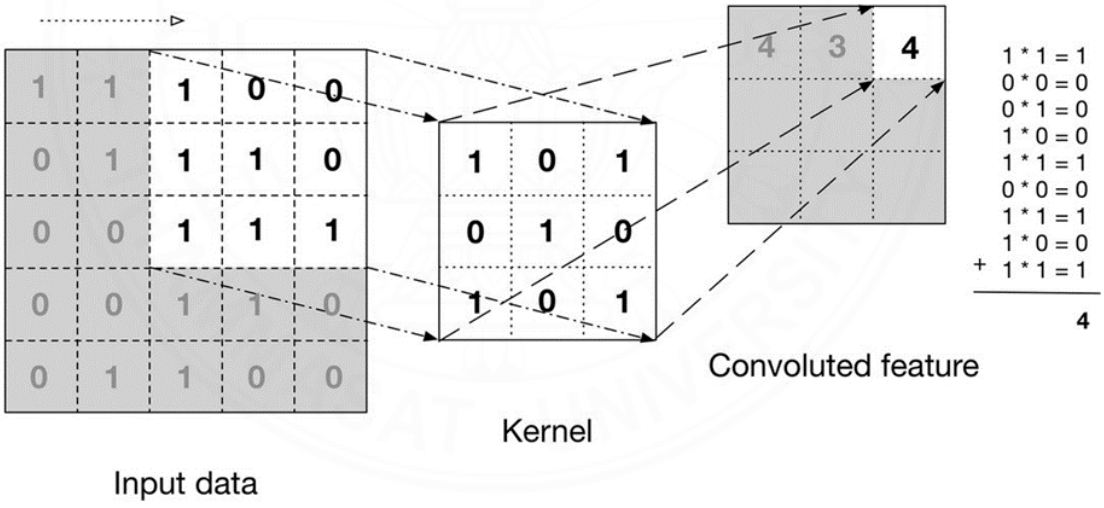
ในการทำงานกับข้อมูลมากกว่า 1 มิตินั้น โดยเฉพาะข้อมูลภาพ ที่มีคุณลักษณะเฉพาะ จึงเป็นเรื่องยากที่จะสกัดคุณลักษณะออกมาจากภาพ ก่อนที่นำไปใช้งานในระบบโครงข่ายประสาทเทียม กระทั่ง Kunihiko Fukushima (1980) ได้คิดค้นกระบวนการในการสกัดคุณลักษณะของภาพแบบอัตโนมัติก่อนที่จะนำคุณลักษณะที่ได้ไปใช้งานในระบบโครงข่ายประสาทเทียมชั้นสังวัตนการ (Convolutional Layer) [16] เป็นกระบวนการสกัดคุณลักษณะของข้อมูลภาพออกมาแบบอัตโนมัติ โดยจะจำลองการมองเห็นของมนุษย์ที่มองพื้นที่เป็นพื้นที่ย่อย ๆ และนำกลุ่มของพื้นที่เหล่านั้นมาประสานรวมกันเพื่อจำแนกว่าสิ่งที่มองเห็นอยู่คืออะไร โดยในการมองพื้นที่ย่อยของมนุษย์จะมีการ

จำแนก คุณลักษณะ (feature) ในพื้นที่นั้นๆ เช่น ลายเส้น หรือการตัดกันของสี โดยในชั้นสังวัตนการจะทำการสกัดคุณลักษณะของภาพออกมาเป็นผังคุณลักษณะ (feature map) ออกมาตามจำนวนตัวกรอง (filter) ที่ใช้ ก่อนจะนำเข้าสู่กระบวนการสุ่มตัวอย่าง (pooling) เพื่อส่งต่อข้อมูลไปยังโครงข่ายประสาทเทียม [19] โดยมีตัวอย่างการทำงานของโครงข่ายประสาทเทียมแบบสังวัตนาการดังที่แสดงในภาพที่ 2.11

2.4.1 การสกัดคุณลักษณะ (Feature Extraction)

ชั้นสังวัตนการสกัดคุณลักษณะเพื่อนำข้อมูลไปใช้ ประกอบด้วย 2 ส่วน คือ การคอนโวลูชัน และการสุ่มตัวอย่าง

การทำคอนโวลูชัน (Convolution) การคำนวนคอนโวลูชันจะใช้ตัวกรอง (kernel) ที่สามารถช่วยดึงคุณลักษณะที่ใช้ในการรู้จำวัตถุออกมา โดยในแต่ละตัวกรองจะสามารถดึงคุณลักษณะที่สนใจออกมาได้เพียงหนึ่งอย่าง จึงจำเป็นจะต้องใช้ตัวกรองหลายตัวเพื่อใช้หาคุณลักษณะในพื้นที่ของภาพได้อย่างครบถ้วน



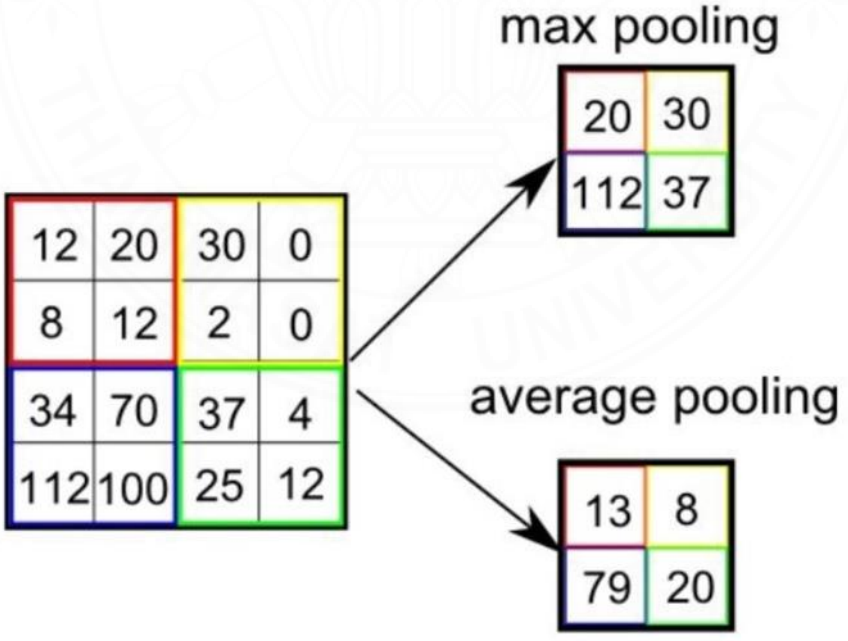
ภาพที่ 2.9 ตัวอย่างการคำนวณคอนโวลูชัน[20]

ในกระบวนการทำคอนโวลูชันจะใช้เมทริกซ์ (matrix) ตัวกรองวางทาบไปบนข้อมูลภาพเริ่มจากด้านซ้ายบนดังที่แสดงในภาพที่ 2.9 จากนั้นจะทำการคูณเมทริกซ์เทียบกับภาพตามขนาดของตัวกรองและรวมผลลัพธ์เข้าด้วยกันเพื่อเก็บเป็นตารางผลลัพธ์ที่แสดงถึงความสอดคล้องของคุณสมบัติของตัวกรองในตำแหน่งบนภาพ โดยเมื่อได้ค่าความสัมพันธ์เทียบกับตัวกรองในแต่ละจุดแล้ว จะทำการเลื่อนตัวกรองไปยังพิกเซลถัดไปทางขวาและ

คำนวนแบบเดียวกันจนสุดด้านขวาแล้วจึงเลื่อนตัวกรองลงมาและเริ่มคำนวนจากฝั่งซ้ายโดยจะทำจนสามารถคำนวนความสัมพันธ์กับตัวกรองได้ครบทั้งภาพ

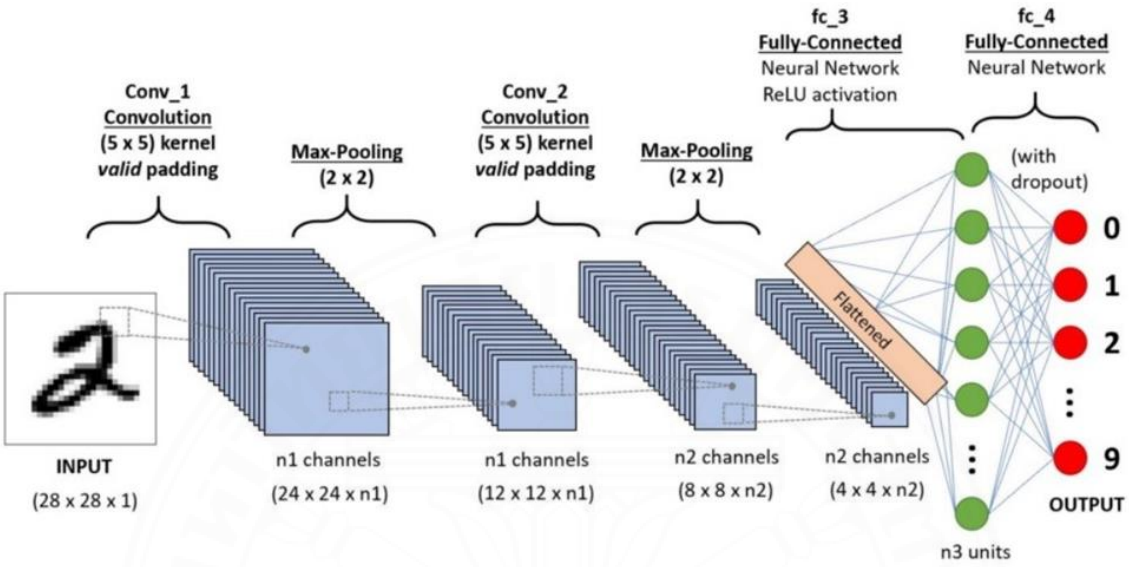
หลังจากที่ทำการคอนโวลูชัน ผลลัพธ์ที่ได้จะเป็นผังคุณลักษณะออกมาตามจำนวนตัวกรองที่ใช้ ซึ่งยิ่งมีจำนวนตัวกรองมากก็จะสามารถสลัดคุณลักษณะของข้อมูลได้หลากหลายรูปแบบยิ่งขึ้น

กระบวนการสุ่มตัวอย่าง (Pooling) กระบวนการสุ่มตัวอย่างเป็นเทคนิคที่ช่วยลดขนาดของข้อมูล โดยรักษาข้อมูลไว้ให้ใกล้เคียงกับค่าเดิมมากที่สุด ในกระบวนการสุ่มตัวอย่างนั้นจะทำการลดจำนวนและความซ้ำซ้อนของข้อมูลที่เกิดขึ้นจากการทำคอนโวลูชัน โดยในชั้นของการสุ่มตัวอย่างจะนิยมใช้วิธีการหาค่าสูงสุด หรือหาค่าเฉลี่ย โดยวิธีการสุ่มตัวอย่างนั้นจะขึ้นอยู่กับจำนวนของพิกเซลผลลัพธ์ที่เราต้องการ และเราต้องการให้ข้อมูลผลลัพธ์จากการสุ่มตัวอย่างมีค่าเท่ากับ 2x2 พิกเซล เราจะต้องทำการแบ่งกลุ่มของข้อมูลเข้า เช่น หากข้อมูลเข้ามีขนาด 4x4 พิกเซล จะสามารถแบ่งเป็น 2x2 ซึ่งสามารถแบ่งออกได้ 4 ส่วน ดังภาพที่ 2.10 ข้อมูลแต่ละส่วนที่เป็นกลุ่มเดียวกันจะถูกคำนวนค่าสูงสุด และค่าเฉลี่ยร่วมกันตามแต่วิธีที่ผู้ใช้เลือก และนำไปใส่ข้อมูลผลลัพธ์ใหม่ตามตำแหน่งของแต่ละกลุ่มข้อมูลที่ถูกแบ่ง



ภาพที่ 2.10 ตัวอย่างการคำนวณการสุ่มตัวอย่างจากผังคุณลักษณะ[20]

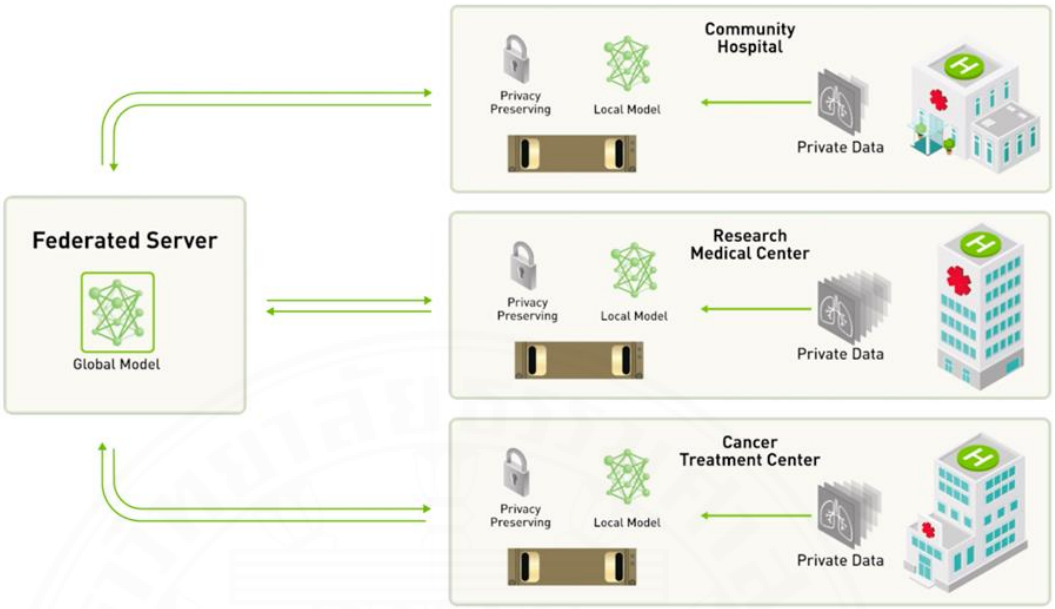
หลังจากเข้าสู่กระบวนการสุ่มตัวอย่าง จะได้ผลลัพธ์ออกมาเป็นคุณลักษณะของภาพที่สามารถนำไปใช้เป็นข้อมูลเข้าสำหรับโครงข่ายประสาทเทียมได้



ภาพที่ 2.11 ตัวอย่างการทำงานของโครงข่ายประสาทเทียมแบบสังวัตนากา [21]

2.5 การเรียนรู้แบบสหพันธ์ (Federated Learning)

การประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึกเพื่อหาความสัมพันธ์ระหว่างข้อมูลเข้าและข้อมูลผลลัพธ์อย่างมีประสิทธิภาพจำเป็นจะต้องใช้ข้อมูลจำนวนมาก แต่เนื่องจากในบางสถานการณ์ข้อมูลอาจถูกกระจายอยู่ในหลายๆแห่ง และไม่สามารถแบ่งปันซึ่งกันและกันได้ เนื่องจากติดปัญหาด้านความเป็นส่วนตัว (privacy) ตัวอย่างเช่น ข้อมูลผู้ป่วย ในแต่ละโรงพยาบาล เป็นต้น กระทั่งในปี 2017 Google Research ได้นำเสนอเทคนิคการเรียนรู้แบบสหพันธ์ [22] โดยเทคนิคนี้จะช่วยให้สามารถ ใช้งานข้อมูลที่หน่วยงานต่างๆ เก็บเอาไว้ได้ โดยที่ไม่กระทบ ความเป็นส่วนตัวของข้อมูล เนื่องจากการเรียนรู้แบบสหพันธ์จะทำการแลกเปลี่ยนเพียงพารามิเตอร์(parameter) ของโมเดลเฉพาะของแต่ละเครื่องในสหพันธ์ (local model) เพื่อปรับปรุงการทำงานของโมเดลหลัก (global model) เท่านั้น โดยที่ไม่มีการเคลื่อนย้ายข้อมูลจริง (raw data) ออกจากเครื่องนั้นๆ ดังที่แสดงในภาพที่ 2.12



ภาพที่ 2.12 ภาพกระบวนการทำงานของการเรียนรู้แบบสหพันธ์[23]

2.5.1 ขั้นตอนการทำงานของการเรียนรู้แบบสหพันธ์

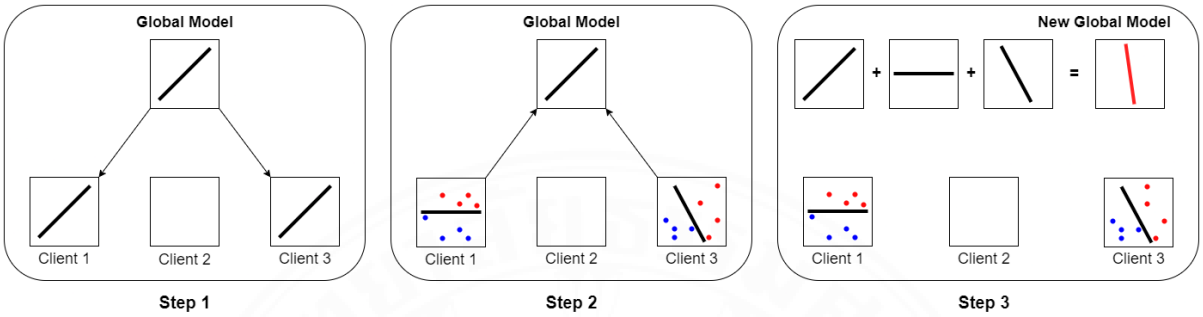
การทำงานของการเรียนรู้แบบสหพันธ์จะประกอบด้วยสองส่วน คือ

- เครื่องเซิฟเวอร์ (server)
- เครื่องในสหพันธ์

โดยขั้นตอนการทำงานของการเรียนรู้แบบสหพันธ์แบ่งออกเป็น 3 ขั้นตอนหลัก ดังที่แสดงในภาพที่ 2.13 คือ

1. เครื่องเซิฟเวอร์จะทำการกำหนดรูปแบบของโมเดลหลักที่จะใช้ในการประมวลผลในครั้งแรก ก่อนที่จะเลือกเครื่องในสหพันธ์ที่จะทำงานในรอบนี้ เพื่อสถาปนาการเชื่อมต่อและส่งโมเดลหลักไปยังเครื่องในสหพันธ์
2. เครื่องในสหพันธ์ได้รับโมเดลหลักจากเครื่องเซิฟเวอร์และทำการใช้โมเดลเรียนรู้กับข้อมูลภายในเครื่อง ก่อนที่จะส่งโมเดลผลลัพธ์ที่ได้จากการเรียนรู้ในรอบนั้นๆกลับไปยังเครื่องเซิฟเวอร์
3. เครื่องเซิฟเวอร์ได้รับโมเดลผลลัพทธ์จากแต่ละเครื่องในสหพันธ์และทำการควบรวมเข้ากับโมเดลหลัก เพื่อให้ได้เป็นโมเดลหลักตัวใหม่ และใช้แทนโมเดลหลักตัวเดิม

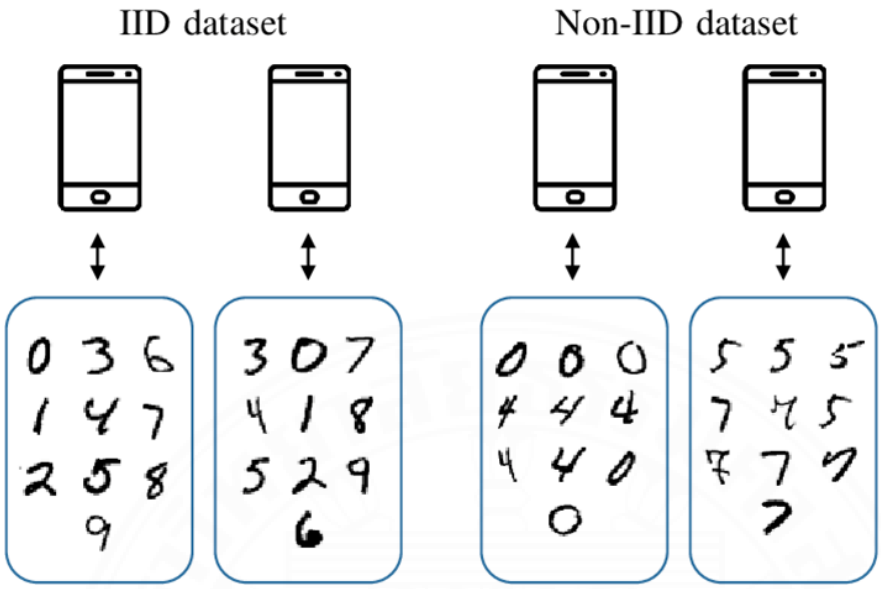
โดยจะทำวนซ้ำในขั้นตอนที่ 1-3 จนกว่าจะได้ผลลัพธ์ของโมเดลหลักที่เป็นที่พอใจจึงจะใช้โมเดลนั้นเป็นโมเดลสำหรับการอนุมาน (Inference)



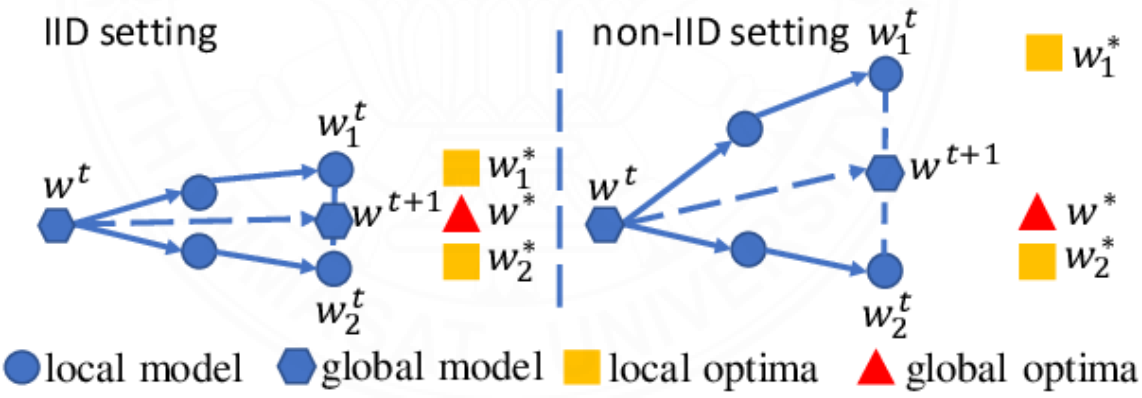
ภาพที่ 2.13 ตัวอย่างขั้นตอนการเรียนรู้ของกระบวนการเรียนรู้แบบสหพันธ์

2.6 ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกัน และไม่เป็นอิสระต่อกัน (Non-Independent and Identically Distributed: Non-IID)

ในการเรียนรู้แบบสหพันธ์หากข้อมูลที่เก็บไว้ในแต่ละแหล่งข้อมูลมีความไม่เป็นอิสระต่อกัน หรือมีการกระจายของข้อมูล ที่ไม่เหมือนกันในแต่ละแหล่งข้อมูล ดังภาพที่ 2.14 ฝั่งขวา ที่ข้อมูลบางประเภทถูกจัดเก็บไว้บนเครื่องเพียงบางเครื่อง หรือมีบางเครื่องที่มีข้อมูลไม่ครบทุกประเภท ซึ่งส่งผลให้ประสิทธิภาพ ของโมเดลหลัก ที่สร้างขึ้นจากการเรียนรู้แบบสหพันธ์ โดยข้อมูลลักษณะนี้นั้น มีประสิทธิภาพที่ลดลง หากข้อมูลในแหล่งเก็บข้อมูลมีการกระจายที่แตกต่างกันหรือไม่เป็นอิสระต่อกัน จะทำให้จุดสมดุลภายในที่ดีที่สุด (local optima) ของแต่ละโมเดลมีความแตกต่างกันสูง และเมื่อนำมารวมกันเป็นโมเดลหลักจะได้โมเดลหลักที่คำนวณได้จุดสมดุลที่ดีที่สุดห่างจากจุดสมดุลที่ดีที่สุดที่เป็นไปได้ของโมเดล (global optima) ดังที่แสดงในภาพที่ 2.15 ส่งผลให้ประสิทธิภาพของโมเดลหลักต่ำกว่าที่ควรจะเป็น



ภาพที่ 2.14 ตัวอย่างการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน[28]



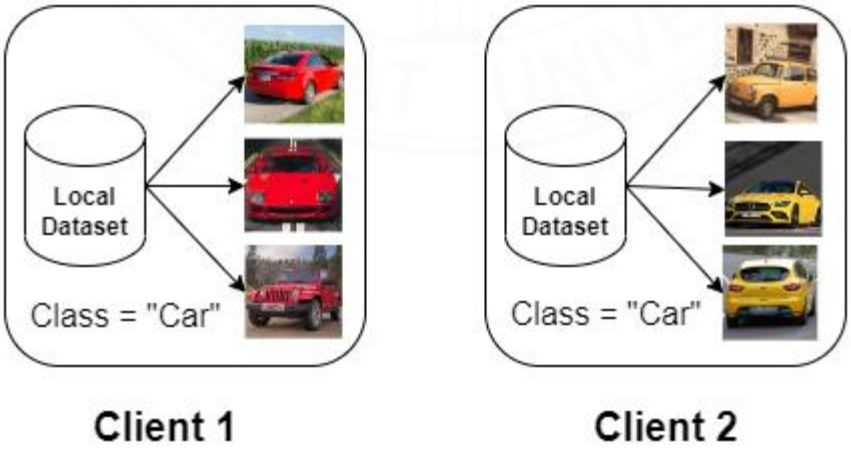
ภาพที่ 2.15 การหาจุดสมดุลในการเรียนรู้แบบสหพันธ์ด้วยข้อมูลที่มีการกระจายตัวสมบูรณ์และเป็นอิสระต่อกัน เทียบกับข้อมูลที่กระจายตัวไม่สมบูรณ์และไม่เป็นอิสระต่อกัน[25]

2.6.1 ประเภทของการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

- การกระจายที่ไม่เหมือนกันในด้านคุณลักษณะของข้อมูล (Feature Distribution Skew) [24] เป็นการกระจายคุณลักษณะของข้อมูลที่ไม่ครบถ้วน กล่าวคือข้อมูลประเภทเดียวกันในบางแหล่งเก็บข้อมูลอาจจะมีคุณลักษณะที่ไม่ครบ หรือแตกต่างจากคุณลักษณะของข้อมูลประเภทเดียวกันในที่อยู่ในแหล่งเก็บข้อมูลอื่นๆ ส่งผลให้โมเดลเฉพาะหน่วยงานในบางเครื่องในสหพันธ์จะมีจุดสมดุลที่แตกต่างจากเครื่องอื่นๆ โดยสามารถแบ่งออกได้เป็นสองระดับ คือ

    1. ความไม่เหมือนกันในด้านคุณลักษณะของข้อมูลบางส่วน โดยระดับนี้ข้อมูลในเครื่องสหพันธ์ จะมีเพียงบางคุณลักษณะของข้อมูลที่เหมือนกันกับข้อมูลในเครื่องอื่น
    2. ความไม่เหมือนกันในด้านคุณลักษณะของข้อมูลทั้งหมด โดยข้อมูลประเภทเดียวกันที่อยู่ในเครื่องในสหพันธ์จะมีคุณลักษณะของข้อมูลที่แตกต่างกันและไม่ซ้ำกับข้อมูลประเภทเดียวกันในเครื่องในสหพันธ์อื่นๆ

ตัวอย่างเช่น การจัดเก็บข้อมูลภาพรถยนต์ดังภาพที่ 2.16 ซึ่งแสดงให้เห็นถึงการกระจายที่ไม่เหมือนกันในด้านคุณลักษณะของข้อมูล โดยข้อมูลในเครื่องทั้งสองเป็นภาพรถยนต์เหมือนกัน แต่แตกต่างที่ในเครื่องแรกจะเก็บเฉพาะภาพรถยนต์สีแดง และเครื่องที่สองเก็บเฉพาะภาพรถยนต์สีเหลือง ซึ่งทำให้ทั้งสองเครื่องขาดคุณลักษณะในด้านสีของรถยนต์ที่มีไม่ครบถ้วน

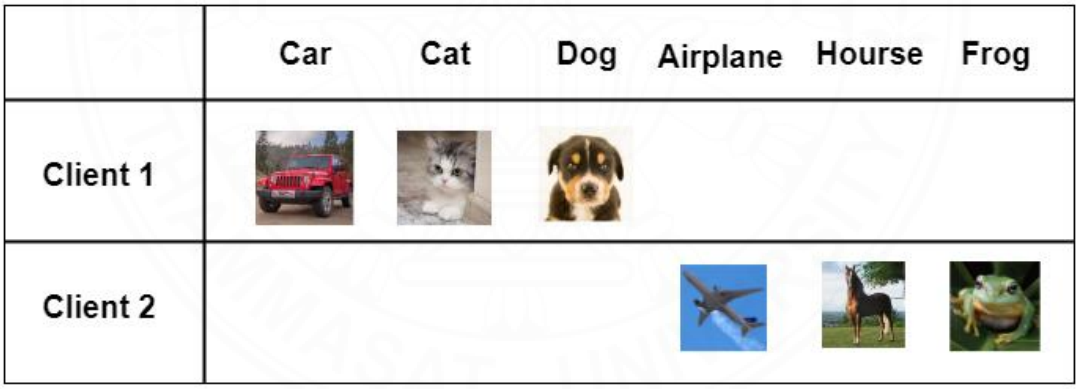


ภาพที่ 2.16 ภาพการกระจายข้อมูลที่ไม่เหมือนกันในด้านคุณลักษณะของข้อมูล

- การกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูล (Label Distribution Skew) [24] เป็นการกระจายข้อมูลที่ข้อมูลในเครื่องในสหพันธ์บางแห่ง มีประเภทของข้อมูลที่ไม่ครบถ้วน ทำให้โมเดลเฉพาะหน่วยงานไม่สามารถเรียนรู้ข้อมูลประเภทนั้นๆ ได้ ส่งผลให้โมเดลเฉพาะเครื่องนั้นๆ มีจุดสมดุลต่างจากที่ควรจะเป็น โดยกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูลนั้นแบ่งเป็นสามระดับ คือ

1. การกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูลในบางเครื่อง โดยจะมีเพียงแค่บางเครื่องในสหพันธ์เท่านั้นที่มีข้อมูลไม่ครบทุกประเภท
2. การกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูลในทุกประเภท โดยข้อมูลแต่ละประเภทจะถูกเก็บไว้ที่เครื่องในสหพันธ์เพียงเครื่องเดียว
3. การกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูลในทุกเครื่อง โดยในแต่ละเครื่องในสหพันธ์จะมีข้อมูลเพียงประเภทเดียว

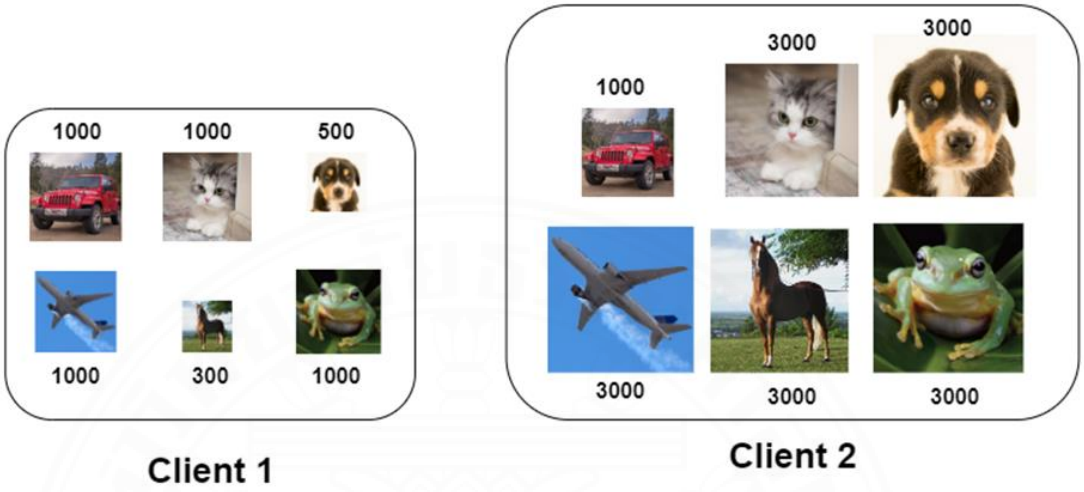
ตัวอย่างเช่น การจัดเก็บภาพข้อมูลภาพวัตถุประเภทต่างๆ ดังภาพที่ 2.17 แสดงให้เห็นถึงการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันในด้านประเภทของข้อมูล โดยทั้งสองเครื่องนั้นมีประเภทของข้อมูลไม่ครบทุกประเภทจากข้อมูลทั้งหมด



ภาพที่ 2.17 ภาพการกระจายข้อมูลที่ไม่เหมือนกันในด้านประเภทของข้อมูลในทุกประเภท

- การกระจายที่ไม่เหมือนกันในด้านจำนวนข้อมูล (Quantity Skew) [24] เป็นการกระจายข้อมูลที่ส่งผลให้จำนวนข้อมูลในข้อมูลทั้งหมดในแต่ละเครื่องในสหพันธ์นั้นมีจำนวนแตกต่างกันเป็นอย่างมาก หรือข้อมูลแต่ละประเภทในเครื่องในสหพันธ์มีจำนวนไม่เท่ากัน ดังที่แสดงในภาพที่ 2.18 ส่งผลให้ประสิทธิภาพในการเรียนรู้ของทั้งสองโมเดล

ไม่เท่ากัน เมื่อนำไปสร้างเป็นโมเดลหลักจึงมีแนวโน้มที่จะถูกลดประสิทธิภาพลงจากโมเดลที่สร้างจากเครื่องในสหพันธ์ที่มีข้อมูลน้อย



ภาพที่ 2.18 ภาพการกระจายข้อมูลที่ไม่เหมือนกันในด้านจำนวนข้อมูล

2.7 การแก้ไขปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

ในการแก้ไขปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันสามารถแก้หลายรูปแบบ เช่น การแก้ด้วยข้อมูล การแก้ด้วยอัลกอริทึม หรือการแก้ที่ระบบที่ใช้ในการเรียนรู้ [24] โดยในงานวิจัยนี้จะมุ่งเน้นที่การแก้ปัญหาโดยใช้ข้อมูล โดยการใช้ข้อมูลจะแบ่งออกเป็นสองหัวข้อ ได้แก่

2.7.1 การแบ่งปันข้อมูล (Data Sharing)

ในปี 2018 Yue Zhao [43] ได้ทดลองเรียนรู้ข้อมูลประเภทที่มีการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยการเรียนรู้แบบสหพันธ์ และพบว่าประสิทธิภาพของโมเดลนั้นค่อนข้างแย่กับข้อมูลประเภทดังกล่าว โดยเขาได้นำเสนอเทคนิคการแบ่งปันข้อมูลโดยแก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระด้วยการแบ่งปันข้อมูลจากเครื่องในสหพันธ์แต่ละเครื่องมาเก็บไว้บนเครื่องเซิฟเวอร์ และทำการกระจายข้อมูลเหล่านั้นไปยังเครื่องในสหพันธ์ที่ขาดข้อมูลเพื่อเพิ่มข้อมูลที่ขาดไปในแต่ละเครื่องในสหพันธ์ โดยผลการ

ทดลองสามารถเพิ่มประสิทธิภาพของโมเดลหลักให้มีประสิทธิภาพใกล้เคียงกับการเรียนรู้จากข้อมูลที่มีการกระจายที่สมดุล โดยแบ่งปันข้อมูลไปยังเซิฟเวอร์เพียง 5% ของข้อมูลทั้งหมด

แต่วิธีการดังกล่าวมีข้อบกพร่องคือ การละเมิดความเป็นส่วนตัวของข้อมูลโดยการใช้ข้อมูลจริง (raw data) จากเครื่องในสหพันธ์ส่งออกไปยังเซิฟเวอร์ ซึ่งขัดต่อวัตถุประสงค์ของการเรียนรู้แบบสหพันธ์ที่ต้องการรักษาความเป็นส่วนตัวของข้อมูล

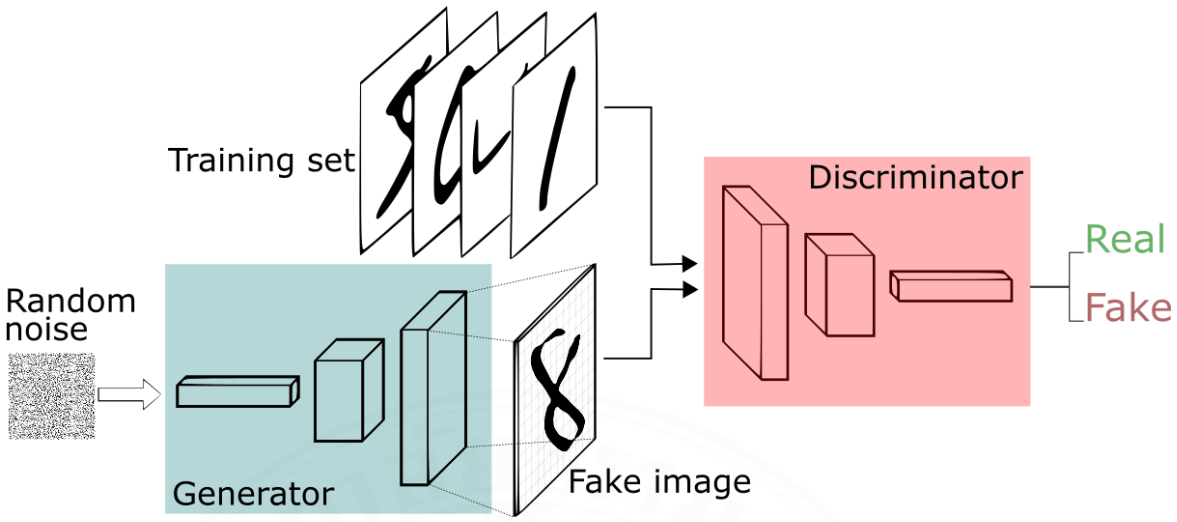
2.7.2 การเพิ่มประสิทธิภาพของข้อมูล (Data Enhancement)

ในปี 2018 Seong-Lyun [40] มุ่งเน้นไปที่การเพิ่มจำนวนข้อมูล (data augmentation) เพื่อแก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน โดยเติมเต็มจำนวนข้อมูลที่ขาดหายไปในแต่ละเครื่องในสหพันธ์และลดปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมด (communication cost) และจะใช้ข้อมูลบางส่วนจากเครื่องในสหพันธ์ส่งไปยังเซิฟเวอร์เพื่อสร้างแกนส์บนเครื่องเซิฟเวอร์ จากนั้นจึงส่งโมเดลผู้สร้าง (generator model) กลับไปยังเครื่องในสหพันธ์ที่ขาดข้อมูล เพื่อใช้ในการสร้างข้อมูลสังเคราะห์เพื่อเติมเต็มข้อมูลที่ขาดหายเพื่อแก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

ซึ่งวิธีการดังกล่าวยังมีการใช้ข้อมูลจริงส่งไปยังเครื่องเซิฟเวอร์เพื่อใช้ในการสร้างโมเดลผู้สร้าง ซึ่งการละเมิดความเป็นส่วนตัวของข้อมูลขัดต่อคุณสมบัติของการเรียนรู้แบบสหพันธ์

2.8 แกนส์ (Generative Adversarial Networks: GANs)

ในปี 2014 Ian J. Goodfellow [27] ได้นำเสนอเทคนิคแกนส์ (Generative Adversarial Nets : GANs) โดยแกนส์จะประกอบด้วย โมเดลผู้สร้าง (generator model) ทำหน้าที่เรียนรู้และสร้างผลงานใหม่จากคุณลักษณะของข้อมูล และโมเดลผู้ตรวจสอบ (discriminator model) ทำหน้าที่ตรวจสอบว่าข้อมูลที่ส่งออกมาจากโมเดลผู้สร้างเป็นข้อมูลจริงหรือข้อมูลที่ถูกสร้างขึ้นมา [7] ดังที่แสดงในภาพที่ 2.19 โดยจุดประสงค์ของการทำงานคือสร้างโมเดลผู้สร้าง ที่มีความสามารถในการสร้างข้อมูลขึ้นมาใหม่ ที่สามารถหลอกโมเดลผู้ตรวจสอบได้ และด้วยการแข่งขันของทั้งสองโมเดลท้ายที่สุดแล้วจะได้โมเดลผู้สร้างที่สามารถสร้างข้อมูลขึ้นมาใหม่จากคุณลักษณะเดิมที่เรียนรู้จากข้อมูลชุดเดิม ด้วยวิธีการดังกล่าวเราจึงสามารถใช้ข้อมูลสังเคราะห์ที่สร้างขึ้นจากแกนส์ในการแก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันได้มีประสิทธิภาพมากกว่าเทคนิคการเพิ่มข้อมูลทั่วไป

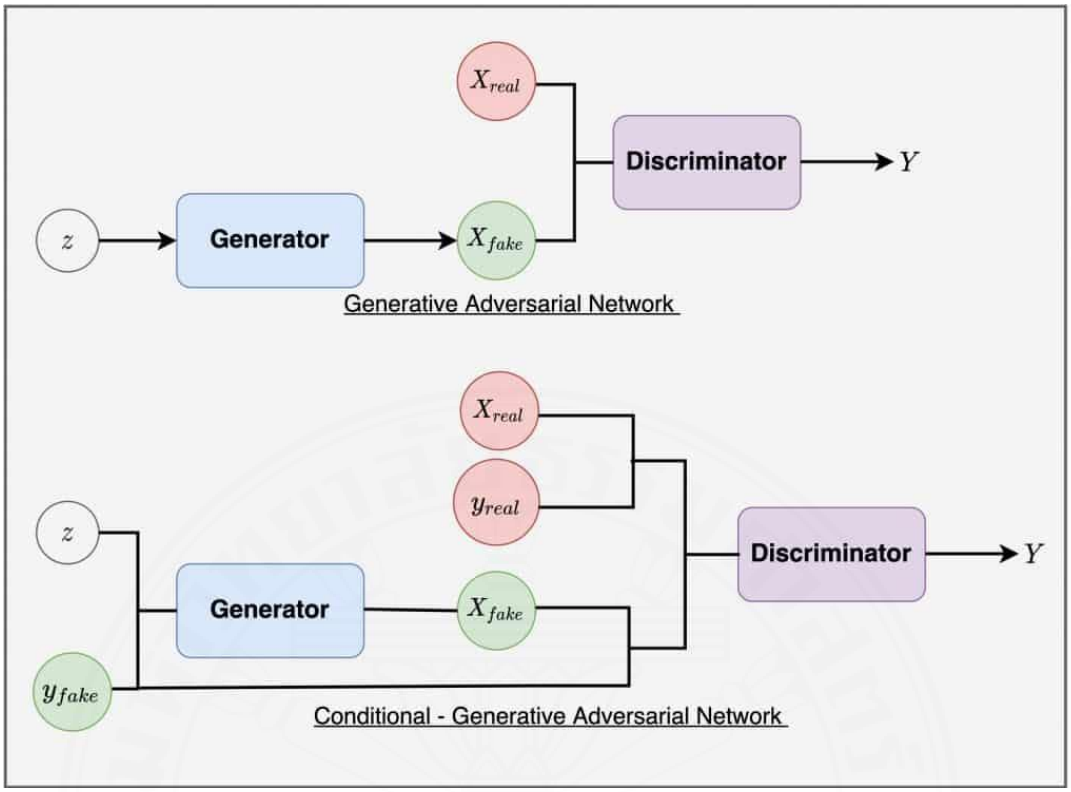


ภาพที่ 2.19 ภาพการทำงานของการเรียนรู้ของโมเดลผู้สร้างและโมเดลผู้ตรวจสอบ[6]

2.8.1 คอนดิชันนอลแกนส์ (Conditional Generative Adversarial Networks: CGANs)

ในปี 2014 Mehdi Mirza ได้นำเสนอเทคนิคคอนดิชันนอลแกนส์ [49] ซึ่งเป็นโมเดลการเรียนรู้เชิงลึกที่ทำงานในลักษณะเดียวกันกับแกนส์ แต่มีความพิเศษในการที่จะสามารถกำหนดประเภทของภาพผลลัพท์ได้จากข้อมูลนำเข้า เพื่อให้ได้ประเภทของข้อมูลภาพที่ต้องการ

โมเดลแกนส์แบบทั่วไปนั้นจะมีข้อมูลนำเข้าเป็นค่ารบกวน (noise) ใดๆ ส่งเข้าไปยังโมเดลผู้สร้าง เพื่อสร้างข้อมูลสังเคราะห์ออกมา แล้วนำไปเทรนโมเดลผู้ตรวจสอบต่อไป แต่ในรูปแบบชองคอนดิชันนอลแกนส์นั้น ข้อมูลเข้าของโมเดลจะสามารถมีได้มากกว่า 1 รูปแบบ โดยสามารถระบุเป็นประเภทของข้อมูลที่ต้องการสร้างได้ ดังภาพที่ 2.20 โดยสามารถป้อนค่า z นอกเหนือจากค่ารบกวนที่ใช้ในการสร้างข้อมูลสังเคราะห์ เพื่อให้สามารถสร้างข้อมูลสังเคราะห์ในประเภทที่ต้องการได้ ตัวเลือกของข้อมูลนำเข้านั้นจะต้องสอดคล้องกับข้อมูลจริงที่ใช้ในการเรียนรู้ของโมเดลผู้ตรวจสอบ เพื่อให้โมเดลสามารถเรียนรู้คุณลักษณะของข้อมูลประเภทต่างๆได้อย่างเป็นอิสระต่อกัน



ภาพที่ 2.20 ภาพการทำงานของการเรียนรู้ของโมเดลผู้สร้างและโมเดลผู้ตรวจสอบเปรียบเทียบระหว่างแกนส์และคอนดิชันนอลแกนส์[49]

2.8.2 การวัดประสิทธิภาพของภาพที่ถูกสร้างขึ้นจากแกนส์ด้วยระบบคอมพิวเตอร์

การวัดประสิทธิภาพการทำงานของโมเดลผู้สร้าง ว่าสร้างภาพผลลัพธ์ออกมาได้สมจริงมากแค่ไหนเป็นเรื่องที่ท้าทาย หากใช้มนุษย์เป็นคนตัดสินจะต้องใช้แรงงานและเวลามาก อีกทั้งมนุษย์แต่ละคนก็ไม่สามารถตัดสินและให้คะแนนความสมจริงของภาพที่ถูกสร้างขึ้นได้ด้วยมาตราฐานเดียวกัน นักวิจัยจึงพยายามค้นหาวิธีการในการวัดผลประสิทธิภาพของโมเดลผู้สร้างโดยการประเมินคุณภาพของภาพที่ถูกสร้างขึ้นด้วยค่าวัดคุณภาพต่างๆ ดังนี้

- Inception Score (IS) ในปี 2016 Tim Salimans [32] ได้คิดค้นวิธีการวัดคุณภาพของข้อมูลภาพที่ถูกสร้างขึ้นโดยแกนส์ โดยจะใช้โมเดลอื่นเป็นตัวตัดสินให้คะแนน ในงานวิจัยดังกล่าวผู้วิจัยได้เลือกใช้โมเดล InceptionV3 ซึ่งเป็นโมเดลตรวจจับวัตถุที่เรียนรู้จากชุดข้อมูล ILSVRC 2012 ครอบคลุมวัตถุมากถึง 1,000 ประเภท โดยจะให้คะแนนจากการที่โมเดล InceptionV3 สามารถรู้จำข้อมูลภาพที่ถูกสร้างขึ้นได้มากแค่ไหน.อย่างไรก็ตาม

คะแนน IS นั้นไม่ได้เป็นการบอกว่าข้อมูลที่ถูกสร้างขึ้นมีความเหมือนกับต้นฉบับมากแค่ไหน แต่เป็นการบอกว่าข้อมูลที่ถูกสร้างขึ้นนั้นเป็นที่จำได้โดยโมเดล InceptionV3 มากแค่ไหน [30] โดยยิ่งคะแนน IS สูงยิ่งแสดงถึงการเป็นที่จดจำได้ของโมเดล InceptionV3 สำหรับชุดข้อมูลนั้นๆ ดังนั้นหากประเภทของข้อมูลที่เราสนใจนั้นไม่มีอยู่ในชุดข้อมูล ILSVRC 2012 ก็จะไม่สามารถวัดประสิทธิภาพของข้อมูลประเภทนั้นด้วยการให้คะแนนแบบ IS ได้

- Frechet Inception Distance (FID) ในปี 2017 Martin Heuselo [33] ได้ปรับปรุงวิธีการวัดคุณภาพของภาพสังเคราะห์จากแกนส์ ด้วยเทคนิคใหม่ที่เรียกว่า FID เพื่อปรับปรุงวิธีวัดคะแนน IS โดยใช้การเปรียบเทียบความเหมือนกับข้อมูลต้นฉบับโดยตรง [31] วิธีการวัดคะแนน FID ยังคงใช้โมเดล InceptionV3 เหมือนเดิม แต่ใช้เพียงแค่ชั้นสุ่มตัวอย่าง(pooling) ชั้นสุดท้าย ด้วยความที่ชั้นนี้จะมีหน่วยประสาทเทียมที่มีสามารถแยกคุณสมบัติเฉพาะของข้อมูลคอมพิวเตอร์วิทัศน์ (computer vision) ได้ FID จะใช้โมเดล InceptionV3 ที่สกัดคุณลักษณะ (feature extraction) ออกมาจากข้อมูล 1000 ประเภท และใช้คุณลักษณะเหล่านั้นในการเปรียบเทียบข้อมูลสองชุดเพื่อทำการเปรียบเทียบความคล้ายและให้คะแนน

2.9 เปรียบเทียบงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่นำเสนอนี้ เป็นการแก้ไขปัญหาความไม่เหมือนกันและไม่เป็นอิสระต่อกันบนกระบวนการเรียนรู้แบบสหพันธ์ด้วยการใช้แกนส์ โดยมีงานวิจัยที่เกี่ยวข้องคือ SDA-FL [15] ซึ่งนำเสนอการแก้ปัญหาเดียวกันด้วยการใช้แกนส์เช่นเดียวกัน โดยในงานวิจัยดังกล่าว จะใช้โมเดลแกนส์เรียนรู้ข้อมูลในแต่ละเครื่องในสหพันธ์ โดยโมเดลแกนส์แต่ละโมเดลนั้นจะถูกเรียนรู้กับข้อมูลเพียงแหล่งเดียว ก่อนที่จะใช้โมเดลนั้นสร้างข้อมูลสังเคราะห์ขึ้นมา เพื่อใช้แทนข้อมูลจริงในกระบวนการเรียนรู้แบบสหพันธ์ ซึ่งการเรียนรู้ของโมเดลแกนส์บนชุดข้อมูลเดียวนั้น ส่งผลให้ข้อมูลสังเคราะห์ที่ถูกสร้างขึ้นมีความเหมือนกับแหล่งข้อมูลที่ถูกใช้ในการเรียนรู้มาก ประกอบกับการที่ทราบว่าข้อมูลถูกเทรนขึ้นจากแหล่งข้อมูลเดียว จึงสามารถทำให้เกิดการสืบสาวไปหาแหล่งต้นทางของข้อมูลได้ จนอาจก่อให้เกิดการละเมิดความเป็นส่วนตัวของข้อมูลได้

ตารางที่ 2.1 ตารางเปรียบเทียบความแตกต่างของงานวิจัยที่เกี่ยวข้อง

หัวข้อ	SDA-FL	งานวิจัยที่นำเสนอ
ใช้ข้อมูลจริงในกระบวนการเรียนรู้แบบสหพันธ์	ไม่	ใช่
ใช้เทคนิคการเพิ่มข้อมูล	ไม่	ใช่
การรวมศูนย์ข้อมูล	ใช่	ไม่
การสร้างและจำแนกข้อมูลสังเคราะห์	GANs + Pseudo Labelling	CGANs
จำนวนโมเดลแกนส์	เท่ากับจำนวนเครื่อง	1
ลำดับการแก้ปัญหา non-iid	ในระหว่างเทรน FL	ก่อนเทรน FL

เพื่อหลีกเลี่ยงการเกิดปัญหาการละเมิดความเป็นส่วนตัวของข้อมูลเนื่องจากข้อมูลสังเคราะห์ที่สร้างจากโมเดลแกนส์เรียนรู้ข้อมูลจากเพียงแหล่งเดียว สามารถถูกสืบหาแหล่งต้นตอของข้อมูลได้ งานวิจัยนี้จึงออกแบบให้โมเดลแกนส์เรียนรู้กับข้อมูลหลายแหล่ง ก่อนจะสร้างข้อมูลสังเคราะห์ออกมา โดยทำการใช้โมเดลคอนดิชันนอลแกนส์เพื่อให้สามารถระบุประเภทของข้อมูลที่ต้องการสร้างได้อย่างแม่นยำ และการใช้ข้อมูลสังเคราะห์ร่วมกับข้อมูลเดิม ส่งผลให้ประสิทธิภาพของโมเดลสหพันธ์ดีขึ้น เนื่องจากมีจำนวนข้อมูลที่มากขึ้น

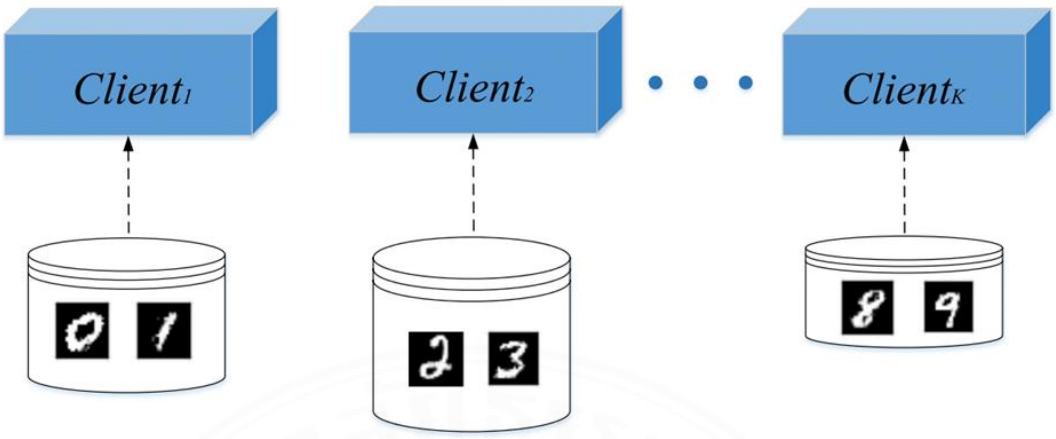
บทที่ 3
วิธีการวิจัย

3.1 ปัญหาวิจัย

งานวิจัยนี้มุ่งเน้นแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลบนการเรียนรู้แบบสหพันธ์ ที่มีจำนวนเครื่องในสหพันธ์ N เครื่อง ซึ่งมีรูปแบบการกระจายของข้อมูลที่ไม่เหมือนกันในด้านประเภทของข้อมูล (label distribution skew) โดยจำลองให้แต่ละเครื่องในสหพันธ์มีจำนวนประเภทของข้อมูลที่ 2 (ดังภาพที่ 3.1) หรือ 5 ประเภท โดยข้อมูลแต่ละประเภทสามารถทับซ้อนกับได้ระหว่างเครื่องในสหพันธ์อื่นๆ แต่ข้อมูลประเภทใดๆ จะต้องถูกจัดเก็บอยู่อย่างน้อยในหนึ่งเครื่องในสหพันธ์ และจำนวนข้อมูลแต่ละประเภทบนเครื่องในสหพันธ์จะมีจำนวนที่ต่างกัน 20 – 40 % โดยประมาณ วัตถุประสงค์หลักคือการแก้ปัญหาการกระจายตัวของข้อมูลโดยไม่ส่งออกข้อมูลภายในแต่ละเครื่องออกไปยังแหล่งข้อมูลใดๆ เพื่อรักษาไว้ซึ่งความเป็นส่วนตัวของข้อมูล (data privacy)

งานวิจัยนี้สนใจการแก้ปัญหาดังต่อไปนี้

1. ปัญหาการกระจายไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลด้านประเภทข้อมูล (label skew) ซึ่งส่งผลต่อประสิทธิภาพองโมเดลการจำแนกภาพในกระบวนการเรียนรู้แบบสหพันธ์
2. การเรียนรู้ของแกนส์ ให้สามารถเรียนรู้คุณลักษณะของข้อมูลจากทุกแหล่งข้อมูลอาจก่อให้เกิดโอเวอร์ฟิตติ้ง โดยเฉพาะข้อมูลจากแหล่งข้อมูลที่เรียนรู้เป็นแหล่งสุดท้าย
3. ความเป็นส่วนตัวของข้อมูลที่ใช้ในการเรียนรู้แบบสหพันธ์อาจลดลงหรือเสียไป หากต้องมีการแบ่งข้อมูลจริงระหว่างแหล่งข้อมูล



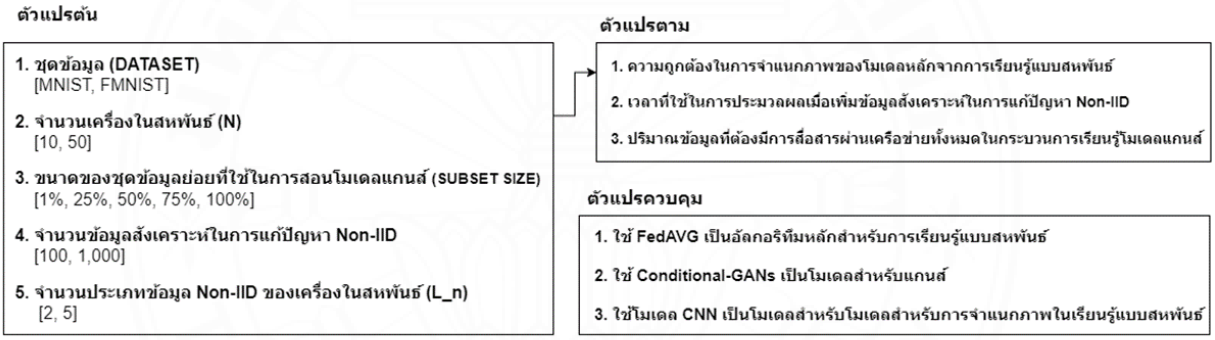
ภาพที่ 3.1 ภาพตัวอย่างการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลในกระบวนการเรียนรู้แบบสหพันธ์[34]

3.2 ภาพรวมขั้นตอนวิธีในการแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

งานวิจัยนี้นำเสนอวิธีการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลบนกระบวนการเรียนรู้แบบสหพันธ์ โดยมีรูปแบบการกระจายของข้อมูลที่ไม่เหมือนกันในด้านประเภทของข้อมูล ซึ่งจะมีจำนวนประเภทของข้อมูลที่ 2 และ 5 ประเภท ในแต่ละเครื่องในสหพันธ์ และใช้เทคนิคการเพิ่มข้อมูลจากแกนส์ในการแก้ไขปัญหาการกระจายตัว โดยโครงสร้างโมเดลแกนส์ที่ใช้จะเป็นคอนดิชั่นนอลแกนส์ [49] ซึ่งจะต้องเรียนรู้ข้อมูลที่กระจายศูนย์อยู่บนเครื่องให้สหพันธ์จำนวน N เครื่อง และเพื่อที่หลีกเลี่ยงการเกิดโอเวอร์ฟิตติ้ง จึงจำเป็นจะต้องแบ่งข้อมูลบนเครื่องในสหพันธ์แต่ละเครื่องออกเป็นชุดข้อมูลย่อย โดยในงานวิจัยนี้จะทดลองที่ขนาดชุดข้อมูลย่อยเท่ากับ 1, 25, 50, 75 และ 100 เปอร์เซ็นต์เทียบกับจำนวนข้อมูลภายในของแต่ละเครื่อง เมื่อกระบวนการเรียนรู้ของแกนส์สิ้นสุดลง จะได้ภาพข้อมูลสังเคราะห์ที่จะถูกส่งไปยังเครื่องในสหพันธ์แต่ละเครื่องในขนาด 100 หรือ 1,000 ภาพ เพื่อแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันในด้านประเภทของข้อมูล

ตารางที่ 3.1 ตัวแปรในการทดลอง

ตัวแปรที่สนใจ	ค่าที่เป็นไปได้
ชุดข้อมูล (DATASET)	MNIST, FMNIST
จำนวนประเภทข้อมูลทั้งหมด	10
จำนวนเครื่องในสหพันธ์ (N)	10, 50
จำนวนประเภทข้อมูลในเครื่องในสหพันธ์ (L_n)	2, 5
ขนาดชุดข้อมูลย่อยที่ใช้ในกระบวนการเรียนรู้โมเดลแกนส์ (SUBSET SIZE)	1, 25, 50, 75, 100



ภาพที่ 3.2 ระเบียบวิธีวิจัย

การวัดผล

1. วัดประสิทธิภาพด้านความถูกต้อง ของการเรียนรู้แบบไม่แก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน กับการเรียนรู้โดยใช้ข้อมูลสังเคราะห์เพื่อแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันบนเครื่องในสหพันธ์

2. เปรียบเทียบประสิทธิภาพของการใช้ข้อมูลสังเคราะห์แก้ปัญหาในการกระจายของข้อมูลในรูปแบบต่างๆ เพื่อดูประสิทธิภาพในการแก้ปัญหาในชุดข้อมูลที่มีความซับซ้อนแตกต่างกัน

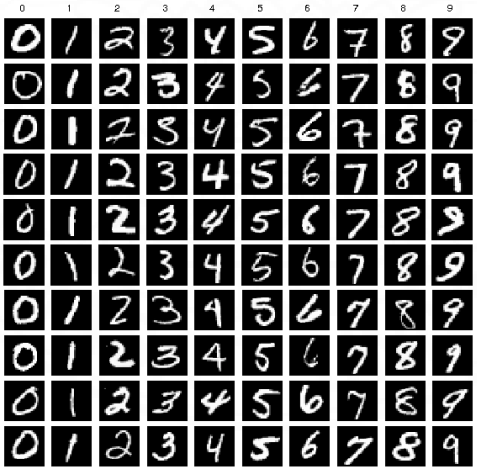
3. เปรียบเทียบประสิทธิภาพของข้อมูลสังเคราะห์เมื่อแบ่งจำนวนชุดข้อมูลย่อยในขนาดต่างๆ (1, 25, 50, 75, 100) ในกระบวนการเรียนรู้ของแกนส์ เพื่อเปรียบเทียบคุณภาพของข้อมูลสังเคราะห์ เทียบกับปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดที่เพิ่มมากขึ้น

4. เปรียบเทียบคะแนนที่ใช้วัดผลคุณภาพของข้อมูลสังเคราะห์ (IS Score, FID Score) กับผลลัพธ์จริงที่ได้จากการใช้งานข้อมูลสังเคราะห์เพื่อปรับปรุงประสิทธิภาพของกระบวนการเรียนรู้แบบสหพันธ์ เพื่อวัดความสอดคล้องของประสิทธิภาพในทางทฤษฎีและในการใช้งานจริง

3.3 ชุดข้อมูลที่ใช้ในงานวิจัย

ในงานวิจัยนี้จะใช้ชุดข้อมูล MNIST [52] , Fashion-MNIST [53] และ โดยใช้สำหรับการทดลองบนการเรียนรู้แบบสหพันธ์ โดยจะกระจายข้อมูลไปยังเครื่องในสหพันธ์ต่างๆ และใช้ในการเรียนรู้ของแกนส์เพื่อสร้างโมเดลผู้สร้างที่สามารถสร้างข้อมูลสังเคราะห์ที่มีคุณลักษณะเหมือนกันกับข้อมูลต้นฉบับ โดยวัดผลจาก FID Score [33] , IS Score [54]

- MNIST ประกอบด้วยภาพลายมือเขียนตัวเลข 0-9 ขนาด 28x28 พิกเซลสีขาวดำ ดังภาพที่ 3.3 โดยมีข้อมูลภาพทั้งสิ้น 70,000 ภาพ แบ่งเป็นข้อมูลภาพที่ใช้ในการเรียนรู้ของโมเดลแกนส์ 60,000 ภาพ และภาพที่ใช้สำหรับทดสอบ 10,000 ภาพ



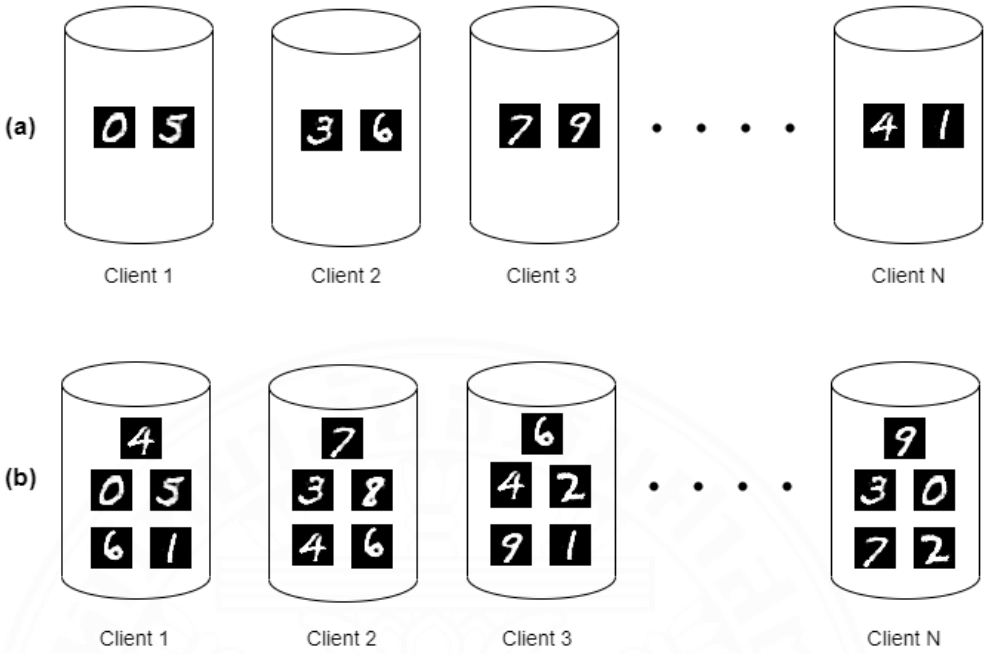
ภาพที่ 3.3 ภาพตัวอย่างชุดข้อมูล MNIST[46]

- Fashion-MNIST เป็นชุดข้อมูลภาพเครื่องแต่งกาย 10 ประเภท ขนาด 28x28 พิกเซลสีขาวดำ ดังภาพที่ 3.4 โดยมีข้อมูลภาพทั้งสิ้น 70,000 ภาพ แบ่งเป็นภาพที่ใช้ในการเรียนรู้ของโมเดลแกนส์ 60,000 ภาพ และภาพสำหรับทดสอบ 10,000 ภาพ

ภาพที่ 3.4 ภาพตัวอย่างชุดข้อมูล Fashion-MNIST[47]

3.4 ข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของเครื่องในสหพันธ์

งานวิจัยนี้ดัดแปลงใช้เครื่องมือเอ็นไอไอดี เบ็นช์ (NIID - Bench) [25] ซึ่งเป็นเกณฑ์มาตรฐานในการวัดประสิทธิภาพการทำงานของโมเดลที่ถูกสร้างโดยการเรียนรู้แบบสหพันธ์ในการจำลองกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันโดยจำกัดเฉพาะ การกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูล (เครื่องในสหพันธ์มีข้อมูลไม่ครบทุกประเภท) โดยจำนวนประเภทข้อมูลในแต่ละเครื่องที่ศึกษา ได้แก่ 2 และ 5 ประเภท จากจำนวนประเภททั้งหมด 10 ประเภท จากสองชุดข้อมูล MNIST และ FMNIST โดยใช้เครื่องในสหพันธ์จำนวน 10 และ 50 เครื่อง

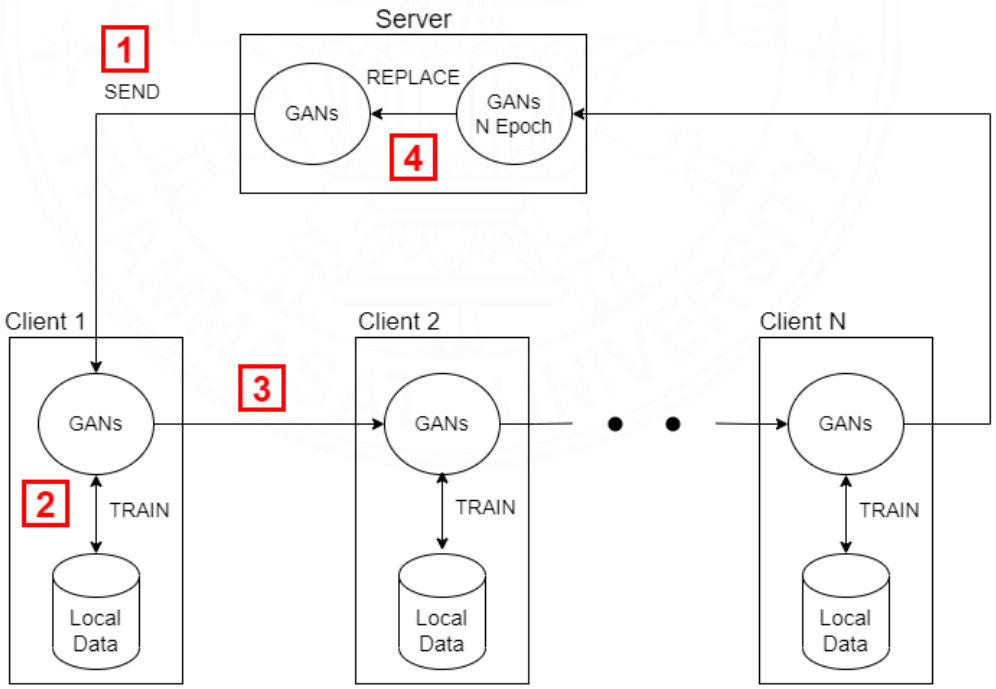


ภาพที่ 3.5 ระดับความรุนแรงของการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันบนเครื่องในสหพันธ์ (a) มีจำนวนประเภทของข้อมูลจำนวน 2 ประเภทต่อหนึ่งเครื่อง (b) มีจำนวนประเภทของข้อมูลจำนวน 5 ประเภทต่อหนึ่งเครื่อง ทั้งนี้ยังไม่ได้พิจารณากรณีที่จำนวนประเภทของข้อมูลภายในเครื่องไม่เท่ากัน

โดยจะทดสอบผลลัพธ์ของทั้งสองรูปแบบทั้งกรณีที่มีจำนวนข้อมูล 2 ประเภทต่อเครื่อง (L_n = 2) และ 5 ประเภทต่อเครื่อง (L_n = 5) ดังภาพที่ 3.5 (a) และ (b) ตามลำดับ โดยมีจำนวนเครื่องในสหพันธ์เท่ากับ N เครื่อง เพื่อวัดระดับความรุนแรงของผลกระทบที่เกิดขึ้นจากการกระจายแต่ละรูปแบบโดยวัดจากประสิทธิภาพของโมเดลในด้านความถูกต้องเปรียบเทียบกับชุดข้อมูลทดสอบ เพื่อใช้ในการเปรียบเทียบประสิทธิภาพของการแก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันในรูปแบบต่างๆ ด้วยข้อมูลสังเคราะห์ และเปรียบเทียบกับทุกชุดข้อมูลทดสอบเพื่อเปรียบเทียบประสิทธิภาพการแก้ปัญหาเทียบกับความซับซ้อนของชุดข้อมูล

3.5 การสอนโมเดลแกนส์เพื่อสังเคราะห์ภาพมาใช้แก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

การสร้างแกนส์โมเดล ในงานวิจัยนี้มีสมมติฐานว่าการแบ่งประเภทของข้อมูลในการสร้างโมเดลแกนส์ สำหรับข้อมูลประเภทใดๆ โดยเฉพาะ จะส่งผลให้ภาพสังเคราะห์ที่ถูกสร้างขึ้นมีคุณลักษณะของภาพที่เหมือนกับภาพจริงมากที่สุด จึงมีการเลือกใช้คอนดิชันนอลแกนส์มาใช้เป็นโมเดลสำหรับสร้างข้อมูลสังเคราะห์ เนื่องจากคอนดิชันนอลแกนส์นั้นใช้ประเภทของข้อมูลเป็นข้อมูลที่ใช้ในการเรียนรู้ด้วย จึงสามารถสร้างโมเดลที่สามารถกำหนดได้ว่าต้องการที่จะสร้างข้อมูลสังเคราะห์ชนิดใด เนื่องจากในการแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน จำเป็นจะต้องสร้างชุดข้อมูลที่มีการกระจายตัวที่สมบูรณ์ที่สุด ซึ่งแต่ละประเภทควรจะมีจำนวนเท่าๆกัน หากใช้โมเดลแกนส์ปกตินั้น จะไม่สามารถกำหนดได้ว่าจะให้โมเดลสร้างข้อมูลประเภทใดออกมา จึงมีโอกาสสร้างข้อมูลประเภทเดิมๆ และเป็นการยากที่จะสร้างชุดข้อมูลเสมือนที่มีการกระจายตัวที่สมบูรณ์ได้ การใช้คอนดิชันนอลแกนส์จึงมีความเหมาะสมกว่า



ภาพที่ 3.6 กระบวนการเรียนรู้ของแกนส์

ขั้นตอนการเรียนรู้ของแกนส์ สำหรับใช้สังเคราะห์ข้อมูลแต่ละประเภท ดังภาพที่ 3.6 สามารถแบ่งออกเป็นขั้นตอนดังนี้

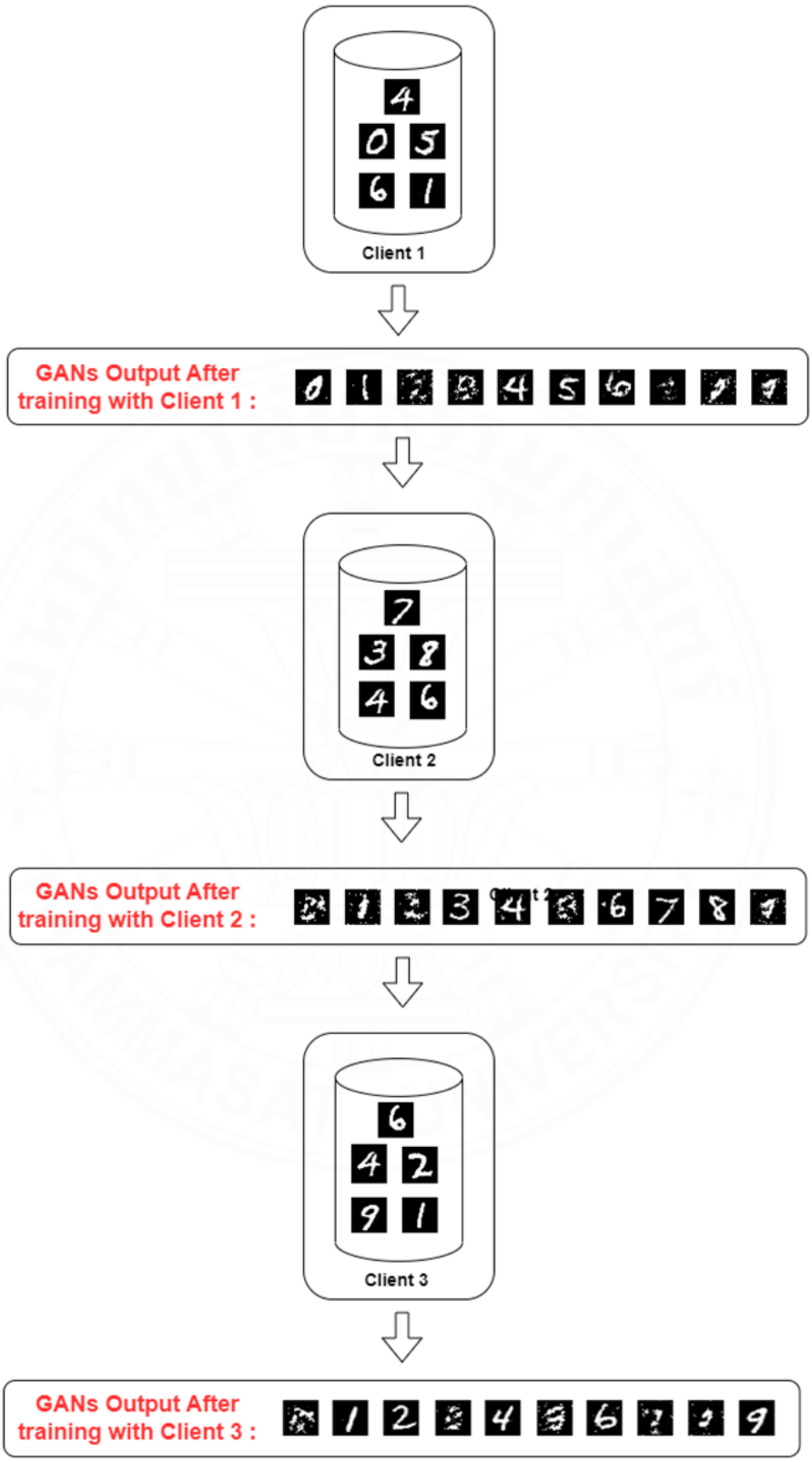
1. เครื่องเซิฟเวอร์สร้างโมเดลแกนส์ขึ้นมา ก่อนที่จะส่งไปยังแต่ละเครื่องในสหพันธ์
2. เครื่องได้รับข้อมูลโมเดลแกนส์ และนำมาเรียนรู้กับข้อมูลภายในของตนเอง (local data)
3. เมื่อกระบวนการเรียนรู้กับข้อมูลภายในเสร็จสิ้น จึงจะส่งต่อโมเดลแกนส์ไปยังเครื่องอื่นๆ
4. เมื่อทำซ้ำในข้อ 2-3 จนสามารถเรียนรู้จากข้อมูลภายในแต่ละเครื่องได้ครบถ้วนแล้ว จึงจะทำการส่งโมเดลแกนส์กลับไปยังเซิฟเวอร์ เพื่อแทนที่โมเดลแกนส์เวอร์ชัน (version) ก่อนหน้า

ทำซ้ำในข้อ 1-4 เพื่อเพิ่มจำนวนรอบในการเรียนรู้ (epoch) เพื่อเพิ่มประสิทธิภาพในการเรียนรู้ของแกนส์

เมื่อสร้างโมเดลแกนส์แต่ละประเภทของข้อมูลจนครบ จะเก็บไว้ในเครื่องเซิฟเวอร์ เพื่อรอกระบวนการสร้างข้อมูลสังเคราะห์ในลำดับถัดไป

การเพิ่มประสิทธิภาพในกระบวนการเรียนรู้ของแกนส์บนระบบกระจายข้อมูลแบบอิสระต่อกันสำหรับระบบที่มีการกระจายของข้อมูลแบบอิสระต่อกันนั้น เมื่อทำการเรียนรู้กับข้อมูลทั้งหมดของแต่ละเครื่องในสหพันธ์แบบจามลำดับจนครบทุกเครื่อง โมเดลผู้สร้างแกนส์ซึ่งเป็นผลลัพธ์ของการเรียนรู้นั้นมีแนวโน้มที่จะโอเวอร์ฟิตติ้งสูง และมักจะสร้างข้อมูลสังเคราะห์ได้ดีเฉพาะประเภทของข้อมูลที่ถูกเก็บไว้ในเครื่องสุดท้ายที่มีการเรียนรู้ ตัวอย่างดังภาพที่ 3.7 ถ้าเรียนรู้ด้วยข้อมูลในเครื่องที่ 1, 2 และ 3 ตามลำดับ โมเดลที่ได้จะสร้างข้อมูลประเภท 1, 2, 4, 6, 9 ได้ดีกว่าประเภทอื่นๆ เนื่องจากเครื่องที่ 3 ซึ่งเป็นเครื่องสุดท้าย มีข้อมูลประเภทนี้อยู่ แต่ไม่มีข้อมูลประเภทอื่นๆ โมเดลมีแนวโน้มที่จะให้ค่าน้ำหนักแก่ประเภทของข้อมูลที่ไม่ถูกนำมาเรียนรู้เป็นเวลานานลดลง ส่งผลให้ประสิทธิภาพในการสร้างข้อมูลสังเคราะห์ประเภทนั้นๆ ลดลงเป็นอย่างมาก

ในการแก้ปัญหาการโอเวอร์ฟิตติ้งของโมเดล จะใช้วิธีแบ่งข้อมูลในแต่ละเครื่องเป็นขนาดชุดข้อมูลย่อย และทำการเรียนรู้เป็นรอบๆ จนกว่าจะเรียนรู้ข้อมูลในแต่ละเครื่องได้ครบถ้วน งานวิจัยนี้แบ่งข้อมูลทั้งหมดในแต่ละเครื่องออกเป็นชุดย่อยๆ แต่ละชุดย่อยมีขนาดคิดเป็น 0, 25, 50, 75, 100 เปอร์เซ็นต์ของข้อมูลทั้งหมดในเครื่องนั้นๆ เพื่อป้องกันการเกิดโอเวอร์ฟิตติ้งที่เกิดจากการเรียนรู้ข้อมูลประเภทซ้ำเดิมมากเกินไป การแบ่งข้อมูลออกเป็นชุดข้อมูลย่อยนั้นจะสามารถเพิ่มการกระจายตัวของชุดข้อมูลที่จะถูกนำมาใช้ในกระบวนการเรียนรู้ของโมเดลแกนส์ โดยผู้วิจัยตั้งข้อสันนิษฐานว่ายิ่งชุดข้อมูลย่อยมีขนาดเล็กมากเท่าไหร่ จะยิ่งทำให้การกระจายตัวดีขึ้นและส่งผลให้คุณภาพของข้อมูลสังเคราะห์นั้นออกมามีคุณภาพมากที่สุด แต่ในขณะเดียวกัน ก็ต้องทำอย่างระมัดระวังเพราะจะเป็นการเพิ่มปริมาณการสื่อสารบนเครือข่ายระหว่างเครื่องตามไปด้วย



ภาพที่ 3.7 ผลลัพธ์ที่ได้จากโมเดลผู้สร้างในแต่ละรอบของการเรียนรู้ในแต่ละรอบของการเรียนรู้

3.6 การวัดผลข้อมูลสังเคราะห์

เมื่อเสร็จสิ้นขั้นตอนในการสร้างโมเดลแกนส์ จะได้มาซึ่งโมเดลผู้สร้างสำหรับสร้างข้อมูลในแต่ละประเภท โดยโมเดลผู้สร้างสามารถสร้างข้อมูลสังเคราะห์ที่มีคุณลักษณะเหมือนกับประเภทของข้อมูลต้นฉบับที่ใช้ในการเรียนรู้ โดยในการที่จะวัดประสิทธิภาพของโมเดลผู้สร้างนั้น ว่าสามารถสร้างข้อมูลสังเคราะห์ที่มีคุณลักษณะเหมือนกันกับข้อมูลจริง จะใช้วิธีวัดผลคุณภาพของข้อมูล ด้วย IS score [33] และ FID Score [54]

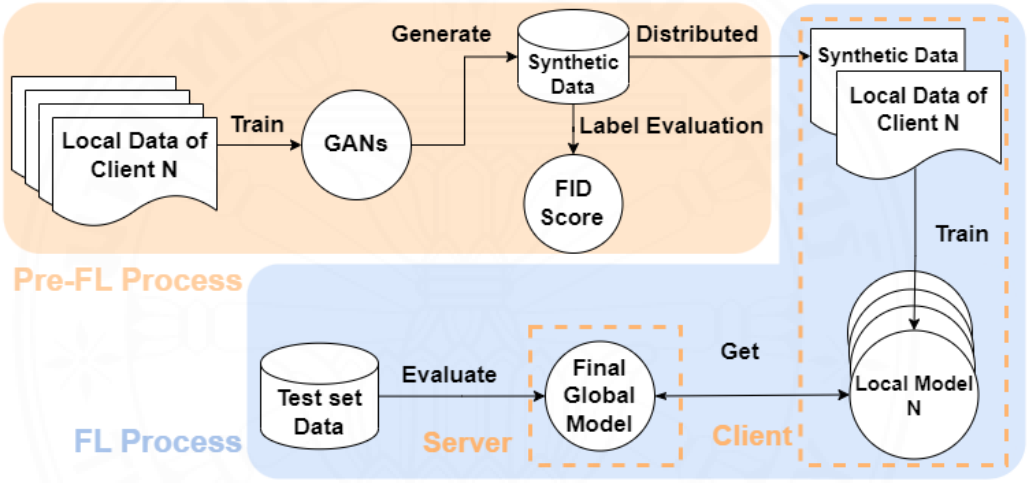
การเปรียบเทียบคะแนน IS Score และ FID Score ของข้อมูลที่ถูกสังเคราะห์ขึ้นด้วยโมเดลแกนส์กับข้อมูลต้นฉบับสำหรับการเรียนรู้ (training set) และข้อมูลสังเคราะห์ของงานวิจัยล่าสุดที่พัฒนาข้อมูลสังเคราะห์ด้วยแกนส์ หากคะแนนใกล้เคียงหรือดีกว่าข้อมูลต้นฉบับสำหรับการเรียนรู้และข้อมูลสังเคราะห์จากงานวิจัยช่วงที่ผ่านมาจึงจะถือว่าได้โมเดลผู้สร้างที่เหมาะสมกับการนำไปใช้งาน

3.7 การแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยการเพิ่มข้อมูลจากข้อมูลสังเคราะห์

ในขั้นตอนการทดลองเมื่อได้โมเดลแกนส์สำหรับข้อมูลแต่ละประเภทแล้ว จะใช้โมเดลแกนส์ที่อยู่บนเครื่องเซิฟเวอร์ในการสร้างชุดข้อมูลสังเคราะห์ขึ้นมาเก็บไว้บนเครื่องเซิฟเวอร์ก่อนที่จะส่งไปยังเครื่องในสหพันธ์ โดยจะสร้างส่งข้อมูลแต่ละประเภทจำนวนเท่ากับ 1,000 ภาพ ต่อประเภทของข้อมูลต่อเครื่องในสหพันธ์ เนื่องจากชุดข้อมูล ILSVRC2012 ได้พิสูจน์แล้วว่าจำนวน 1,000 ภาพในแต่ละประเภทเป็นจำนวนที่เพียงพอในการสร้างโมเดลจำแนก (classification model) ที่มีประสิทธิภาพ (ตัวอย่างเช่น AlexNet , InceptionV3) [36] โดยขั้นตอนนี้จะเป็นขั้นตอนก่อนเริ่มกระบวนการเรียนรู้แบบสหพันธ์ดังที่แสดงในภาพที่ 3.8 ในส่วน Pre-FL

หลังจากนั้นเครื่องเซิฟเวอร์จะทำการส่งข้อมูลสังเคราะห์ไปยังเครื่องในสหพันธ์เพื่อแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันก่อนเริ่มกระบวนการเรียนรู้แบบสหพันธ์ โดยจุดประสงค์คือทำให้ชุดข้อมูลในแต่ละเครื่องมีข้อมูลครบทุกประเภทและแต่ละประเภทมีจำนวนที่มากเพียงพอที่จะใช้ในการดึงคุณลักษณะพิเศษของประเภทข้อมูลนั้นๆออกมาได้ ในขั้นตอนกระบวนการเรียนรู้แบบสหพันธ์

โดยในขั้นตอนก่อนเริ่มกระบวนการเรียนรู้แบบสหพันธจะเป็นขั้นตอนการแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันบนเครื่องในสหพันธ์โดยจะเริ่มกระบวนการเรียนรู้ของโมเดลแกนส์บนระบบเก็บข้อมูลแบบกระจายศูนย์ เมื่อได้ข้อมูลสังเคราะห์มาแล้ว จะทำการกระจายข้อมูลเหล่านั้นไปยังเครื่องในสหพันธ์ทุกเครื่อง โดยจะทำการรวมข้อมูลภายในเข้ากับข้อมูลสังเคราะห์ที่ได้รับ เพื่อเป็นการแก้ไขปัญหาการกระจายตัวของข้อมูลก่อนที่จะเริ่มกระบวนการเรียนรู้แบบสหพันธ์ ดังภาพที่ 3.8 ในส่วนของ FL Process ที่เป็นการเริ่มต้นการเรียนรู้แบบสหพันธ์แบบปกติ หลังจากที่ทำการแก้ไขปัญหาการกระจายตัวของข้อมูลโดยใช้ข้อมูลสังเคราะห์แล้ว



ภาพที่ 3.8 ภาพรวมสถาปัตยกรรมของระบบ

3.8 การประเมินประสิทธิภาพของการแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

หลังจากได้ผลลัพธ์จากกระบวนการเรียนรู้แบบสหพันธ์แล้ว จะทำการเปรียบเทียบประสิทธิภาพของโมเดลหลักซึ่งได้จากการใช้ภาพสังเคราะห์เพื่อแก้ปัญหาการกระจายของข้อมูล กับประสิทธิภาพของโมเดลหลักที่ได้จากกระบวนการเรียนรู้แบบสหพันธ์ที่ยังไม่ได้แก้ปัญหาการกระจายของข้อมูล ประสิทธิภาพของโมเดลหลักวัดจากค่าความถูกต้องในการจำแนกชุดข้อมูลทดสอบ

3.9 สมมติฐานในการทดลอง

1. เมื่อแก้ปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันบนโดยใช้ของมูลสังเคราะห์ที่สร้างจากโมเดลแกนส์ โมเดลกลางที่ได้จากกระบวนการเรียนรู้แบบสหพันธ์จะมีประสิทธิภาพดีกว่าโมเดลกลางที่ได้จากกระบวนการเรียนรู้แบบสหพันธ์ที่ยังไม่แก้ไขปัญหาการกระจายตัวของข้อมูล

2. คะแนนคุณภาพของข้อมูลสังเคราะห์ (FID Score) ชุดข้อมูลที่มีคะแนนที่ต่ำจะเป็นชุดข้อมูลที่มีคุณลักษณะตรงกับข้อมูลจริงมากที่สุด จึงมีแนวโน้มที่จะปรับปรุงประสิทธิภาพของกระบวนการเรียนรู้แบบสหพันธ์ได้มากกว่า

3. การเพิ่มข้อมูลสามารถเพิ่มประสิทธิภาพของการเรียนรู้แบบสหพันธ์ได้ แม้ว่าจะมีการกระจายตัวของข้อมูลที่เป็นปกติ

4. การใช้จำนวนข้อมูลสังเคราะห์จำนวน 1,000 ภาพต่อประเภทข้อมูลในหนึ่งเครื่องในสหพันธ์นั้น มากเพียงพอที่จะแก้ปัญหาการกระจายตัวของข้อมูล และสามารถแก้ปัญหาการโอเวอร์ฟิตติ้งของกระบวนการเรียนรู้แบบสหพันธ์เนื่องจากขาดข้อมูลบางประเภทบนเครื่องในสหพันธ์

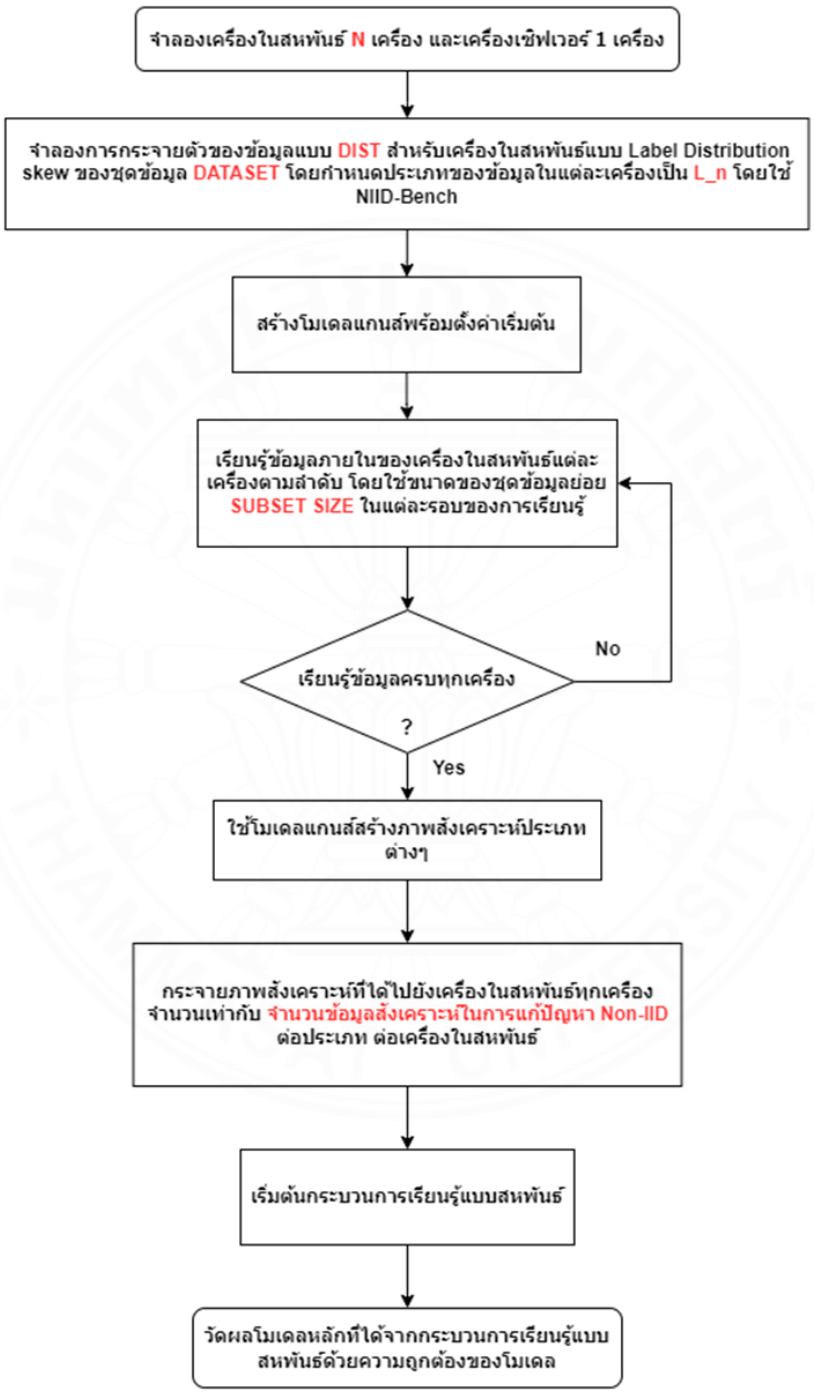
3.10 การออกแบบการทดลอง

ในงานวิจัยนี้จะใช้เครื่องมือเอ็นไอไอดี เบ็นช์ (NIID-Bench) ในการจำลองการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน งานวิจัยนี้จะสนใจรูปแบบการกระจายที่ไม่เหมือนกันในด้านประเภทของข้อมูล (Label Distribution Skew) เป็นหลัก โดยจะมีตัวแปรที่สามารถกำหนดค่าได้ในงานวิจัยคือ

1. ชุดข้อมูล (MNIST, FMNIST)
2. จำนวนเครื่องในสหพันธ์ที่ใช้ในการทดลอง (10, 50)
3. จำนวนประเภทข้อมูลที่มีในแต่ละเครื่องในสหพันธ์ (2, 5)
4. ขนาดของชุดข้อมูลย่อยที่ใช้ในการเรียนรู้ของโมเดลแกนส์ (1, 25, 50, 75, 100)

ในการเลือกชุดข้อมูลที่ใช้ในการทดลองนั้น ได้มีการคำนึงถึงความซับซ้อนของประเภทข้อมูลภาพในชุดข้อมูล เพื่อต้องการแสดงให้เห็นถึงประสิทธิภาพในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของชุดข้อมูลที่มีความซับซ้อนน้อยกับชุดข้อมูลที่มีความซับซ้อนมากขึ้น ในการทดลองจะใช้จำนวนเครื่องในสหพันธ์และจำนวนประเภทของข้อมูลในเครื่องในสหพันธ์ในจำนวนที่แตกต่างกัน เพื่อแสดงประสิทธิภาพเมื่อมีการเพิ่มขึ้นของความไม่เหมือนกันและไม่

เป็นอิสระต่อกันของข้อมูลเมื่อมีจำนวนเครื่องในสหพันธ์เพิ่มมากขึ้น หรือมีจำนวนประเภทของข้อมูลในแต่ละเครื่องในสหพันธ์ลดลง



ภาพที่ 3.9 ผังงานแสดงภาพรวมการทำงานของระบบ

ในการทดลองจะใช้ขนาดของชุดข้อมูลย่อยในหลากหลายขนาด เพื่อเลือกขนาดของชุดข้อมูลย่อยที่เหมาะสมที่สุด โดยทำการเปรียบเทียบคุณภาพของข้อมูลสังเคราะห์เทียบกับค่าใช้จ่ายในการโอนถ่ายข้อมูล เพื่อให้ได้ขนาดที่เหมาะสมที่สุดในการใช้ในกระบวนการเรียนรู้ของแกนส์แบบคอนดิชันนอล โดยโมเดลที่ใช้ในกระบวนการเรียนรู้ของแกนส์แบบคอนดิชันนอลนั้นจะใช้ขนาดชุดข้อมูลในการเรียนรู้ (batch size) เท่ากับ 32

สำหรับโมเดลหลักที่ใช้ในกระบวนการเรียนรู้แบบสหพันธ์นั้นจะใช้โครงข่ายประสาทเทียมแบบสังวัตนาการที่ออกแบบมาสำหรับการจำแนกข้อมูลประเภทภาพสำหรับข้อมูล MNIST และ FMNIST โดยมีชั้นคอนโวลูชันจำนวนสองชั้นเชื่อมด้วยแม็กพูลลิ่ง (max pooling) และชั้นเชื่อมต่ออย่างสมบูรณ์ (fully connected layer) จำนวนสามชั้น โดยชั้นคอนโวลูชันชั้นแรกจะทำการเชื่อมข้อมูลเข้ากับคุณลักษณะของภาพจำนวน 6 คุณลักษณะ และ 16 คุณลักษณะในคอนโวลูชันชั้นที่สอง ชั้นแม็กพูลลิ่งจะทำการลดขนาดของข้อมูลที่ทำการเชื่อมคุณลักษณะแล้ว และชั้นเชื่อมต่ออย่างสมบูรณ์จำนวนสามชั้นที่เหลือจะทำการใช้ลิเนียร์ทรานฟอร์เมชัน (linear transformation) และแอคทิเวชันฟังก์ชัน (activation function) กับข้อมูลเข้าโดยมีจำนวนชั้นซ่อนเท่ากับ 256, 120 และ 84 ตามลำดับ เพื่อให้ได้ออกมาเป็นผลลัพธ์ในชั้นสุดท้ายที่มีจำนวน 10 ค่า เท่ากับจำนวนของประเภทข้อมูล เพื่อใช้ในการจำแนกประเภทของข้อมูล โดยโมเดลหลักจะมีค่าอัตราการเรียนรู้ (learning rate) อยู่ที่ 0.01 และมีขนาดชุดข้อมูลที่ใช้ในการเรียนรู้ (batch size) เท่ากับ 64 และในกระบวนการเรียนรู้จะใช้จำนวนรอบของการเรียนรู้แบบสหพันธ์ (round) เท่ากับ 10-15 รอบ ซึ่งเพียงพอต่อการสร้างโมเดลที่มีประสิทธิภาพ โดยประสิทธิภาพของโมเดลมักจะคงที่ภายในรอบที่ 10

โดยเริ่มต้นด้วยการจำลองเครื่องสหพันธ์จำนวน N เครื่อง ดังที่แสดงในภาพที่ 3.9 จากนั้นจะทำการจำลองการกระจายตัวของข้อมูลแบบเหมือนกันและไม่เป็นอิสระต่อกัน โดยกำหนดให้จำนวนประเภทของข้อมูลต่อหนึ่งเครื่องเท่ากับ L_n โดยใช้ NIID-Bench ในการจำลอง จากนั้นจะทำการสร้างโมเดลแกนส์และทำการจำลองการส่งไปยังแต่ละเครื่อง เพื่อเรียนรู้กับข้อมูลภายในของแต่ละเครื่องด้วยขนาดเท่ากับ SUB_SIZE ก่อนจะส่งไปยังเครื่องอื่นๆ เมื่อเสร็จสิ้นกระบวนการเรียนรู้ของแกนส์ จะนำโมเดลผู้สร้างที่ได้มาใช้ในการสร้างชุดข้อมูลสังเคราะห์ในแต่ละประเภทจำนวน 100 ภาพ เพื่อส่งข้อมูลสังเคราะห์ในแต่ละชุดไปยังเครื่องในสหพันธ์ทุกๆเครื่อง เพื่อแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน ก่อนที่จะเริ่มกระบวนการเรียนรู้แบบสหพันธ์ตามปกติ และวัดประสิทธิภาพในการใช้ข้อมูลสังเคราะห์ในการแก้ปัญหาด้วยความถูกต้องของโมเดลหลัก ซึ่งเป็นผลลัพธ์ที่ได้จากกระบวนการเรียนรู้แบบสหพันธ์

บทที่ 4
ผลการวิจัยและอภิปรายผล

4.1 การทดลองพื้นฐาน (Baseline Experiment)

ในการทดลองนี้จะทำการทดสอบประสิทธิภาพของโมเดลหลักที่ได้จากกระบวนการเรียนรู้แบบสหพันธ์ โดยจะใช้ชุดข้อมูล MNIST และ FMNIST ในการทดลอง ซึ่งในการทดลองจะแบ่งการกระจายตัวออกเป็น 3 รูปแบบ คือ

1. การกระจายตัวของข้อมูลแบบปกติ (IID)
2. การกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน โดยแต่ละเครื่องในสหพันธ์มีจำนวนข้อมูลเท่ากับ 2 (Non-IID L_n = 2)
3. การกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน โดยแต่ละเครื่องในสหพันธ์มีจำนวนข้อมูลเท่ากับ 5 (Non-IID L_n = 5)

โดยมีจำนวนเครื่องในสหพันธ์ (N) ทั้งหมดเท่ากับ 10 จะได้ผลลัพธ์ดังตารางที่ 4.1

จากการทดลองพบว่าชุดข้อมูล MNIST ที่มีความซับซ้อนน้อยนั้น เมื่อมีการกระจายตัวของข้อมูลแบบปกติ (IID, L_n = 10) จะได้ผลลัพธ์ใกล้เคียงกับการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันในระดับปานกลาง (Non-IID, L_n = 5) แต่เมื่อข้อมูลมีการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันมากขึ้นเป็นระดับสูง (Non-IID, L_n = 2) จะทำให้ความถูกต้องของโมเดลหลักลดลงจากโมเดลหลักในกรณีปกติ ซึ่งได้ค่าความถูกต้องจาก 96.95% ลงมาเป็น 63.58% (ลดลงมากถึง 33.37)

สำหรับชุดข้อมูล FMNIST ที่มีความซับซ้อนของภาพมากกว่านั้น เมื่อเปรียบเทียบประสิทธิภาพของโมเดลหลักที่ได้จากการเรียนรู้ด้วยข้อมูลที่มีการกระจายตัวแบบปกติกับข้อมูลที่มีการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันแล้ว พบว่าความถูกต้องของโมเดลหลักลดลงอย่างมีนัยยะสำคัญในทุกระดับ และมากขึ้นเมื่อมีการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันที่มากขึ้น

ตารางที่ 4.1 ตารางแสดงความถูกต้องของโมเดลหลักที่ได้จากการทดลองพื้นฐาน

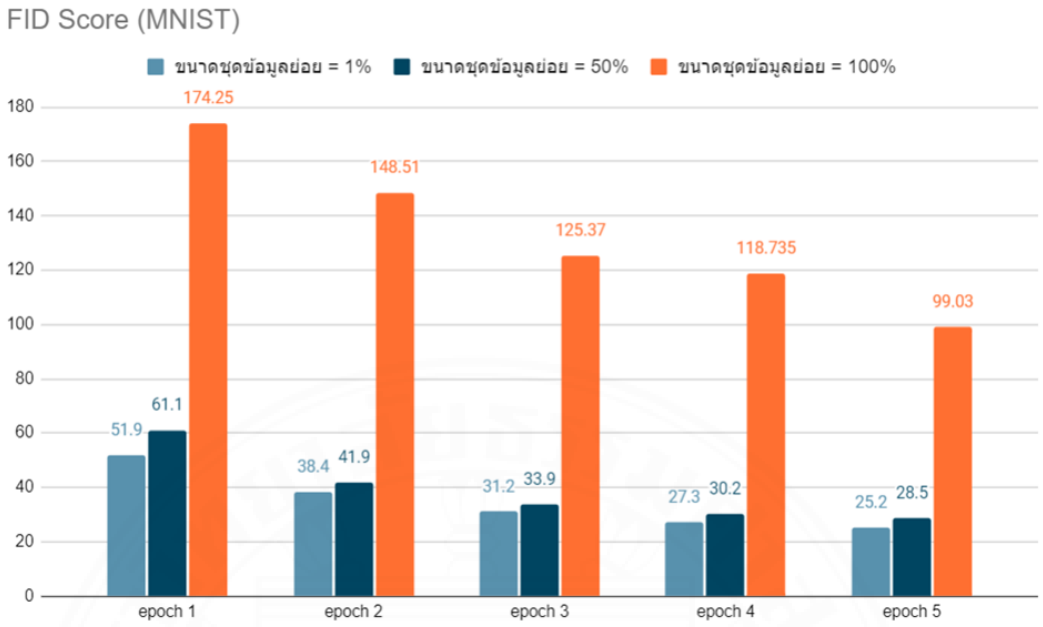
จำนวนเครื่องในสหพันธ์ (N)	ชุดข้อมูล (DATASET)	การกระจายของข้อมูล (DIST)	ประเภทข้อมูลที่มีในแต่ละเครื่อง (L_n)	ความถูกต้องของโมเดลหลักซึ่งเป็นผลจากกระบวนการเรียนรู้แบบสหพันธ์ (MAcc)
5	MNIST	IID	10	98.41
		Non-IID	5	96.35
		Non-IID	2	83.41
	FMNIST	IID	10	85.35
		Non-IID	5	67.91
		Non-IID	2	58.48
10	MNIST	IID	10	96.95
		Non-IID	5	96.55
		Non-IID	2	63.58
	FMNIST	IID	10	82.07
		Non-IID	5	67.77
		Non-IID	2	53.82

4.2 การทดลองเพื่อศึกษาขนาดของชุดข้อมูลย่อยในการเรียนรู้ของโมเดลแกนส์และคุณภาพของข้อมูลสังเคราะห์ของโมเดลแกนส์ที่ได้

เมื่อมีปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของเครื่องในสหพันธ์ ซึ่งทำให้เราต้องสร้างโมเดลแกนส์เพื่อเพิ่มข้อมูลสังเคราะห์ให้เครื่องเหล่านั้น การทดลองนี้มุ่งเน้นศึกษาผลกระทบของขนาดชุดข้อมูลย่อยที่ใช้ในการฝึกสอนโมเดลแกนส์ที่มีต่อประสิทธิภาพของโมเดลแกนส์ที่ได้และปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายโดยพิจารณากรณีที่ มีเครื่องในสหพันธ์จำนวน 10 เครื่อง (N = 10) แต่ละเครื่องนั้นมีประเภทของข้อมูลอยู่ 5 ประเภท (L_n = 5) โดยกำหนดให้ขนาดของชุดข้อมูลย่อยมีค่าเป็น 1, 25, 50, 75 และ 100 เปอร์เซ็นต์ของข้อมูลทั้งหมดในแต่ละเครื่อง และกำหนดให้ปริมาณข้อมูลที่ต้องสื่อสารผ่านเครือข่ายต่อครั้งอยู่ที่ 83.25 MB ซึ่งเท่ากับขนาดของโมเดลแกนส์แบบคอนดิชันนอล โดยจะใช้จำนวนรอบในการเรียนรู้ (epoch) เท่ากับ 5 รอบ เนื่องจากเป็นจำนวนรอบที่ประสิทธิภาพของโมเดลแกนส์เริ่มเข้าสู่จุดเสถียร (ดังที่แสดงในภาพ 4.1)

ตารางที่ 4.2 ตารางแสดงคุณภาพของข้อมูลเทียบกับปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดสำหรับชุดข้อมูลย่อยในขนาดต่างๆ

ขนาดของชุดข้อมูลย่อย (SUBSET SIZE)	จำนวนรอบสหพันธ์ที่ต้องใช้ (ROUND)	ปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมด (MB)	คะแนนคุณภาพข้อมูลสังเคราะห์จากโมเดลแกนส์ที่ได้ (FID Score)
100%	10	832	N/A
75%	20	1,665	79.5
50%	20	1,665	28.5
25%	40	3,330	26.3
1%	1,870	155,677	25.2



ภาพที่ 4.1 คะแนนคุณภาพของข้อมูลสังเคราะห์ (FID Score) ที่สร้างจากโมเดลแกนส์หลังแต่ละรอบของการเรียนรู้ (epoch) ด้วยชุดข้อมูล MNIST

ผลลัพธ์จากการทดลองดังตารางที่ 4.2 พบว่าเมื่อใช้ข้อมูลทั้งหมดที่มีในเครื่องฝึกสอนโมเดลแกนส์นั้น (ขนาดข้อมูลย่อย = 100%) จะส่งผลให้ข้อมูลสังเคราะห์ที่ได้มีเพียงบางประเภทซึ่งปรากฏในเครื่องหลังสุดที่ใช้ในการฝึกสอนเท่านั้น (อธิบายในภาพที่ 3.7) ซึ่งทำให้ไม่สามารถคำนวนคะแนนคุณภาพของข้อมูลสังเคราะห์ที่เกิดจากโมเดลผู้สร้างนี้ได้ในภาพรวมอย่างถูกต้อง (แทนตารางด้วย N/A) และเมื่อใช้ขนาดของข้อมูลย่อยเท่ากับ 1% นั้น ส่งผลให้ภาพสังเคราะห์ที่ได้มีค่า FID Score น้อยที่สุดในการทดลองนี้ เท่ากับ 25.2 ซึ่งถือได้ว่ามีคุณภาพที่ดีมาก เมื่อเทียบกับงานวิจัยล่าสุดในด้านนี้ อย่างไรก็ตามการใช้ขนาดของชุดข้อมูลย่อยเล็กขนาดนั้นจำเป็นต้องใช้จำนวนรอบในการเรียนรู้มาก และส่งผลให้ปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดสูงขึ้นอย่างทวีคูณ ขนาดของข้อมูลย่อยเท่ากับ 50% นั้นทำให้ได้คุณภาพของข้อมูลสังเคราะห์ใกล้เคียงกับขนาดข้อมูลย่อย 1% ในขณะที่ใช้จำนวนรอบที่น้อยกว่ามาก จึงส่งผลให้มีปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดที่น้อยลงตามไปด้วย จึงสรุปผลได้ว่าขนาดของข้อมูลย่อยที่ 50% ของข้อมูลในเครื่องทั้งหมดในเครื่องเป็นค่าที่เหมาะสมที่จะใช้ในกระบวนการเรียนรู้ของแกนส์ เนื่องจากสามารถสร้างข้อมูลสังเคราะห์ที่มีคุณภาพที่ดี ในขณะที่ใช้ปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดที่ต่ำ

4.3 การทดลองเพื่อศึกษาจำนวนข้อมูลสังเคราะห์ที่เหมาะสมจะนำมาใช้ในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน

จากสมมติฐานที่ว่าจำนวนข้อมูลสังเคราะห์ 1,000 ภาพในแต่ละประเภทข้อมูลต่อเครื่องในสหพันธ์ จะสามารถแก้ปัญหาการกระจายตัวของข้อมูลได้ การทดลองเปรียบเทียบประสิทธิภาพด้านความถูกต้องของโมเดลหลักและเวลาที่ใช้ในกระบวนการเรียนรู้แบบสหพันธ์ ซึ่งมีทั้งหมด 50 เครื่อง (N = 50) เมื่อใช้ชุดข้อมูล MNIST และ FMNIST มีปัญหาการกระจายของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน โดยมีจำนวนประเภทของข้อมูลในเครื่อง 2 และ 5 ประเภท และใช้จำนวนข้อมูลสังเคราะห์ต่อประเภทต่อเครื่องในสหพันธ์อยู่ที่ 100 และ 1,000 ภาพ

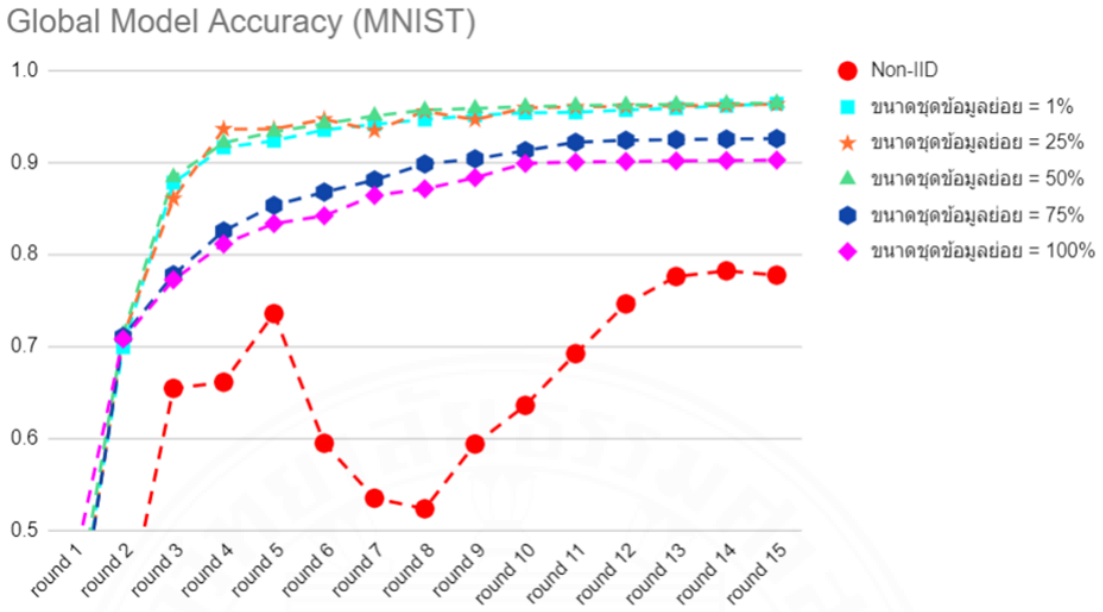
ตารางที่ 4.3 ตารางความถูกต้องของโมเดลเมื่อใช้ข้อมูลสังเคราะห์จำนวนต่างๆ (N=50 , Round=10)

ชุดข้อมูล (DATASET)	การกระจายของข้อมูล (DIST)	ประเภทข้อมูลที่มีในแต่ละเครื่อง (L_n)	จำนวนข้อมูลสังเคราะห์ต่อประเภทต่อเครื่องในสหพันธ์	ความถูกต้องของโมเดลหลักซึ่งเป็นผลจากกระบวนการเรียนรู้แบบสหพันธ์ (MAcc)	เวลาที่ใช้ในกระบวนการเรียนรู้แบบสหพันธ์ (นาที)
MNIST	IID	10	0	79.59	30
	Non-IID	5	0	62.14	32
	Non-IID	5	100	92.74	43
	Non-IID	5	1,000	97.12	147
	Non-IID	2	0	43.16	33
	Non-IID	2	100	91.68	47
	Non-IID	2	1,000	96.91	151
FMNIST	IID	10	0	63.81	32
	Non-IID	5	0	50.77	33
	Non-IID	5	100	73.01	45
	Non-IID	5	1,000	80.73	153
	Non-IID	2	0	34.69	35
	Non-IID	2	100	72.40	44
	Non-IID	2	1,000	80.20	159

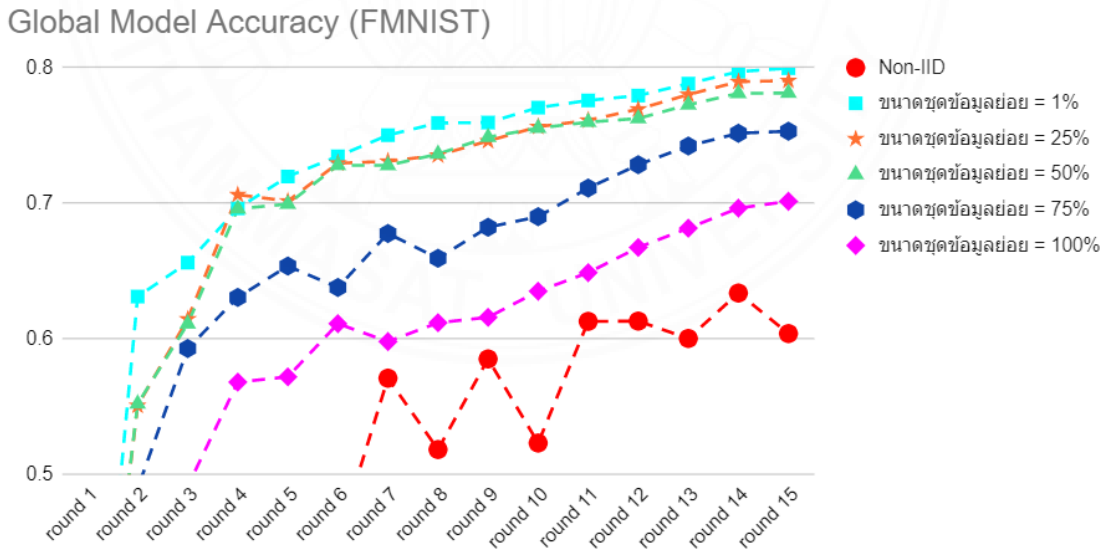
ผลการทดลองแสดงให้เห็นว่าในการใช้จำนวนข้อมูลสังเคราะห์ที่เพิ่มมากขึ้น ส่งผลให้ความถูกต้องของโมเดลหลักเพิ่มมากขึ้น เมื่อใช้จำนวนข้อมูลสังเคราะห์ 1,000 ภาพต่อประเภทข้อมูลต่อเครื่องในสหพันธ์ จะเพิ่มประสิทธิภาพของโมเดลหลักด้านความถูกต้องอยู่ที่เฉลี่ย 5-7% เมื่อเทียบกับการใช้ข้อมูลสังเคราะห์เพียง 100 ภาพต่อประเภทต่อเครื่องในสหพันธ์ ในขณะเดียวกันการใช้ข้อมูลสังเคราะห์จำนวนมาก ส่งผลให้ระยะเวลาที่ใช้ในกระบวนการเรียนรู้แบบสหพันธ์เพิ่มขึ้นมากกว่า 3 เท่า เมื่อเพิ่มจำนวนข้อมูลสังเคราะห์จาก 100 ภาพ เป็น 1,000 ภาพต่อประเภทต่อเครื่องในสหพันธ์ ผลสรุปจากการทดลองนี้จึงทำให้ผู้วิจัยแนะนำการใช้จำนวนข้อมูลสังเคราะห์เพียง 100 ภาพต่อประเภทข้อมูลต่อหนึ่งเครื่องในสหพันธ์ เพื่อลดระยะเวลาในการประมวลผล

4.3.1 การทดลองเพื่อศึกษาประสิทธิภาพในการแก้ปัญหาการกระจายตัวของข้อมูลในภาพรวม

ในการทดลองนี้จะทำการเปรียบเทียบประสิทธิภาพของข้อมูลสังเคราะห์ที่สร้างด้วยชุดข้อมูลย่อยในขนาดต่างๆ ในการแก้ปัญหาความไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลบนกระบวนการเรียนรู้แบบสหพันธ์ โดยในการทดลองจะใช้ชุดข้อมูล MNIST และ FMNIST โดยจำลองการกระจายตัวของข้อมูลในรูปแบบแบบที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน และมีจำนวนประเภทของข้อมูลในเครื่องในสหพันธ์เพียง 2 ประเภทต่อหนึ่งเครื่องในสหพันธ์ โดยมีจำนวนเครื่องในสหพันธ์ทั้งหมด 10 เครื่อง และข้อมูลสังเคราะห์ที่ใช้ในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน จะถูกสร้างขึ้นด้วยการเรียนรู้แบบชุดข้อมูลย่อยในกระบวนการเรียนรู้ของแกนส์ โดยจะมีขนาดชุดข้อมูลย่อยในแต่ละรอบของการเรียนรู้อยู่ที่ 1, 25 ,50, 75 และ 100 เปอร์เซ็นต์ เพื่อเปรียบเทียบประสิทธิภาพของโมเดลหลักในกระบวนการเรียนรู้แบบสหพันธ์เมื่อใช้ข้อมูลสังเคราะห์ที่ถูกสร้างขึ้นจากการเรียนรู้ด้วยชุดข้อมูลย่อยในขนาดต่างๆ



ภาพที่ 4.2 ความถูกของโมเดลหลักในกระบวนการเรียนรู้แบบสหพันธ์ในแต่รอบของการเรียนรู้ ในขนาดชุดข้อมูลย่อยแต่ละขนาด บนชุดข้อมูล MNIST (N=10, L_n=2)



ภาพที่ 4.3 ความถูกของโมเดลหลักในกระบวนการเรียนรู้แบบสหพันธ์ในแต่รอบของการเรียนรู้ ในขนาดชุดข้อมูลย่อยแต่ละขนาด บนชุดข้อมูล FMNIST (N=10, L_n=2)

ผลลัพธ์ที่ได้แสดงให้เห็นว่าสำหรับชุดข้อมูลที่มีความซับซ้อนต่ำอย่าง MNIST จะเห็นได้ว่าเมื่อใช้ข้อมูลสังเคราะห์ในการแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน ด้วยข้อมูลที่ถูกสร้างขึ้นด้วยการเรียนรู้จากชุดขข้อมูลย่อยในขนาดต่างๆจะมีประสิทธิภาพด้านความถูกของโมเดลหลักเพิ่มขึ้นเป็นอย่างมากเทียบกับการไม่ใช้ข้อมูลสังเคราะห์ แต่ในการเรียนรู้ด้วยชุดข้อมูลย่อยในขนาดต่างๆนั้นแตกต่างกันไม่มากในแง่ของประสิทธิภาพของโมเดลหลัก โดยขนาดชุดข้อมูลย่อย 1, 25, และ 50 เปอร์เซ็นต์นั้นไม่มีความแตกต่างอย่างมีนัยยะสำคัญ กล่าวคือมีประสิทธิภาพเทียบเท่ากัน ในขณะที่ชุดข้อมูลย่อย 75 และ 100 เปอร์เซ็นต์นั้นมีประสิทธิภาพลดลงเล็กน้อยตามลำดับ โดยสำหรับชุดข้อมูล MNIST ค่าความถูกต้องของโมเดลจะเริ่มคงที่ตั้งแต่รอบที่ 10 ของการเรียนรู้แบบสหพันธ์เป็นต้นไป เว้นเพียงแต่โมเดลเรียนรู้กับที่ยังไม่ถูกแก้ปัญหาการกระจายตัวของข้อมูลจะได้ค่าความถูกต้องของโมเดลที่ไม่คงที่ ดังภาพที่ 4.2

สำหรับชุดข้อมูล FMNIST ที่มีความซับซ้อนของข้อมูลมากขึ้นนั้นจะได้ผลลัพธ์ในการเพิ่มประสิทธิภาพของโมเดลหลักเช่นเดียวกับชุดข้อมูล MNIST โดยมีประสิทธิภาพใกล้เคียงกันสำหรับชุดข้อมูลย่อย 1, 25 และ 50 เปอร์เซ็นต์ แต่สำหรับชุดข้อมูลย่อย 75 และ 100 เปอร์เซ็นต์นั้นจะมีประสิทธิภาพลดลงมากกว่า เมื่อเปรียบเทียบชุดข้อมูล MNIST ซึ่งผลลัพธ์ทั้งหมดสอดคล้องกับคะแนนคุณภาพของข้อมูล โดยหากมีคะแนนคุณภาพของข้อมูลสังเคราะห์ที่ดี ก็จะส่งผลให้สามารถแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันได้ดียิ่งขึ้น

4.4 การทดลองใช้ข้อมูลสังเคราะห์ในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของกระบวนการเรียนรู้แบบสหพันธ์

ในการทดลองนี้จะทำการใช้ข้อมูลสังเคราะห์เพื่อแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของข้อมูลบนการเรียนรู้แบบสหพันธ์ หลังจากที่ได้ขนาดของชุดข้อมูลย่อยและจำนวนข้อมูลสังเคราะห์ที่เหมาะสมแล้ว โดยจะทำการเปรียบเทียบโมเดลหลักที่ได้จากการเรียนรู้แบบสหพันธ์ ทั้งในรูปแบบที่มีการกระจายตัวของข้อมูลแบบปกติ ในรูปแบบที่มีการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกัน และในรูปแบบที่ใช้ข้อมูลสังเคราะห์ในการแก้ปัญหาดังกล่าว โดยจะใช้ชุดข้อมูล MNIST และ FMNIST ในการทดสอบ โดยทำการจำลองการกระจายตัวของข้อมูลที่ไม่

เหมือนกันและไม่เป็นอิสระต่อกัน ซึ่งจำนวนประเภทข้อมูลในเครื่องในสหพันธ์จะมีจำนวน 2 และ 5 ประเภท โดยมีจำนวนเครื่องในสหพันธ์เท่ากับ 10 และ 50 เครื่อง และใช้ข้อมูลสังเคราะห์ที่เรียนรู้ที่ขนาดชุดข้อมูลย่อยเท่ากับ 50% โดยใช้จำนวน 100 ภาพต่อประเภทข้อมูลต่อเครื่องในสหพันธ์

ตารางที่ 4.4 ตารางแสดงผลการใช้ข้อมูลสังเคราะห์ในการแก้ไขปัญหาการกระจายตัวของข้อมูล (N = 10, Round=15)

ชุดข้อมูล (DATASET)	การกระจายตัวของข้อมูล (DIST)	จำนวนประเภทข้อมูลในเครื่องในสหพันธ์ (L_n)	จำนวนข้อมูลสังเคราะห์ต่อประเภท ต่อเครื่อง	ความถูกต้องของโมเดลหลัก (%)
MNIST	IID	10	0	96.95
	Non-IID	5	0	96.55
		5	100	97.51
		2	0	63.58
		2	100	96.06
FMNIST	IID	10	0	82.07
	Non-IID	5	0	67.77
		5	100	78.11
		2	0	53.82
		2	100	79.83

ตารางที่ 4.5 ตารางแสดงผลการใช้ข้อมูลสังเคราะห์ในการแก้ไขปัญหาการกระจายตัวของข้อมูล (N = 50, Round = 15)

ชุดข้อมูล (DATASET)	การกระจายตัวของข้อมูล (DIST)	จำนวนประเภทข้อมูลในเครื่องในสหพันธ์ (L_n)	จำนวนข้อมูลสังเคราะห์ต่อประเภท ต่อเครื่อง	ความถูกต้องของโมเดลหลัก (%)
MNIST	IID	10	0	79.59
	Non-IID	5	0	62.14
		5	100	92.43
		2	0	43.16
		2	100	91.25
FMNIST	IID	10	0	63.81
	Non-IID	5	0	50.77
		5	100	72.68
		2	0	34.69
		2	100	72.11

ผลลัพธ์ที่ได้จากการทดลองแสดงให้เห็นว่าในกระบวนการเรียนรู้แบบสหพันธ์นั้น การที่มีเครื่องในสพันธ์เป็นจำนวนมากแม้จะมีการกระขายของข้อมูลที่ปกติ ยังส่งผลให้การควบคุมประสิทธิภาพของโมเดลมีความท้าทายมากขึ้น โดยสังเกตจากค่าความถูกต้องที่ลดลงเมื่อเพิ่มจำนวนเครื่องในสหพันธ์เป็น 50 เครื่อง โดยเฉพาะอย่างยิ่งกับข้อมูลที่มีการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันถึงแม้จะเป็นกรณีที่มีระดับของความรุนแรงที่ต่ำ เช่นชุดข้อมูล MNIST ที่มีความซับซ้อนน้อย หรือกรณีที่มีจำนวนประเภทข้อมูลเท่ากับ 5 (L_n = 5) ประสิทธิภาพของโมเดลหลักก็ยังลงอย่างมีนัยยะสำคัญ

สำหรับข้อมูลที่มีความซับซ้อนต่ำอย่าง MNIST นั้น เมื่อทำการแก้ปัญหาการกระจายที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันด้วยข้อมูลสังเคราะห์นั้น จะมีการเพิ่มขึ้นของประสิทธิภาพด้านความถูกต้องของโมเดลหลังสูงกว่าข้อมูลที่มีความซับซ้อนมากกว่าอย่าง FMNIST โดยทั้งสองชุดข้อมูลเมื่อมีการใช้ข้อมูลสังเคราะห์ในการแก้ปัญหาการกระจายตัวของข้อมูลแล้ว จะทำให้ความถูกต้องของโมเดล

สูงใกล้เคียงกับโมเดลหลักที่เรียนรู้บนระบบที่มีการกระจายตัวของข้อมูลแบบปกติสำหรับระบบที่มีจำนวนเครื่องในสหพันธ์เท่ากับ 10 เครื่อง

สำหรับข้อมูลที่มีความซับซ้อนมากขึ้นอย่าง FMNIST การแก้ไขปัญหาการกระจายตัวของข้อมูลด้วยภาพสังเคราะห์นั้นไม่เพียงแต่แก้ปัญหาความถูกต้องของโมเดลที่ต่ำลงเท่านั้น แต่ประสิทธิภาพที่ได้รับการปรับปรุงนั้นสูงขึ้นมาก จนสามารถได้ค่าความถูกต้องสูงกว่าโมเดลหลัก ที่เรียนรู้ด้วยข้อมูลที่มีการกระจายตัวแบบปกติบนกระบวนการเรียนรู้แบบสหพันธ์ ซึ่งแสดงให้เห็นว่าการเพิ่มข้อมูลนั้นไม่เพียงแต่จะสามารถแก้ปัญหาการกระจายตัวของข้อมูลได้แล้ว ยังสามารถใช้ในการเพิ่มประสิทธิภาพของโมเดลบนการเรียนรู้แบบสหพันธ์ได้อีกด้วย

ประสิทธิภาพด้านความถูกต้องของโมเดลหลักจะเพิ่มขึ้นเป็นอย่างมากเมื่อระบบมีเครื่องในสหพันธ์มากขึ้น เช่น 50 เครื่องในการทดลอง โดยผลลัพธ์ที่ได้นั้นโมเดลหลักจะมีความถูกต้องสูงกว่าโมเดลหลักที่เรียนรู้จากข้อมูลที่มีการกระจายตัวแบบปกติ เนื่องจากเมื่อมีจำนวนเครื่องมากขึ้นจะส่งผลให้จำนวนข้อมูลสังเคราะห์สูงขึ้นตามจากเทคนิคการเพิ่มข้อมูล จึงส่งผลให้โมเดลหลักนั้นมีประสิทธิภาพสูงขึ้นจากจำนวนข้อมูลที่มากขึ้น

4.5 สรุปสมมติฐานในการทดลอง

สมมติฐานที่ 1 เป็นจริง โดยการใช้ข้อมูลสังเคราะห์ที่ได้จากโมเดลแกนส์ในการแก้ปัญหาการกระจายตัวของข้อมูลบนกระบวนการเรียนรู้แบบสหพันธ์สามารถเพิ่มประสิทธิภาพของโมเดลให้เทียบเท่าหรือมากกว่าโมเดลที่เรียนรู้กับข้อมูลที่ไม่มีปัญหาการกระจายตัวของข้อมูล วัดผลได้จากตารางที่ 4.3

สมมติฐานที่ 2 เป็นจริง โดยเมื่อมีคะแนน FID Score ที่มาก จะส่งผลให้สามารถเพิ่มประสิทธิภาพของโมเดลที่ได้จากการเรียนรู้แบบสหพันธ์ได้มากขึ้น วัดผลจากภาพที่ 4.2 และ 4.3

สมมติฐานที่ 3 เป็นจริง แม้ว่าจะนำโมเดลที่มีปัญหาการกระจายตัวของข้อมูลมาเพิ่มข้อมูลด้วยด้วยภาพสังเคราะห์จากแกนส์ ก็สามารถทำให้มีประสิทธิภาพสูงกว่าโมเดลที่เรียนรู้กับข้อมูลที่ไม่มีปัญหาการกระจายตัวของข้อมูล ดังตารางที่ 4.5 จึงสรุปได้ว่าการเพิ่มข้อมูลไม่เพียงแต่สามารถแก้ไขปัญหาการกระจายตัวของข้อมูล แต่ยังสามารถเพิ่มประสิทธิภาพของโมเดลในกระบวนการเรียนรู้แบบสหพันธ์ได้อีกด้วย

สมมติฐานที่ 4 เป็นจริง เนื่องจากในการทดลองใช้ข้อมูลภาพที่น้อยกว่า 1,000 ภาพต่อประเภท ก็ยังสามารถแก้ไขปัญหาการกระจายตัวของข้อมูลได้เป็นอย่างดี วัดผลจากตารางที่ 4.5 และ 4.5

4.6 สรุปผลการทดลอง

ในการทดลองหาวิธีการในการแก้ปัญหาการกระจายตัวของข้อมูลที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันนั้น เทคนิคการเพิ่มข้อมูลเป็นวิธีที่ง่ายและตรงไปตรงมาที่สุดในการแก้ปัญหาดังกล่าว ในงานวิจัยนี้เลือกในแกนส์ในการสร้างข้อมูลสังเคราะห์เพื่อใช้ในการเพิ่มข้อมูลเพื่อแก้ไขปัญหาการกระจายตัวของข้อมูล แต่ในกระบวนการเรียนรู้ของแกนส์บนระบบการเก็บข้อมูลแบบกระจายศูนย์นั้นมีข้อจำกัด ทำให้แกนส์สามารถสร้างข้อมูลที่มีคุณภาพได้จากประเภทข้อมูลที่อยู่ในเครื่องในสหพันธ์เครื่องสุดท้ายที่ทำการเรียนรู้ได้เท่านั้น จึงจำเป็นต้องออกแบบกระบวนการเรียนรู้ของแกนส์บนระบบเก็บข้อมูลแบบกระจายศูนย์ เพื่อหลีกเลี่ยงการเกิดโอเวอร์ฟิตติ้ง

เพื่อให้สามารถสร้างข้อมูลที่มีคุณภาพได้ทุกประเภท จึงเป็นที่มาของการออกแบบระบบการเรียนรู้แบบชุดย่อย เพื่อให้ข้อมูลที่ป้อนเข้าโมเดลมีการคละกันอยู่ตลอดเวลา โดยที่ข้อมูลไม่จำเป็นต้องถูกโอนย้าย โดยเมื่อเปรียบเทียบคุณภาพของข้อมูลสังเคราะห์กับปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดแล้ว จะได้ขนาดของชุดข้อมูลย่อยที่เหมาะสมที่สุดคือ 50% เมื่อได้ข้อมูลสังเคราะห์ที่มีคุณภาพแล้ว ลำดับต่อมาคือการออกแบบการกระจายข้อมูลเพื่อแก้ปัญหา โดยในการทดลองจะทำการทดสอบประสิทธิภาพในการแก้ปัญหาการกระจายตัวของข้อมูลสำหรับขนาดข้อมูลสังเคราะห์ที่ 100 และ 1,000 ภาพต่อประเภทต่อเครื่องในสหพันธ์ โดยวัดผลจากความถูกต้องของโมเดลหลัก พบว่าจำนวนภาพที่มากขึ้นนั้น ส่งผลให้มีประสิทธิภาพของโมเดลหลักเพิ่มขึ้นเล็กน้อย ในขณะที่ใช้เวลาในการประมาณผลมากกว่าถึง 3 เท่า

ในการเปรียบเทียบคุณภาพของข้อมูลด้วยคะแนนนั้น ทางผู้วิจัยได้ทำการเปรียบเทียบคะแนนคุณภาพของข้อมูลสังเคราะห์ กับการนำไปใช้จริงในการแก้ปัญหาการกระจายตัวของข้อมูล โดยวัดผลจากความถูกต้องของโมเดลหลักที่เพิ่มมากขึ้น พบว่าการที่มีคะแนนคุณภาพข้อมูลที่สูงนั้นมีแปรผันกับความถูกต้องของโมเดลที่มีการพัฒนาเพิ่มขึ้น และในการแก้ปัญหาการกระจายตัวของข้อมูลด้วยข้อมูลสังเคราะห์นั้น วิธีการดังกล่าวสามารถแก้ปัญหาได้อย่างมีประสิทธิภาพ โดยวัดจาก

ความถูกต้องของโมเดลหลักที่เพิ่มขึ้นเทียบเคียงกับโมเดลหลักที่เรียนรู้บนระบบที่มีการกระจายตัวของข้อมูลแบบปกติ ซึ่งจะสามารถแก้ไขปัญหาได้ดีเมื่อชุดข้อมูลมีความซับซ้อนที่ต่ำ หรือระบบมีเครื่องในสหพันธ์เป็นจำนวนมาก

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้นำเสนอวิธีการแก้ปัญหาการกระจายตัวที่ไม่เหมือนกันและไม่เป็นอิสระต่อกันของกระบวนการเรียนรู้แบบสหพันธ์โดยใช้คอนดิชันนอลแกนส์ในการเพิ่มข้อมูลเพื่อแก้ปัญหา โดยได้ออกแบบกระบวนการเรียนรู้ของแกนส์บนระบบเก็บข้อมูลแบบกระจายศูนย์ ซึ่งทำการลดจำนวนข้อมูลในแต่ละรอบของการเรียนรู้ออกเป็นชุดข้อมูลย่อย เพื่อให้ได้มาซึ่งข้อมูลสังเคราะห์ที่มีคุณภาพโดยใช้ปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดน้อยที่สุด โดยจากการทดลองพบว่าขนาดชุดข้อมูลย่อย 50% เป็นค่าที่เหมาะสมที่สุดในการนำมาใช้ เพื่อให้ได้ข้อมูลสังเคราะห์ที่มีคุณภาพสูงโดยไม่ใช้ปริมาณข้อมูลที่ต้องมีการสื่อสารผ่านเครือข่ายทั้งหมดมากจนเกินไป และสำหรับจำนวนข้อมูลที่ใช้ในการแก้ปัญหาการกระจายข้อมูลสำหรับจำนวน 100 ภาพ และ 1,000 ภาพต่อประเภทข้อมูลต่อเครื่องในสหพันธ์นั้นมีประสิทธิภาพต่างกันไม่มาก โดยเปรียบเทียบจากความถูกต้องของโมเดลหลัก แต่การใช้ 1,000 ภาพนั้นจะเป็นการเพิ่มระยะเวลาของการเรียนรู้แบบสหพันธ์อย่างมาก การเลือกใช้จำนวนข้อมูลสังเคราะห์เท่ากับ 100 ภาพจึงมีความเหมาะสมมากกว่า โดยเมื่อนำข้อมูลสังเคราะห์ไปใช้ในการแก้ปัญหาการกระจายตัวของข้อมูลพบว่าสามารถแก้ไขปัญหาดังกล่าวได้อย่างมีประสิทธิภาพ โดยความถูกต้องของโมเดลหลักเพิ่มขึ้นสูงสุดถึง 45% โดยความถูกต้องของโมเดลหลักนั้นเพิ่มขึ้นจนใกล้เคียงหรือมากกว่าโมเดลหลักที่เรียนรู้กับข้อมูลบนระบบที่มีการกระจายตัวแบบปกติ และยังคงรักษาไว้ซึ่งความเป็นส่วนตัวของข้อมูล

5.2 ข้อจำกัดและแนวทางการวิจัยในอนาคต

ในงานวิจัยนี้จะใช้ชุดข้อมูลเพียง 2 ชุด ซึ่งเป็นชุดข้อมูลที่มีความซับซ้อนไม่มาก ซึ่งหากมีการใช้ชุดข้อมูลอื่นในการวิจัย โดยเฉพาะชุดข้อมูลที่มีความซับซ้อนสูง จึงอาจจะได้ผลลัพธ์ที่แตกต่างกันในแง่ของการเพิ่มประสิทธิภาพ เนื่องจากวิธีการที่นำเสนอมีแนวโน้มที่จะเพิ่มประสิทธิภาพได้ดีเมื่อทำงานกับชุดข้อมูลที่มีความซับซ้อนต่ำ รวมถึงการหาค่าที่เหมาะสมสำหรับตัวแปรควบคุมที่ใช้อาจมีการเปลี่ยนแปลงเพื่อให้เหมาะกับชุดข้อมูลใหม่ เช่น จำนวนชุดข้อมูลย่อย จำนวนข้อมูลสังเคราะห์ที่ใช้ เป็นต้น

รายการอ้างอิง

[1] Alex Castrounis, AI Explained (Online), 2022. Available: https://www.whyofai.com/blog/ai-explained

[2] Kritsada Arjchariyaphat, Federated Learning (Online), 2020. Available: https://medium.com/deaware/federated-learning-การเรียนรู้ของ-machine-learning-โดยไม่ต้องส่ง-raw-dataset-ไปที่ส่วนกลาง-c2a4c3538079

[3] พิชชากร วงศ์ตัะ , การอ่านเลขสายรถประจำทางจากภาพ, Undergraduate Thesis, วิศวกรรมคอมพิวเตอร์, วิศวกรรมศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย, 2560.

[4] Connor Shorten, DCGANs (Deep Convolutional Generative Adversarial Networks) (Online), 2018, Available: https://towardsdatascience.com/dcgans-deep-convolutional-generative-adversarial-networks-c7f392c2c8f8

[5] Aaron Pereira , TOWARDS FEDERATED LEARNING OVER LARGE-SCALE STREAMING DATA, Undergraduate Thesis, Department of Computer Science , Colorado State University, 2020.

[6] Nut Chukamphaeng, Generative Adversarial Networks (Online), 2019. Available: https://medium.com/@nutorbitx/gans-อะไรคือ-generative-adversarial-networks-7973ae70db70

[7] Ian J. Goodfellow, Generative Adversarial Nets, 2014. arXiv:1406.2661 [stat.ML]

[8] Neural Network History (Online), 2019. Available: https://medium.com/mmp-li/deep-learning-แบบฉบับคนสามัญชน-ep-1-neural-network-history-f7789236a9a3

[9] Satya Ganesh, The Role Of Weights And Bias In a Neural Network (Online), 2020. Available: https://towardsdatascience.com/whats-the-role-of-weights-and-bias-in-a-neural-network-4cf7e9888a0f

[10] The Neural Network (Online), 2020. Available: https://phusitsom.medium.com/พื้นฐาน-deep-learning-ทฤษฎี-how-does-ann-learn-d3b24ce2778a

[11] Rahul Jayawardana, ANALYSIS OF OPTIMIZING NEURAL NETWORKS AND ARTIFICIAL INTELLIGENT MODELS FOR GUIDANCE, CONTROL, AND NAVIGATION SYSTEMS, 2021. IRJETS e-ISSN: 2582-5208

[12] David E. Rumelhart, Backpropagation, 1995. Psychology Press

[13] Frederik Kratzert, Understanding the backward pass through Batch Normalization Layer (Online), 2016. Available: https://kratzert.github.io/2016/02/12/understanding-the-gradient-flow-through-the-batch-normalization-layer.html

[14] Pisit Bee, Backpropagation (Online), 2018. Available: https://medium.com/boobeejung/backpropagation-a0c8c6363192

[55] Zijian Li, Federated Learning with GAN-based Data Synthesis for Non-IID Clients, 2022. Available: https://arxiv.org/pdf/2206.05507.pdf

[16] Kunihiko Fukushima, Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position, 1980. NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

[17] Vithan Minaphinant, Machine Learning (Online), 2018. Available: https://medium.com/investic/machine-learning-คืออะไร-fa8bf6663c07

[18] Vithan Minaphinant, Deep Learning (Online), 2018. Available: https://medium.com/investic/deep-learning-คืออะไร-อาชีพไหนจะตกงานบ้าง-499c250784a1

[19] Natthawat Phongchit, Convolutional Neural Network (CNN) คืออะไร (Online), 2018. Available: https://medium.com/@natthawatphongchit/มาลองดูวิธีการคิดของ-cnn-กัน-e3f5d73eebaa

[20] Ankita Wadhawan, Deep learning-based sign language recognition system for static signs, 2020. Neural Computing and Applications 32(2):1-12

[21] Manpreet Kaur, When Cartoon Meets Anime Distinguishing Animation Styles with Convolutional Neural Networks, 2020. International Conference on Intelligent Engineering and Management (ICIEM)

[22] Brendan McMahan and Daniel Ramag, Federated Learning: Collaborative Machine Learning without Centralized Training Data, 2017. Google Research

[23] NICOLA RIEKE, What Is Federated Learning? (Online), 2019. Available: https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/

[24] Hangyu Zhu, Federated Learning on Non-IID Data: A Survey, 2021. arXiv:2106.06843 [cs.LG]

[25] Qinbin Li, Federated Learning on Non-IID Data Silos: An Experimental Study, 2021. arXiv:2102.02079v4 [cs.LG]

[26] Brendan McMahan, Communication-efficient learning of deep networks from decentralized data, 2017. International Conference on Artificial Intelligence and Statistics (AISTATS)

[27] Ian J. Goodfellow, Generative Adversarial Nets, 2017. Advances in Neural Information Processing Systems, Curran Associates, Inc.

[28] Henrik Hellström, Wireless for Machine Learning, 2020. Student Member IEEE.

[29] Jason Brownlee, How to Evaluate Generative Adversarial Networks (Online),2019. Available: https://machinelearningmastery.com/how-to-evaluate-generative-adversarial-networks/

[30] Jason Brownlee, How to Implement the Inception Score (IS) for Evaluating GANs, 2019. Available: https://machinelearningmastery.com/how-to-implement-the-inception-score-from-scratch-for-evaluating-generated-images/

[31] Jason Brownlee, How to Implement the Frechet Inception Distance (FID) for Evaluating GANs, 2019. Available: https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/

[32] Tim Salimans, Improved Techniques for Training GANs, 2016. arXiv:1606.03498 [cs.LG]

[33] Martin Heusel, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2017. arXiv:1706.08500 [cs.LG]

[34] Wei Li, IFL-GAN: Improved Federated Learning Generative Adversarial Network With Maximum Mean Discrepancy Model Aggregation, 2022. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

[35] Alex Krizhevsky, The CIFAR-10 dataset, 2009. The Department of Computer Science , University of Toronto

[36] Jia Deng, ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), Stanford University

[37] Agrawal, Sachin. Leaf Features Extraction for Plant Classification using CNN, 2021 International Journal of Advanced Research in Science, Communication and Technology. 10.48175/IJARSCT-807.

[38] Qian, Bin & Su, Jie & Wen, Zhenyu & Yang, Renyu & Zomaya, Albert & Rana, Omer. Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey, 2019. Research Gate

[39] Xiaodong Ma, A state-of-the-art survey on solving non-IID data in Federated Learning, 2022. Future Generation Computer Systems, Volume 135, 2022, Pages 244-258

[40] Seong-Lyun Kim, Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data, 2018. arXiv:1811.11479 [cs.LG]

[41] Liyang Xie, Differentially Private Generative Adversarial Network, 2018. arXiv:1802.06739v1 [cs.LG]

[42] Xingjian Cao, PerFED-GAN: Personalized Federated Learning via Generative Adversarial Networks, 2022. IEEE Internet of Things Journal, doi: 10.1109/JIOT.2022.3172114

[43] Yue Zhao, Federated Learning with Non-IID Data. 2018. arXiv:1806.00582 [cs.LG]

[44] Cynthia Dwork, The Algorithmic Foundations of Differential Privacy, 2014. University of Pennsylvania, USA

[45] PHILIPP JOSEPH, Differential Privacy – Privacy-preserving data analysis (Online), 2018. Available: https://blog.mi.hdmstuttgart.de/index.php/2018/08/14/differential-privacy/

[46] Seung-Hwan Lim, An analysis of image storage systems for scalable training of deep neural networks, 2016. ResearchGate

[47] Saeed Reza Kheradpisheh, BS4NN: Binarized Spiking Neural Networks with Temporal Coding and Learning, 2020. ResearchGate

[48] Mehdi Mirza, Conditional Generative Adversarial Nets, 2014.

Available : https://arxiv.org/abs/1411.1784

[49] Aditya Sharma, Conditional GAN (cGAN) in PyTorch and TensorFlow, 2021. Available : https://learnopencv.com/conditional-gan-cgan-in-pytorch-and-tensorflow/

[50] MNIST (Online), 2022. Available : https://datasets.activeloop.ai/docs/ml/datasets/mnist/

[51] Hello FASHION MNIST! (Online), 2022. Available : https://torchfusion.readthedocs.io/en/latest/training/hello.html

[52] Deng, L., 2012. The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), pp.141–142.

[53] Michael Weiss, 2022. Simple Techniques Work Surprisingly Well for Neural Network Test Prioritization and Active Learning. Proceedings of the 31th ACM SIGSOFT International Symposium on Software Testing and Analysis.

[54] Shane Barratt, 2018. A Note on the Inception Score. ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models.

ประวัติผู้เขียน

ชื่อ	ฐิติ ชื่นบุบผา
วุฒิการศึกษา	ปีการศึกษา 2561 : วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ประยุกต์ มหาวิทยาลัยธรรมศาสตร์
ทุนการศึกษา	ปีการศึกษา 2562 – 2563: ทุนบัณฑิตเรียนดีเพื่อศึกษา ต่อระดับบัณฑิตศึกษา คณะวิทยาศาสตร์และเทคโนโลยี

ผลงานทางวิชาการ

[1] Chuenbubpha Thiti et al., Solving Non-IID in Federated Learning for Image Classification using GANs, 20th The International Joint Conference on Computer Science and Software Engineering (JCSSE), Phitsanulok, Thailand, 2023, pp. 333-338.