



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

# AN OPTIMIZED FEDERATED LEARNING FRAMEWORK FOR INTERNET OF VEHICLES BASED ON NOISE DATA AND INCREMENTAL DATA PROCESSING



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

LEI YUAN

SULTAN IDRIS EDUCATION UNIVERSITY

2024



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

AN OPTIMIZED FEDERATED LEARNING FRAMEWORK FOR INTERNET OF  
VEHICLES BASED ON NOISE DATA AND INCREMENTAL DATA  
PROCESSING

LEI YUAN



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi

THESIS PRESENTED TO QUALIFY FOR A DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTING & META-TECHNOLOGY  
SULTAN IDRIS EDUCATION UNIVERSITY

2024



05-4506832



pustaka.upsi.edu.my



Perpustakaan Tuanku Bainun  
Kampus Sultan Abdul Jalil Shah



PustakaTBainun



ptbupsi



Please tick (✓)

Project Paper

Masters by Research

Master by Mixed Mode

PhD

✓

## INSTITUTE OF GRADUATE STUDIES

### DECLARATION OF ORIGINAL WORK

This declaration is made on the **11<sup>TH</sup> JUNE 2024**

#### i. Student's Declaration:

I, **LEI YUAN (P20202001386) FACULTY OF COMPUTING & META-TECHNOLOGY** (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled **AN OPTIMIZED FEDERATED LEARNING FRAMEWORK FOR INTERNET OF VEHICLES BASED ON NOISE DATA AND INCREMENTAL DATA PROCESSING** is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

Signature of the student

#### ii. Supervisor's Declaration:

I **ASSOC. PROFESSOR DR. WANG SHIR LI** (SUPERVISOR'S NAME) hereby certify that the work entitled **AN OPTIMIZED FEDERATED LEARNING FRAMEWORK FOR INTERNET OF VEHICLES BASED ON NOISE DATA AND INCREMENTAL DATA PROCESSING** (TITLE) was prepared by the above-named student, and was submitted to the Institute of Graduate Studies as a \* partial/full fulfillment for the conferment of **DOCTOR OF PHILOSOPHY** (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

11 JUNE 2024

Date

Prof. Madya Dr. Wang Shir Li  
Pensyarah  
Fakulti Komputeran dan Meta-Teknologi  
Universiti Pendidikan Sultan Idris  
35900 Tanjung Malim, Perak.

Signature of the Supervisor

INSTITUT PENGAJIAN SISWAZAH /  
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK  
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / Title: AN OPTIMIZED FEDERATED LEARNING FRAMEWORK FOR  
INTERNET OF VEHICLES BASED ON NOISE DATA AND  
INCREMENTAL DATA PROCESSING

No. Matrik / Matric's No.: P20202001386

Saya / I : LEI YUAN

(Nama pelajar / Student's Name)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)\* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-

*acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-*

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.

*The thesis is the property of Universiti Pendidikan Sultan Idris*

2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.

*Tuanku Bainun Library has the right to make copies for the purpose of reference and research.*

3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.

*The Library has the right to make copies of the thesis for academic exchange.*

4. Sila tandakan ( ✓ ) bagi pilihan kategori di bawah / Please tick ( ✓ ) for category below:-

☐ **SULIT/CONFIDENTIAL**


Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / Contains confidential information under the Official Secret Act 1972

☐ **TERHAD/RESTRICTED**

Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / Contains restricted information as specified by the organization where research was done

☒ **TIDAK TERHAD/OPEN ACCESS**

  
(Tandatangan Pelajar/ Signature)

  
(Tandatangan Penyelia / Signature of Supervisor)  
& (Nama & Cop Rasmi / Name & Official Stamp)

Tarikh: 11 JUNE 2024

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

## ACKNOWLEDGEMENT

Here, I find myself overflowing with genuine gratitude for Assoc. Prof. Dr. Shir Li Wang. Her relentless pursuit of academic excellence, her painstaking attention to detail when responding to journal reviewers' comments, have illuminated my path in academic research. More than her profound understanding of research directions and trends, it's her radiant positivity and optimism about life that have deeply touched my soul. The opportunity to learn from such an inspiring supervisor has been a privilege, an experience that has imparted me with valuable knowledge and insight. This profound influence she has had on me, will undoubtedly ripple through my future research and beyond, shaping my life in ways I can only begin to appreciate. In addition, my heartfelt thanks go to my alma mater—Universiti Pendidikan Sultan Idris (UPSI). Despite the barriers of the pandemic that made physical exploration impossible, the spirit and charm of UPSI were never out of reach, ever present in the digital realm. It wasn't until I had spent a few months here that I truly experienced UPSI in its full splendor. UPSI, a sanctuary that seems untouched by time, provides an inspiring backdrop for learning and research. The enriching environment and atmosphere here have been instrumental in my academic journey, allowing me to delve deeper into my research.



## ABSTRACT

Internet of Vehicles (IoV) technology has been rapidly advancing, making intelligent transportation systems the future trend. This research revolves around building efficient and secure vehicular networks using data processing mechanisms backed by machine learning and information security. However, noise and incremental data present challenges to vehicular network development. This study proposes two novel federated learning frameworks, namely the Outlier Detection and Exponential Smoothing Federated Learning (OES-FED) and Federated Learning Framework Based on Incremental Weighting and Diversity Selection for IoV (FED-IW&DS), to overcome the above problems. The OES-FED framework leveraged anomaly detection and exponential smoothing to filter noise data, thus, improving model robustness and enhancing communication efficiency. In terms of accuracy, it outperformed the existing Federated Learning-Average (FED-AVG) and FED-SGD models on three datasets by 44.46% and 2.36%, respectively. Furthermore, the FED-IW&DS framework that integrates incremental weights and diversity selection to effectively deal with issues of growing data scale was able to achieve rapid information sharing while preserving user privacy. The superiority of FED-IW&DS was clearly proven through its performance on two data sets, which found its accuracy to exceed that of the Fed-prox model by 30-35%. Ultimately, integrating the OES-FED and FED-IW&DS frameworks unveiled two critical integration points: the execution order and transition point of the two frameworks. By synergistically integrating the two frameworks, the proposed strategy unlocked new federated learning solutions for IoV as it yielded up to 5-10% higher accuracy compared to employing either framework individually. This study highlights novel approaches that address noise and incremental data challenges in IoV, yielding substantial advancements in both theoretical research and practical applications. The research outcomes have several implications, of which the proposed solutions play an essential role in improving communication efficiency, enhancing data processing capabilities, protecting user privacy, and providing crucial theoretical support and practical reference for future research and optimization of data processing mechanisms in IoV.





## **KERANGKA PEMBELAJARAN BERSEKUTU YANG DIOPTIMUMKAN UNTUK INTERNET KENDERAAN BERDASARKAN PEMPROSESAN DATA HINGAR DAN DATA TAMBAHAN**

### **ABSTRAK**

Teknologi Internet Kenderaan (IoV) telah pesat membangun yang menjadikan system pengangkutan pintar sebagai trend masa depan. Penyelidikan ini berkisar tentang pembinaan rangkaian kenderaan yang cekap dan selamat dengan menggunakan mekanisme pemprosesan data yang disokong oleh pembelajaran mesin dan keselamatan maklumat. Walau bagaimanapun, data hingar dan data tambahan merupakan cabaran kepada pembangunan rangkaian kenderaan. Kajian ini mencadangkan dua kerangka pembelajaran bersekutu yang baharu, iaitu Pembelajaran Bersekutu Pengesanan Unsur Luaran dan Pelicin Eksponen (OES-FED) dan Pembelajaran Bersekutu Berdasarkan Pemberat Tambahan dan Pemilihan Kepelbagaian untuk IoV (FED-IW&DS), untuk mengatasi masalah di atas. Kerangka OES-FED menggunakan pengesanan penyimpangan dan pelicin eksponen untuk menapis data hingar yang dapat meningkatkan keteguhan model dan kecekapan komunikasi. Dari sudut ketepatan, ia mengatasi model Pembelajaran Bersekutu Purata (FED-AVG) dan FED-SGD sedia ada berdasarkan tiga set data sebanyak 44.46% dan 2.36% masing-masing. Tambahan pula, kerangka FED-IW&DS yang menggabungkan pemberat tambahan dan pemilihan kepelbagaian untuk menangani isu pertambahan skala data secara berkesan telah berjaya mencapai perkongsian maklumat pantas sambil memelihara kerahsiaan pengguna. Kelebihan FED-IW&DS jelas terbukti melalui prestasinya pada dua set data yang mendapati ketepatannya melebihi model Fed-prox sebanyak 30-35%. Akhirnya, pengintegrasian kerangka OES-FED dan FED-IW&DS telah mendedahkan dua titik integrasi kritikal: turutan pelaksanaan dan titik peralihan kedua-dua kerangka tersebut. Dengan menggabungkan kedua-dua kerangka secara sinergi, strategi yang dicadangkan dapat menyediakan penyelesaian pembelajaran bersekutu yang baharu untuk IoV memandangkan ia menghasilkan ketepatan sehingga 5-10% lebih tinggi berbanding menggunakan mana-mana kerangka secara individu. Kajian ini menyerlahkan pendekatan baharu untuk menangani cabaran data hingar dan data tambahan dalam IoV yang dapat menghasilkan kemajuan besar dalam penyelidikan teori dan aplikasi praktikal. Hasil kajian mempunyai beberapa implikasi, antaranya adalah meningkatkan kecekapan komunikasi, memantapkan keupayaan pemprosesan data, melindungi kerahsiaan pengguna, serta menyediakan sokongan teori penting dan rujukan praktikal untuk penyelidikan masa depan dan pengoptimuman mekanisme pemprosesan data dalam IoV.



## TABLE OF CONTENTS

	Page
<b>DECLARATION OF ORIGINAL WORK</b>	ii
<b>DECLARATION OF THESIS SUBMISSION</b>	iii
<b>ACKNOWLEDGEMENT</b>	iv
<b>ABSTRACT</b>	vi
<b>TABLE OF CONTENT</b>	vii
<b>LIST OF TABLES</b>	xiii
<b>LIST OF FIGURES</b>	xiv
<b>LIST OF ABBREVIATIONS</b>	xvi
<b>CHAPTER 1 INTRODUCTION</b>	
1.1 Introduction	1
1.2 Research Background	2
1.2.1 Federated Learning	2
1.2.2 Image Processing in the Internet of Vehicles	7
1.2.3 Why Federated Learning Combined with IoV?	14
1.3 Problem Statement	25
1.4 Research Objectives	28
1.5 Research Questions	29
1.6 Research Significance	30
1.7 Research Scope	33
1.8 Thesis Structure	35



1.9 Chapter Summary	35
---------------------	----

## CHAPTER 2 LITERATURE REVIEW

2.1 Introduction	38
2.2 Systematic Review	40
2.2.1 Method	42
2.2.2 Literature Collection	42
2.2.3 Literature Inclusion Criteria	44
2.2.4 Literature Exclusion Criteria	44
2.3 Taxonomy Analysis	48
2.3.1 Review of Overview Articles	48
2.3.2 Insights from Review Articles	50
2.3.3 Review Articles Combining Federated Learning and IoV	59
2.3.3.1 Perceptual Layer	63
2.3.3.2 Network Layer	64
2.3.3.3 Application Layer	66
2.3.4 Research Trends of IoV	67
2.3.5 Privacy Protection in IoV	69
2.3.6 Optimization Method/Technology of Federated Learning Framework in IoV	78
2.3.6.1 Federated Learning Framework in IoV	80
2.3.6.2 Noise Data Problem with IoV	86
2.3.6.3 Incremental Data and Data Diversity in IoV	90
2.4 Discussion	96
2.4.1 Challenges	96

2.4.1.1	Privacy Data Protection in Internet of Vehicles (IoV)	96
2.4.1.2	Noise Data Processing in the Internet of Vehicles	109
2.4.1.3	The Conflict between Incremental Data and Existing Data in IoV	111
2.4.2	Motivations	113
2.4.3	Suggestions	115
2.4.3.1	Federated Learning Framework	115
2.4.3.2	Optimization of Image Processing in Federated Learning	119
2.4.3.3	Advantages of Federated Learning	129
2.4.3.4	On Noise Data in IoV	130
2.4.3.5	Incremental Data Challenge	132
2.5	Critical Review	132
2.6	Chapter Summary	134

### CHAPTER 3 METHODOLOGY

3.1	Introduction	136
3.2	Methodology Overview	137
3.3	Experimental Procedure	139
3.4	Experimental Data Set	141
3.4.1	Mnist	141
3.4.2	CIFAR-10	142
3.4.3	Vehicle Classification	144
3.5	Noise Data Processing	147
3.5.1	Parameter Introduction	147

3.5.2

Outlier-Based K-Means Algorithm

149

3.5.2.1

Outlier Algorithm

150

3.5.2.2

Improved Outlier-Based K-Means Algorithm

151

3.5.2.3

Current Perspective on Noise Data: The Improved K-Means Scheme based on the Outlier Algorithm

153

3.5.3

Exponential Smoothing Algorithm Based on Kalman Filtering

155

3.5.3.1

Kalman Filtering Algorithm

155

3.5.3.2

Exponential Smoothing Algorithm

162

3.5.3.3

Historical Perspective of Noise Data: The Improved Exponential Smoothing Scheme based on Kalman Filter

163

3.5.4

Outlier Detection and Exponential Smoothing Federated Learning (OES-Fed) Framework

166

3.6

Incremental Data Processing

168

3.6.1

Section Background

168

3.6.2

Cosine Distance

171

3.6.3

Diversity Score Calculation

173

3.6.3.1

Diversity Score Calculation - Initial Process

173

3.6.3.2

Diversity Score Calculation - Improvement Based on Penalty Factor

174

3.6.4

Incremental Weight

178

3.6.4.1

Incremental Weighted Calculation

179

3.6.5	Improved Incremental Learning Scheme based on Cosine Distance and Diversity Score	180
3.6.5.1	Motivation of Adjustment variable $\alpha$ for Parameter Depth-Value	181
3.6.5.2	Incremental Parameter Depth-Value Calculation	183
3.6.6	Federated Incremental Weighting & Diversity Score (Fed-IW&DS) Framework	188
3.7	Federated Learning Combining Noise Data and Incremental Data in Iov	190
3.8	Evaluation Criterion	194
3.9	Research Process of The Proposed Method	202
3.10	Description of Experimental Parameters	204
3.11	Chapter Summary	206

## CHAPTER 4 FINDINGS

4.1	Introduction	209
4.2	Experimental Modeling	210
4.2.1	Noise Data Processing Experiment	210
4.2.2	Results of Accuracy, Loss, and AUC of OES-FED Framework	213
4.2.3	Incremental Data Processing Experiments	225
4.2.4	Fed-IW&DS: The Acc, Loss, MCC Results and Computational Time Cost	234
4.2.5	Possibilities Testing	246
4.2.6	Possibilities Validation	249
4.3	Chapter Summary	259

## CHAPTER 5 CONCLUSIONS AND FUTURE WORK

5.1	Summary of Research Findings	265
-----	------------------------------	-----

5.1.1	OES-FED Framework	267
5.1.2	Fed-IW&DS Framework	271
5.2	Contributions to The Field	277
5.2.1	Theoretical Contributions	277
5.2.2	Practical Implications	280
5.3	Challenges	284
5.3.1	Scalability and Robustness	285
5.3.2	Communication Overhead and Latency	288
5.4	Future Research Directions	291
5.4.1	Advanced Machine Learning Techniques	291
5.4.2	Novel Privacy-Preserving Mechanisms	293
5.4.3	Integration with Emerging Technologies	295
5.5	Concluding Remarks	301

## REFERENCES

306



## LIST OF TABLES

Table No.		Page
1.1	IoV by Authoritative Statistics	10
2.1	Comparison of Federal Learning Algorithms for Different Domains	89
2.2	Comparison of Different Frameworks	93
3.1	Mathematical Symbols and Explanations	148
3.2	K-Means Clustering Step	152
3.3	Parameter Setting of the Experiment	206
4.1	Comparison of Accuracy, Loss Values and AUC Values for the MNIST Dataset, CIFAR-10 and the Vehicle Classification Dataset Using FedAVG, FedSGD and OES-FED	227
4.2	Each Specific Parameter of the Experiment	228
4.3	Impact of Different Diversity-Ratio Values	233
4.4	Time Cost	244
4.5	Proportion of Incremental Data in Each Round	255



## LIST OF FIGURES

Figure No.		Page
1.1	Federal Learning (Source: Wikipedia)	4
1.2	Internet of Vehicles (Source: Author's Own Photograph)	8
2.1	Flowchart of Study Selection, Search Query, and Inclusion Criteria	47
2.2	Graph of the Number of Literatures Analyzed by Database	49
2.3	Keyword Analysis	50
2.4	Telematics Architecture Description	62
2.5	The Citation Relationship between "Federated Learning", "Incremental Learning" and "Internet of Vehicles"	94
2.6	The Citation Relationship between "Federated Learning", as well as the Number of Publications and Citations in the Last Five Years	94
2.7	The Citation Relationship between "Incremental Data", as well as the Number of Publications and Citations in the Last Five Years	95
2.8	The Citation Relationship between "Internet of Vehicles", as well as the Number of Publications and Citations in the Last Five Years	95
3.1	Research Process Framework	140
3.2	Display of CIFAR-10 Data	144
3.3	Experimental Images Part 1	145
3.4	Experimental Images Part 2	146
3.5	Experimental Images Part 3	146
3.6	OES-Fed Framework	149
3.7	Our Fed-IW&DS Framework	169

3.8	Improved Arctangent Function Incremental Parameter Depth-Value	186
3.9	Threshold Conjecture	194
4.1	Statistics on the number of Actual Clients in the OES-Fed Model Using the MNIST Dataset, the CIFAR-10 Dataset and the Vehicle Classification Dataset	212
4.2	Comparison of the Accuracy of Each Model for MNIST, CIFAR-10 and the Vehicle Classification Datasets Using Different Models and Non-IID Data Settings	215
4.3	Comparison of Loss Values for Each Model for MNIST, CIFAR-10 and the Vehicle Classification Datasets Using Different Models and Non-IID Data Settings	216
4.4	Comparison of AUC Values for Each Model for MNIST Using Different Models and Non-IID Data Settings	217
4.5	Comparison of AUC Values for Each Model for CIFAR-10 Using Different Models and Non-IID Data Settings	217
4.6	Comparison of AUC Values for Each Model for Vehicle Classification Dataset Using Different Models and Non-IID Data Settings	221
4.7	Accuracy Comparison of all Clients in the Last Round for MNIST Dataset, CIFAR-10 Dataset and the Vehicle Classification Dataset Using OES-Fed Model and FedAVG Model	223
4.8	Comparison of Accuracy	236
4.9	Comparison of Loss	238
4.10	Comparison of MCC	242
4.11	Algorithm Conversion Function	247
4.12	Experimental Threshold Validation Results	253





## LIST OF ABBREVIATIONS

ACC	Accuracy
AI	Artificial intelligence
AP	Access Point
AUC	Area Under the Curve
BGD	Batch Gradient Descent
CA	Cloud Aggregator
CF	Catastrophic Forgetting
CIFAR-10	Canadian Institute For Advanced Research
CLCPPA	Certificate-Less CPPA
CNIL	French National Commission for Information and Freedom
CNN	Convolutional Neural Networks
DAG	Directed Acyclic Graph
DANN	Domain Adaptive Neural Networks
DNN	Deep Neural Networks
D-R	Diversity-Ratio
DRL	Deep Reinforcement Learning
EDPB	European Data Protection Committee
ES	Exponential Smoothing
ET	Edge Trainer
FedAvg	Federated Averaging
Fed-IW&DS	Federated Incremental Weighting & Diversity Score





Fed-prox	Federated Optimization in Heterogeneous Networks
FedSGD	FederatedSGD
FedSLD	Federated Learning of Shared Label Distribution
FL	Federated learning
FSD	Fully Automatic Driving
FVN	Federated Vehicle Network
GAN	Generative Adversarial Network
HGNAS	Heterogeneous Graph Neural Architecture Search
HT	Hierarchical Tucker
IL	Imitation learning
IoT	Internet of Things
ITS	Intelligent Transportation System
KD	Knowledge Distillation
LDP	Local Differential Privacy
LOSS	loss value
IoV	Internet of Vehicles
MCC	Matthews Correlation Coefficient
MEC	Mobile Edge Computing
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology database
MPC	Multi-Party Computation
NAS	Neural Architecture Search
non-IID	Non-Independently Identically Distribution
OES-FED	Outlier Detection and Exponential Smoothing federated learning
PID	Pseudo-Identity



PPFL	Privacy-Preserving Federated Learning
QDNNs	Quantized Deep Neural Networks
RSU	Roadside Unit
SAE	Society of Automotive Engineers
SAFH	Adaptive Feedback Handover
SC	Smart Cars
SDN	Software Defined Network
SES	Single Exponential Smoothing
SFLEC	Secure Federated Learning and Efficient Communication
SJR	SCI Mago Journal Rank
SMC	Secure Multiparty Computation
SNIP	Source Normalized Impact per Paper
TAs	Trusted Authorities
TR	Tensor Ring
TT	Tensor Training
TTP	Trusted Third Party
V2C	Between Vehicles and Data Centers
V2I	Vehicle to Infrastructure
V2V	Vehicle to Vehicle
V2X	Vehicle-to-Everything
VANET	Vehicular Ad-hoc Network
VFC	vehicle platform computing
VI	Vehicle Interactor
VRF	Verified Random Function

## CHAPTER 1

### INTRODUCTION

This chapter mainly introduces the research background, problem statement, research questions, research objects, research significance, research scope, and thesis organization. Nowadays, with the gradual increase in privacy awareness, a large amount of private data cannot be shared. Against this background, federated learning, which can perform machine learning training without sharing data, has attracted much attention since its emergence. It is applicable in many fields, and one of them is the Internet of Vehicles (IoV). However, privacy is not the only issue in IoV. In the era of big data, IoV can generate vast volumes of information in real time, but the effectiveness of the produced data is poor and full of noisy data. In addition, the global machine learning (ML) model cannot cover everything during training (Ardabili et al.,



2020; Rasouli & Yu, 2021). For example, the performance deteriorates when faced with new types of incremental data.

This study combines related techniques by reviewing the advancement of federated learning and machine learning. We effectively improve the noise data filtering ability, better integrate incremental data and existing data, and help alleviate the contradiction between the privacy requirements of car owners and the performance improvement the ML model.

## 1.2 Research Background



### 1.2.1 Federated Learning

Artificial intelligence (AI) has been widely used in recent years, for instance, speech recognition, recommender systems, computer vision, vehicle-assisted driving, etc (Arias-Otalora et al., 2022; Bhatia & Singh, 2022; Das, 2019). The common feature of these technical applications is that they are based on a substantial body of evidence. From these data, AI models learn specific abilities to complete various complex tasks and even perform operations that are difficult for human beings. Training an AI model requires massive amounts of data. For example, Facebook's face detection system was trained on 350 million images (Kosinski, 2021). Another example is that Tesla, an automobile company, is trying to realize the Society of Automotive Engineers (SAE) Level 5 fully automatic driving (FSD) method, which uses a few million Tesla drivers'





behaviors to train neural networks (Li et al., 2020). These drivers primarily make use of visible light cameras and images from other automotive components, such as the ultrasonic sensors for parking and the coarse-grained two-dimensional maps utilized for navigation.

However, with the continuous occurrence of data leaks and privacy violations in the recent past, people have started to pay serious attention to whether their private data is being used without their consent or being used by others for commercial or political purposes. People gradually realize the importance of protecting data privacy for the data being used to train AI models. Moreover, the gradual strengthening of people's privacy awareness has prompted the formulation of relevant data laws and guidelines governing privacy. According to General Data Protection Regulation in Europe, top Internet companies have been fined continuously for leaking personal data. In 2021, Irish authorities fined WhatsApp €225 million (Rossini et al., 2021). In 2022, the French data regulator CNIL (French National Commission for Information and Freedom) fined Google and Facebook 150 million euros each and 60 million euros each for making it difficult for French users to opt out of cookie tracking (Nicholsono et al., 2022). Luxembourg's main data privacy authority fined Amazon a record-breaking 746 million euros (Li et al., 2021). In such a situation, organizing and collecting data will be extremely difficult. On the one hand, people will keep data confidential because of privacy, and companies cannot obtain enough data for training machine learning. On the other hand, due to regulatory reasons, different organizations and departments cannot share their small data and then cannot form big data, thus forming isolated data islands.



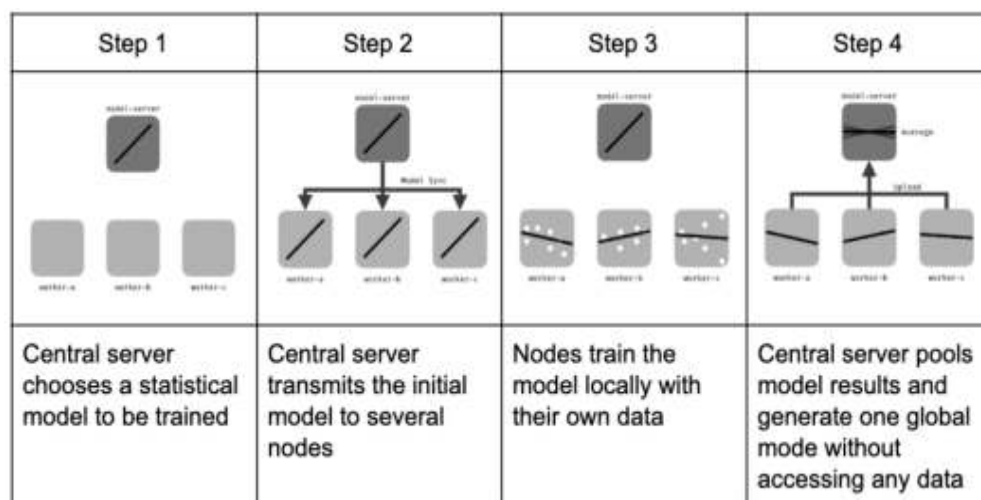


Most training data are produced and owned by individuals and departments in different organizations. The conventional approach to machine learning collects and transmits the data to a central server, which can read and control it. Therefore, this central server needs high-performance computing power to build machine learning models and handle sensitive data to avoid revealing data privacy. However, this method requires complete trust in the central server, which is unrealistic. In this case, data owners tend to hold their data in their own hands, thus forming isolated data islands. As a result, the foundation of vast amounts of data has disappeared, and artificial intelligence will enter a desperate situation. Therefore, how to extract useful information from scattered information with the idea that observing strict protection of privacy laws and regulations has become the primary problem in the development of contemporary AI.



**Figure 1.1**

*Federated Learning (Source: Wikipedia)*



A great solution is provided by federated learning. Federated learning is a distributed machine learning framework first provided by Google in 2016. Multiple





clients can be trained global ML models collaboratively in federated learning, thus preventing the training data from leaving the local client (as shown in Figure 1-1). As a new framework of distributed machine learning was initially proposed by the Google team, aiming at solving the problem of global input predictions that are updated for many mobile phone users. Most telephone users do not want to share their input, which is very private, while the global model needs to get input from many users (the more, the better). Federated learning aims to improve global ML predictions by sharing local ML model parameters while ensuring local data privacy. Traditional centralized learning upload local data from mobile devices to cloud computing centers or edge servers for global ML model education. However, each mobile device in the federated learning framework performs local ML model training and uploads the local ML models' parameters to the central server. Federated learning completes global ML model training in a distributed fashion. The mobile device only needs to exchange the local ML model parameters in this process. However, it does not need to share the local data, which protects the user's security and privacy (Chen et al., 2020; Choi & Pokhrel, 2020; Zhang et al., 2021). In addition, the model of federated learning only needs to upload the ML model parameters and does not involve uploading local data, significantly reducing the communication cost. It effectively solves the problems of high data transmission cost and privacy leakage risk in traditional centralized learning (Pillutla et al., 2022; Tseng et al., 2022; Xu et al., 2021).

In federated learning, there are many clients and one server (for aggregating global ML model parameters). It differs from centralized machine learning in many ways: (1) The central server is no longer trusted and is replaced by an ML model parameter aggregator. Aggregating the local ML model parameters submitted by the







clients is the central server's primary function. (2) The clients always save the data locally and only upload the parameters of the local ML model. The central server cannot collect the private data of each client. Federated learning is a form of networked machine learning that realize collaborative learning between clients who do not trust each other without sharing original data between data owners. Federated learning can fully use each participant's computing power, improve learning efficiency, and provide better privacy solutions for data owners. In addition, federated learning uses ML model parameter sharing to prevent data owners from migrating their data, thus significantly reducing privacy problems and communication expenses by traditional centralized ML. In recent years, federated learning has attracted significant attention in research and application fields. With the McMahan et al.'s Federated Averaging (FedAvg) algorithm (Konen et al., 2016), the idea of federated has come into the public view and has been widely used in various fields of artificial intelligence, such as the Internet of Vehicles, blockchain, medical care, banking (Lo et al., 2022; Ma et al., 2022; Rodriguez-Barroso et al., 2020). From the perspective of technological development, the current research on federated learning mainly focuses on data selection, data heterogeneity, communication costs and system robustness. Based on these federated learning researches, many excellent branching algorithms have been formed (from the very beginning to directly share private data for training, then to propose federated learning that only shares ML model parameters, and finally to various optimized practical deep learning combined with federated learning algorithms) (Mothukuri et al., 2021; Tuli et al., 2022; Verbraeken et al., 2020). These researches make federated learning technology gradually mature.





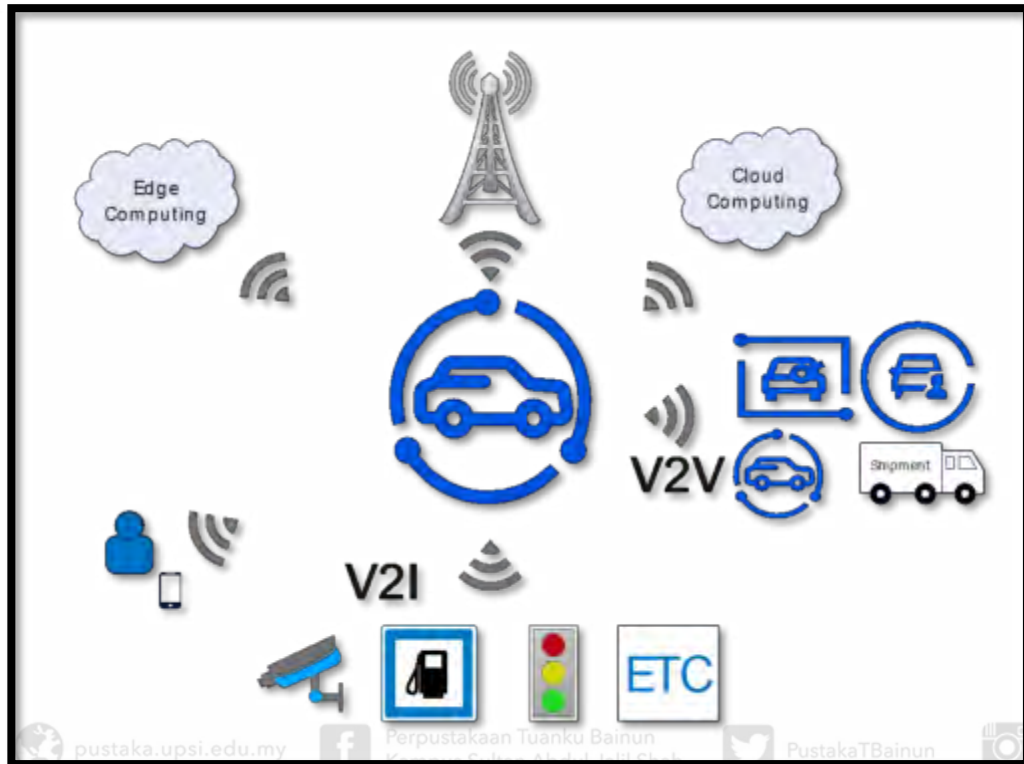
### 1.2.2 Image Processing in the Internet of Vehicles

The Internet of Vehicles (IoV) is a network of vehicles with sensors, software and technology designed to connect and exchange data over the Internet based on agreed standards. IoV is an essential application of the Internet in the automotive industry and an emerging application of intelligent transportation. IoV can be defined in broad and narrow senses. In the broad sense, IoV uses advanced sensing technology, 5G network technology, cloud computing technology, intelligent technology, and automation control technology to fully sense road and traffic information for smooth, safe, and efficient operation. IoV can achieve intelligent collaboration between vehicles and vehicles, vehicles and roads, vehicles and people, and vehicles and environment. In a narrow sense, IoV can be defined as a VANET (vehicular ad-hoc network), an electronic multi-hop network with fast mobility. IoV transforms all moving vehicles into mobile hotspots or routers and forms a large-scale wireless mobile communication network using short-term communication between vehicles. Therefore, as shown in Figure 1-2, there are communications between vehicles (Vehicle to Vehicle, V2V), between vehicles and the roadside infrastructure (Vehicle to Infrastructure, V2I), and between vehicles and data centers (V2C) through this network, while providing intelligent applications and safe traffic services for vehicles.



**Figure 1.2**

*Internet of Vehicles (Source: Author's own photograph)*



Many governments or organizations have formulated IoV's strategic development plans to advance IoV research. Since the 1990s, the global IoV industry has developed for more than 30 years. Relying on a mature computer and big data technology, some developed countries have become leaders in the IoV field, such as North America, Europe, Japan, and South Korea. They have formed relatively complete and stable IoV technical standards, policies, and regulations. For example, the European Union has made specific specifications for the development of autonomous driving in "Putting European transport on track for the future": to achieve autonomous driving in some scenarios by 2022, and it is expected to achieve fully autonomous driving by 2030 (Tuli et al., 2022). In 2020, the U.S. Department of Transportation published "Ensuring American Leadership in Automated Vehicle Technologies: Automated Vehicles 4.0",



which emphasizes supporting user safety and cybersecurity in autonomous driving, promoting efficient market operation, and improving transportation systems (Tuli et al., 2022). In August of the same year, the Korean Ministry of Communications Technology released the "Future Mobile Communication R&D Strategy Leading the 6G Era", planning to list self-driving cars as one of the five main areas of the pilot project (Tuli et al., 2022). The German Federal Government issued a draft law on autonomous driving on February 10, 2021 (Frigo et al., 2019). The law aims to establish a suitable legal framework to start the normal operation of self-driving driverless cars in Germany by supplementing the existing provisions of the Road Traffic Act.

The introduction of 5G networks and the progress of network function virtualization, software-defined network, and other technologies have made the technical obstacles to the development of IoV disappear. Currently, most IoV models are developed on cloud servers, which is convenient for collecting Vehicle classification, analyzing and processing them, and returning the integrated global model to the driving vehicle. However, many studies speculate that a vehicle equipped with sensors and cameras, at least 10 TB, will be collected of data every day soon when everything is connected (Ambroziak et al., 2022; Gan et al., 2023; Li et al., 2022). According to many official, authoritative statistics, as shown in Table 1-1, the global IoV market size surpassed 245.42 billion yuan in 2015, rose to 643.44 billion yuan in 2020, and is expected to surpass 1.5 trillion yuan in 2025. At that time, the percentage of global IoV penetration rose from 30.7% in 2018 to 45% in 2020, and it is estimated that it will cover nearly 60% of vehicles by 2025. In the information explosion age, the data generated in IoV is increasing exponentially.



**Table 1.1***IoV by Authoritative Statistics*

IoV market statistics	It is estimated that more than one billion motor vehicles worldwide use networked devices. (Statista)
	By 2020, the global connected vehicles market will generate about \$54 billion.
	It is estimated that in 2025, this figure will grow to 166 billion dollars. (Statista)
	As of 2018, there were 119 million connected vehicles on the road.
	It is estimated that in 2023, the number of networking vehicles will reach 353 million. (Statista)
	In 2019, the global sales of connected vehicles with embedded telematics functions reached about 28.5 million. (Statista)
	The share of connected vehicles in new vehicle sales will increase from 35% in 2015 to 100% in 2025. (Mordor Intelligence)
Statistics of IoV telematics	From 2021 to 2026, the telematics market of IoV is expected to achieve a compound annual growth rate of 20.7%. (Mordor Intelligence)
	As of 2018, the United States has 32.7% of the world's connected vehicles.
	It is expected that by 2023, Europe will surpass the United States in the number of connected vehicles when they own 31% of the connected vehicles in the world.
	Intelligent information technology can reduce the high-risk behaviors of young drivers by more than 30%. (Mordor Intelligence)
	Risk reduction can reduce the claim cost of this age group by at least 30%. (Mordor Intelligence)
	Sales of advanced driver assistance systems (ADAS) and related technology products have increased from 45 million in 2014 to 54 million in 2018. (Mordor Intelligence)

	Usage-based insurance (UBI) and telematics are changing the expectations of the insurance industry.
	It is estimated that in 2023, the global users' demand for UBI will increase to more than 140 million. (Mordor Intelligence)
	It is estimated that by 2030, UBI and insurance telematics will make a profit through automobile data, and the revenue will exceed 700 billion USD. (Mordor Intelligence)
	In 2019, 86% of fleets used long-distance communication, compared with 48% in 2017 and 82% in 2018. (Teletrac Navman)
	74% of vehicle fleets use telematics to monitor the location of vehicles. (Teletrac Navman)
	66% of fleets use telematics to track service time. (Teletrac Navman)
	61% of vehicle fleets use telematics to control and optimize vehicle speed. (Teletrac Navman)
Long-distance communication for vehicle security	Intelligent telematics can effectively improve the safety of vehicle queues. More than a quarter of freight companies list driver monitoring (32%), speed prevention (26%), and driver fatigue prevention (30%) as the most prominent safety advantages related to telematics. (Teletrac Navman)
	Telematics reduces accidents by 45%. (Driver's Alert)
	Telematics reduces speeding events by up to 75%. (Driver's Alert)
	Telematics can increase the seat belt utilization rate by 90%. (Driver's Alert)
Remote management for vehicle queue	Telematics reduces aggressive driving by 80%. (Driver's Alert)
	Although fuel is considered the largest expenditure of 32% of the fleet, the fuel cost has been reduced by 55% due to telematics software. (Teletrac Navman)
	Tele information processing reduces the travel time of vehicles to their destinations by 68%. The procedure can reduce carbon dioxide emissions by 75%, or about 36 million tons annually. (Mordor Intelligence)



These explosive growth data have important social significance. The pinnacle of IoV image data is various things that people encounter in travel. As the conduit for all information encountered, it contains rich information and has great value for promoting the development of IoV. IoV image data has penetrated people's travel, life, entertainment, and other aspects. It provides crucial support for objectively understanding the appearance of the real world in a data-driven manner and has turned into the focus of study for academics, business, and governments. (Cuzzocrea, 2021; Hao et al., 2019; He et al., 2021). Effectively mining the value of IoV image data, objectively classifying images, and summarizing the characteristics of images are of great significance for promoting social progress, improving people's quality of life, and maintaining sustainable social development. However, IoV image data is different from traditional large-scale data. In addition to its mass, it also has low quality and real-time characteristics. These properties pose severe challenges for mining the potential value of IoV image data (Cao et al., 2021; Diu et al., 2021; Xu et al., 2021). The specific characteristics are as follows.

- i. Large scale. As the most intuitive data feature, the massiveness of IoV image data refers to the vast amount and sparse data value patterns. For example, Intel estimates that each self-driving car will produce almost 4,000GB of data per day, which is comparable to the amount of mobile data produced by just under 3,000 cell phone users. Effectively revealing complete value patterns in massive data requires efficient data analysis methods. However, the current data mining methods based on high-performance computing paradigms include the cloud computing and edge computing ignore the processing efficiency of massive data. Finding a data analysis method for the massive IoV image data improves the





processing effectiveness of algorithm according to the high-performance computing paradigm, efficiently mines the total value of the massive data, and comprehensively performs the massive image data classification.

- ii. Low quality. The low quality of IoV image data refers to the low pattern density caused by the massive data, data containing missing items, and data containing noisy, random, and fuzzy data. To accurately mine the correct value of low-quality data and avoid huge losses caused by unnecessary errors requires data analysis methods to have high robustness. However, most data analysis methods can not reveal the value of low-quality data. Therefore, a data analysis method for low-quality IoV image data can build robust features of low-quality data, reveal value patterns in low-quality data, and serve knowledge discovery and decision-making in a data-noisy environment.

- iii. Real-time. Real-time is another typical characteristic of IoV image data. The real-time nature of IoV image data is reflected in the fact that the data is generated rapidly and has an obvious flow pattern and the property of dynamically changing distribution. Image data value is time-sensitive, showing a decreasing trend over time. For example, regarding intelligent navigation, navigation software such as AutoNavi Maps and Google Maps must collect real-time road traffic and vehicle location information to ensure accurate route planning (Belschner & Pereira, 1995; Furfaro & Nigro, 2003; Liu et al., 2014). Effectively capturing time-sensitive value patterns in real-time data requires data analysis methods with dynamic processing capabilities. However, most of the current data-driven computing methods are static methods. The model's







parameters are based on historical data and cannot cope with incremental image data. Therefore, there is an urgent need to study data analysis methods oriented toward the real-time characteristics of IoV image data. It needs to dynamically adjust the parameters in the ML model based on ensuring accurate discriminative analysis of knowledge patterns in historical data and, simultaneously, realize the mining of potential patterns in real-time incremental data.

It is worth mentioning that massiveness, low quality, and real-time characteristics are not independent. They are an interconnected unity. Our research needs to study all the characteristics individually and finally integrate them to form an excellent unified framework.



### 1.2.3 Why Federated Learning Combined with IoV?

This subsection explains why federated learning and IoV can be integrated based on these three aspects: the challenge of IoV image data, the solution of IoV security and the advantages of federated learning.

Intel CEO Brian Krzanich said: "Data is the new driving force for the future of autonomous driving." The method to implement traditional big data machine learning technology is to upload all data to a single server for global ML training. However, as mentioned above, with the significant increase of image data in IoV, this centralized machine learning method has many shortcomings, such as privacy leakage of car





owners, high communication costs and high transmission delay. These shortcomings gradually form IoV's challenges with isolated data islands, noise data, issues related to diversity and more data. These challenges, caused by resource constraints, seriously hinder the further development of IoV.

#### **i. The Challenge of Isolated Data Islands ---**

Corresponding to characteristic: massive "Isolated data island" refers to the closed and semi-closed phenomena such as asymmetry and redundancy caused by the subject's initiative, the technicality of the object, and the incompleteness of the policy environment and system construction during the formation, analysis, and use of the data set. Many fields have developed so they will build their own corresponding data management systems. These data management systems can standardize business processes, form standardized business models, and automatically accumulate relevant the system database for data to accumulate data resources for relevant individuals. When the value of data is prominent, this accumulated data can enable individuals in the field to develop further. However, the data in these databases of different individuals and different information data systems are often unable to communicate and can only be stored in their respective databases. Therefore the databases cannot be used uniformly. In this way, each individual and each database are separated, like isolated islands overseas, unable to connect and communicate, thus, leading to the occurrence of an isolated data island.



Several reasons for the challenge of isolated data islands in IoV include vehicle owner reluctance, strict policies, and technical difficulties. (1) Vehicle owners' reluctance. IoV data includes some sensitive data about personal privacy, such as travel tracks, navigation information, in-vehicle recordings, and camera images. Most vehicle owners are reluctant to share these sensitive data. (2) Strict policies. Once these sensitive data are stored and processed in a centralized server, the risk of data leakage increases significantly. In 2018, the General Data Protection Regulation was processed to protect user data privacy in Europe (Riazi et al2019). In 2019, China launched the "Guide to Internet Personal Information Security Protection." The European Data Protection Committee (EDPB) adopted the "Guidelines on Processing Personal Data in the Context of Connected Vehicles and Mobility-Related Applications" in 2021. The guidelines introduced over the years explain the privacy protection, data risks, and countermeasures in different scenarios of IoV. These laws or guidance indicate that data owners must be supervised and be obligated to protect data. (3) Technical difficulties. Last but not least, there were about 119 million intelligent vehicles in 2018, and by 2023, this number will nearly triple to 353 million. Meanwhile, as mentioned above, every smart vehicle will generate about 4,000 GB of data every day. Therefore, even though 5G is widely applied on this planet, it is still impossible to realize data sharing and real-time calculation in IoV.

## **ii. The Challenge of Noise Data---Corresponding to Characteristic: Low Quality**

When a variable is measured, noise data is any potential inaccuracy or variation that might impact how accurately and effectively following analytic processes are



performed. Erroneous, misleading, and anomalous data make up most noise data. Abnormal data is a term used to describe discontinuous data that significantly affects the outcomes of data analysis.

Statistical learning methods for labels aim to design theoretically robust loss functions. However, after 2015, deep convolutional networks and recurrent neural networks became mainstream due to their better generalization performance in computer vision (Onan, 2022; Wang et al., 2019; Zhou, 2020). The research on eliminating noise has also gradually shifted from statistical learning to learning representations, from traditional statistical learning models to deep learning models. Despite achieving excellent generalization performance in many tasks, deep learning models still face many challenges in practical applications. The success of deep learning models relies heavily on large-scale and accurately labeled training sets, which is challenging to meet in many real-world tasks. For example, Image Net contains millions of annotated images, and these massive image data require much technical personnel to annotate. The knowledge limitation of each technician, these samples cannot be labeled 100% accurately, thus introducing noise to the dataset. Deep learning models cannot accurately eliminate noise in some fields, which can easily cause irreparable consequences. For example, in the field of medical applications, although medical data sets are usually relatively small, labeling these medical data requires expert knowledge in this field. At the same time, due to the influence of subjective bias among different experts, sample annotations often produce noise. Medical label noise can lead to incorrect predictions of deep learning models, which can negatively impact human health.



Another example is that vehicle-generated data will significantly rise., and the vast local data of vehicles is a significant load for IoV. These image data are easily affected by problems including unstable network environment and transmission speed of driving. In addition, due to the low quality of IoV image data, the local vehicle data used for training has much redundancy for IoV. Noise is pervasive in IoV. However, as the issue of data privacy has been taken seriously, less and fewer data is being collected. In order to succeed excellent accuracy, for training, deep learning models require a lot of high-quality data. (Xiaoqi et al., 2021; Xie et al., 2018). Most technicians lack domain expertise to accurately express the defects of IoV datasets, which will generate much noise. Therefore, there is an urgent need to establish a robust learning algorithm with theoretical guarantees to handle the noise data challenge of IoV.

An excellent IoV deep learning model balances car owners' privacy protection requirements and high-quality data requirements. Under this premise, a practical IoV data-denoising framework is vital.

### **iii. The Challenge of Incremental Data - Corresponding to Characteristic: Real-Time**

In the IoV data storage scenario, there are two ends: the data generation end is the source end, and the data storage end is the destination end. In traditional storage systems, the source and destination are often on the same node or the same network, such as traditional local databases, and local parallel storage. The distinction between the source and destination ends is not apparent in the traditional storage system. the source is local, and the destination is in the remote cloud in the cloud storage scenario.



Therefore, cloud storage faces challenges including high latency, high bandwidth, and high concurrency brought about by this particularity when ensuring data consistency between the source and destination. The cloud storage data center is one of the major sources of resource consumption, which includes not only the power consumed by the data center to maintain operations but also the consumption of hardware resources and network bandwidth during data upload, download, and synchronization. In cloud storage, when data is backed up to the destination, the bandwidth cost caused by data synchronization is a problem that must be considered. For multiple small changes to a large file, if the entire file is sent to the server in complete synchronization, it will inevitably waste a lot of network bandwidth.

However, IoV's cameras are deployed on vehicles, which makes it difficult for traditional storage systems to meet practical requirements. From the perspective of camera data acquisition, data acquisition is not a one-time process but a steady, gradual acquisition process. The increasing data will inevitably improve target recognition ability and expand the application range. A prerequisite, however, is the need to deal successfully with a larger amount of image data of increasing size. The process requires that the recognition algorithm be able to train the recognition model when how many samples there are, and categories is sufficient and to learn additional target categories and training samples.

The real-time incremental data generated in IoV is huge. Different road conditions result in its incremental data such as uphill and downhill in urban and suburban areas, day and night, rain, or snowy and sunny days. It is necessary to update the old ML models to better accommodate driver assistance or safety warning features



in IoV. Key research focuses on fusing incremental data with older ML models and ensuring that all vehicles can effectively participate in global ML model training. Thus, the ability to process incremental data in IoV is to the forefront.

#### **iv. The Challenge of Data Diversity---Corresponding to Characteristic: Massive & Real-Time**

The IoV system needs to learn new knowledge from new image data continuously and requires good preservation of the old categories and old knowledge learned before. Many effective multi-instance methods (Breivold & Rizvanovic, 2018; Jiang & Liu, 2017; Singh et al., 2014) have been created in the recent years to address the computational difficulties brought about by data diversity. Many techniques have been provided to solve the issue of catastrophic forgetting. However, almost no frameworks are proposed to meet the challenge data diversity in IoV.

However, traditional federated learning methods do not take into account the weights of incremental data and rely heavily on the repetition of the training process, even leading to a serious degradation of the accuracy and bias of the global ML model. The common goal of federated and incremental learning is to acquire more reliable prediction results from local ML model parameters of multiple vehicles. Theoretically, the more significant difference of each local ML model parameter, the better the result will be. If the local ML models are highly homogeneous and non-complementary, there is no point in training global ML models, which just replicate them and also increase the computational cost. Our desired framework combines the strengths of different local



ML models in IoV to bridge the existing gap and better address the data diversity challenge.

With the popularization of IoV technology, vehicles collect many location and image data through positioning technology and photography technology. This data contains a lot of personally identifiable information about users. Once leaked, it may lead to severe consequences. Adequate protection of people's privacy can eliminate concerns about personal privacy data leakage, increase people's confidence, and help realize information sharing in smart cities (Ceballos & Larios, 2016). At present, vehicle privacy protection technology can be divided into three groups: based on data about the user's attributes, based on user behavior information and based on user relationship network. Privacy protection technology based on data about the user's attributes mainly refers to realizing user identity protection through anonymity and concealment (He, 2017). Using user-based technology, preserve privacy behavior knowledge can stop attackers from employing the correlation between user attributes and behavior information to construct user data models. Therefore, the technology comprehensively protects user location privacy, mainly including the pseudo-location and generalization methods (Baker-Eveleth et al., 2022). User relationship networks-based privacy protection technology primarily uses the potential relationship between users and multi-hop users to protect user privacy data. The most popular method is to select trusted users. As a result, privacy protection technology aims to reduce the risk of leakage of user privacy information while ensuring the quality of related services.

The above privacy protection technologies are mainly realized by hiding the real information of vehicle users. However, in the traditional centralized machine







learning training process, the training set data are all raw data uploaded by vehicle users. Although the above method can protect the privacy of users, it cannot solve the data security problem of centralized training well. The development of AI technology has made a qualitative leap in information extraction technology. We hope to use deep learning in distributed machines to meet these challenges, that is, distributed learning. Distributed learning has the following characteristics.

- i. In the current research, distributed learning requires the central server to have high control over the client and its data. Clients accept instructions from the central server and is fully under its control. For example, in the distributed computing model of MapReduce, the central server can issue an instruction to allow clients to exchange data with each other.
- ii. Fully distributed learning is peer-to-peer technology. Most of the research direction is data security. There is very little research on IoV.
- iii. Clients under distributed learning are usually located in dedicated computer rooms, interconnected with high-speed broadband, and the network and operating environment are very stable.
- iv. In distributed learning, the data from various clients is usually divided evenly and randomly dispersed. They have independent and identical distribution characteristics, which is very suitable for designing efficient training algorithms.

Most distributed learning assumes a well-resourced environment where clients are stable nodes. However, overly ideal environment assumptions lead to algorithms that are often difficult to implement in IoV in practical applications. IoV is an interactive network composed of data information, such as the vehicle itself and things



outside the vehicle. IoV collects its environment and status information by wireless communication devices including GPS, RFID, sensors, and cameras of the vehicle to the server. After receiving the local data, the server will analyze and process it and finally provide different functional services to the running vehicles. IoV was required to provide a large amount of user privacy data because distributed learning has absolute control over the client's data. As people increasingly focus on the privacy protection of personal data, the amount of data that can be publicly used in IoV is minimal, thus, leading to unsatisfactory effects of some ML models applied in IoV.

Federated learning technology will be an excellent solution. It can improve the precision of the global ML model while saving the data locally in the vehicle. Federated learning is an encrypted distributed machine learning technology. Introducing it into the IoV architecture makes it possible to build a safe and efficient data-sharing framework with the help of the local training of all parties involved in federated learning without the need to migrate local data. Federated learning can combine vehicles and service providers in IoV to jointly train a ML model for a specific task. During the training of the ML model, local ML parameters and global ML parameters are exchanged between the vehicle and the server. In addition, the vehicle does not need to migrate local data during the training process, effectively avoiding leakage during data transmission. Federated learning has the following advantages over traditional machine learning methods.

- i. Address the issue of isolating data islands. Individual dataset owners are reluctant to share their data. In federated learning, data from each client is stored

locally and won't be leaked to the outside, which can effectively protect user privacy.

- ii. Efficient use of bandwidth resources. The client only sends the local ML model parameters to the server, not the dataset. The approach reduces the cost of data communication and reduces the burden on the network.
- iii. Ensure the integrity of the global ML model. Federated learning is to send the complete global ML model to the client without splitting the model, which can guarantee that there will be no negative transfer.
- iv. Ensure client equality. In federated learning, each local client is a fair, cooperative relationship, and there is no situation of unequal status, and each participant will get a performance improvement.

situations, including finance, medical care, and mobile phones (Luo et al., 2021; Wainakh et al., 2020; Yang et al., 2019). With the rapid development of Mobile Edge Computing (MEC) technology, The Roadside Unit (RSU) acts as the central server and the vehicle as a local client. It is no longer a dream to deploy a framework for federated learning in the IoV scenario.

Although the federated learning in IoV can obtain the above benefits, it also faces many challenges. On the one hand, the federated learning process that the vehicle and the central server require certain computing power, storage capacity, energy, and wireless resources. The central server connects wired devices and has sufficient resources. However, vehicles are terminal devices in MEC, and their computing resources and energy are limited. The balance of computing and wireless transmission



resource allocation can directly affect the federated learning training process. On the other hand, the channel bandwidth between the vehicle and the central server is usually limited, and wireless transmission will have error rate and delay problems. Therefore, how is an urgent problem to be solved is to reasonably allocate limited computing and wireless resources in IoV to achieve better global ML model training.

Due to the real-time mobility of vehicles, differences in computing power, and differences in communication infrastructure in different regions, directly combining federated learning technology with IoV may bring excessive computing burdens to some vehicle nodes. In addition, dishonest service providers privately use vehicle data to mine as much information as possible, causing the privacy of vehicle users to leak. Therefore, it is necessary to carry out a targeted transformation of federated learning technology to provide more efficient and secure solutions.

### 1.3 Problem Statement

In the contemporary domain of the IoV, the importance of frameworks dedicated to vehicle privacy protection is increasingly underscored due to their direct connection to the safeguarding and security of user privacy. With the rapid advancement of technology, vehicles are generating an ever-increasing volume of data. The challenge of protecting this data from misuse or breach has become a pressing issue to address. Despite the emergence of numerous privacy protection frameworks, the majority still require users to submit their original private data, thereby elevating the risk of privacy leaks and causing significant concern and resistance among many users.





The reliance on original data not only compromises the effectiveness of privacy protection but also leads to the formation of isolated data silos within the IoT. These silos restrict the flow and sharing of data, impeding the further development of IoV technology. Hence, striking a balance between protecting user privacy and enabling efficient data flow and sharing presents a significant challenge in the current IoV landscape.

Regrettably, despite the proliferation of privacy protection frameworks in the vehicle privacy protection sector, most still necessitate the submission of original private data by users. This requirement poses a substantial privacy leak risk, eliciting strong resistance from them. The reluctance of users to cooperate has gradually led to the creation of isolated data silos within the IoT, significantly limiting the effective utilization and circulation of data. Moreover, in the field of IoV image data processing, finding a vast amount of meaningful picture data is inherently challenging. Additionally, improving the performance of global Machine Learning (ML) models not only requires ample data support but also an efficient training methodology. However, existing methods often rely on datasets, which considerably limits the models' generalization ability and practicality. Evidently, finding a straightforward and effective global ML model training method without depending on original private data emerges as one of the urgent issues to be addressed in the vehicle privacy protection field.

Furthermore, in the implementation and operation of IoV image data, the issue of noise data gradually surfaces, becoming a critical factor constraining further technological progress. Deep learning, especially supervised learning, has achieved significant success in applications like image classification. However, it's essential to





acknowledge that supervised deep learning models' performance heavily relies on training sets with minimal noise data. In practice, obtaining large-scale training sets with little noise is challenging due to unavoidable subjective biases and measurement errors in the manual annotation process. This noise not only affects the training effectiveness of models but also severely interferes with the convergence direction of ML models, leading to a notable decline in model performance.

Thus, addressing how to manage noise data in IoV images and effectively improve model accuracy under such conditions becomes another pressing issue for this study. This not only relates to the further development and application of IoV technology but also holds significant practical significance for enhancing deep learning's performance in image processing domains.



Beyond these concerns, the rapid advancement of IoT technology daily generates a vast volume of image data, posing substantial storage pressures on IoT systems. The explosive growth of image data makes storing real-time incremental data particularly challenging. More complexly, these additional data often highly resemble existing old data, with a considerable overlap in the information conveyed. Therefore, finding ways to reduce the storage burden of incremental data while effectively integrating new and old data to construct an efficient and stable IoT framework also emerges as a critical issue to address.

Lastly, research must filter more reliable image data classification results from the numerous participating vehicle clients' local ML model parameters. It's evident that different types of incremental data significantly influence the local ML model





parameters. If the parameters of various local ML models exhibit significant differences, their complementarity increases, thereby aiding in enhancing the training effect of the global ML model. Conversely, when local ML models show high homogeneity and lack sufficient complementarity, the training of the global ML model becomes meaningless, merely amounting to repetition. This not only fails to improve model performance but also incurs unnecessary computational costs.

Therefore, selecting and utilizing sufficiently diverse local ML model parameters to optimize the global ML model's training effect is crucial. Solving this issue not only pertains to the efficiency of IoT image data processing but also significantly impacts enhancing the performance and stability of the entire IoT system.



#### 1.4 Research Objectives

The overarching goal of this study is to address the challenges identified in IoV image data processing through federated learning. The specific research objectives (ROs) are:

- RO1: To develop an IoV noise data processing framework on top of the federated learning model.
- RO2: To establish an incremental data processing framework based on federated learning, to mitigate data homogeneity.
- RO3: To integrate the noise data processing framework with the incremental data framework to efficiently handle both noise and incremental data in IoV.



RO4: To explore the interplay between noise data processing and incremental data processing within the overall framework, investigating whether the sequence of data processing impacts the proportion of incremental data types.

## 1.5 Research Questions

Aligned with the stated objectives, this study poses the following research questions (RQs) to guide the investigation.

RQ1: Is the IoV noise data processing framework based on the federated learning model effectively constructed?

RQ1.1: Do the outlier detection, Kalman filtering, and exponential smoothing algorithms within the framework effectively filter noise data in IoV?

RQ2: Is the incremental data framework based on federated learning effectively constructed for the IoV context?

RQ2.1: Does the integration of federated and incremental learning address the performance balance issue between incremental and existing data within the global ML model?

RQ3: Does combining the noise data processing framework with the incremental data processing framework achieve effective handling of both noise and incremental data?

RQ4: Is the sequence of processing data types related to the proportion of incremental data types?





## 1.6 Research Significance

In the context of the IoV, this study introduces a federated learning-based framework designed to efficiently process both noise and incremental data. The construction of this framework not only offers a novel approach to data management within the IoV domain but also significantly supports the practical application and expansion of federated learning models. As a vast platform for data generation and exchange, the IoV ecosystem produces an enormous amount of data daily, including a substantial proportion of noise and incremental data. Noise data may arise from sensor inaccuracies, transmission errors, among other sources, while incremental data is generated by factors such as the addition of new vehicles and the opening of new roads. Addressing the effective management of these data types to enhance the performance and accuracy of IoV systems is an urgent challenge. The proposed framework employs a combination of outlier detection, Kalman filtering, and exponential smoothing algorithms for effective noise data management. Outlier detection identifies and eliminates aberrant data points, Kalman filtering smooths noisy data through prediction and update steps, and exponential smoothing further reduces random errors in the data. This amalgamation of algorithms significantly enhances data quality and reliability. For incremental data, this study adopts incremental learning techniques, which allow for the assimilation of new data without the need for retraining the entire model. This method enables the framework to rapidly adapt to the inclusion of new data, maintaining the model's timeliness and accuracy without compromising system performance.





Overall, the federated learning-based framework proposed in this research, through the integrated application of outlier detection, Kalman filtering, exponential smoothing algorithms, and incremental learning techniques, offers an effective solution for data processing in the IoV context. The implementation of this framework is anticipated to improve the performance and accuracy of IoV systems, further advancing IoV technology.

Firstly, the federated learning framework ensures data privacy while enhancing model training efficiency and accuracy, catering to the dynamic data and privacy protection requirements within the IoV environment.

Secondly, by integrating outlier detection, Kalman filtering, and exponential smoothing algorithms, the framework effectively processes noise data, enhancing the accuracy and reliability of IoV data under this framework.

Thirdly, employing incremental learning technology allows the framework to adapt in real-time to the incorporation of new data, maintaining timeliness and accuracy to meet the dynamic changes in IoV data. This capacity is crucial for advancing real-time traffic management, route prediction, and other applications within the IoV, highlighting the framework's importance in improving IoV system performance and reliability and propelling the development of intelligent transportation and autonomous driving.

Fourthly, the combination of noise data processing and incremental data handling in the IoV context provides a comprehensive and efficient data management





solution for a large-scale, highly dynamic data environment. This approach not only addresses data quality and dynamic changes but also supports intelligent transportation, autonomous driving, and the construction of smart cities by enhancing system performance and reliability.

Fifthly, comparing the proposed noise data processing framework with baseline models validates its effectiveness and superiority in handling noise data within the IoV environment. This comparison offers new solutions for data processing in the IoV field and serves as a valuable reference for similar data challenges in other domains.

Sixthly, the incremental data processing framework proposed for the IoV environment aims to prevent homogenization of local ML models, achieve data diversity, and demonstrate superior incremental data handling capabilities. This design addresses the challenges of data dynamics and distributed processing in the IoV, ensuring personalized local models can optimize performance using new data through incremental learning.

Seventhly, the study suggests that the sequence of processing noise and incremental data may indeed relate to the proportion of incremental data types. This highlights the need for flexibility and adaptability in data processing strategies, essential for dynamic environments like the IoV.

Eighthly, the framework employs advanced techniques and methodologies, including outlier detection, Kalman filtering, exponential smoothing algorithms, and incremental learning, to ensure dual enhancement of data quality and model





performance. This comprehensive approach demonstrates exceptional performance in addressing noise and incremental data challenges within the IoV, supporting the evolution of IoV technology and the improvement of intelligent transportation systems.

This research framework presents a solution for handling incremental and noise data challenges in the IoV, not only advancing IoV technology and enhancing intelligent transportation system performance but also supporting the sustainable development of smart cities. Its universal applicability also offers beneficial insights for addressing similar data challenges in other fields.

## 1.7 Research Scope



This research is committed to innovating and improving the processing of noise and incremental data in the Internet of Vehicles (IoV) domain through the comprehensive application of advanced technologies such as machine learning algorithms, federated learning, and incremental learning. The scope of this study is broad yet detailed, encompassing a range from foundational algorithmic research to exploration of practical applications.

Federated Learning in IoV: The focus within federated learning, particularly in the context of IoV, centers on enabling intelligent collaboration and efficient learning among vehicles while safeguarding data privacy and security. This includes designing federated learning algorithms suitable for IoV environments, optimizing system architectures for efficient vehicle-to-vehicle communication and model updates,





exploring data privacy and security preservation during federated learning processes, and investigating practical application scenarios and potential challenges within IoV.

**Noise Data in IoV:** Addressing the prevalent issue of noise data within IoV, this study concentrates on devising effective noise data filtering methods using machine learning, K-Means algorithms, and Kalman filter techniques. This entails researching accurate identification and classification of noise data in IoV, as well as algorithmic enhancements to improve the efficiency and accuracy of data filtering, thereby elevating data quality.

**Incremental Learning in IoV:** Given the challenge of handling massive data volumes in IoV, incremental learning technology is adopted as a core solution.

Traditional machine learning models often struggle with the vast amounts of data due to storage constraints and fail to efficiently process new incremental data. This research investigates the design and implementation of incremental learning algorithms tailored for the IoV environment to achieve continuous, efficient processing of massive data sets. Incremental learning enables ongoing learning and adaptation to new data, maintaining model accuracy and robustness and addressing the challenges of large-scale, rapidly changing data in IoV, thus enhancing data processing efficiency and accuracy.

Overall, by integrating technologies such as Internet of Things (IoT) algorithms, joint learning, machine learning, and incremental learning, this study aims to provide innovative solutions for various challenges in the IoT domain. Our research, with its





extensive and in-depth scope, seeks to advance IoT technology development and application, contributing to the progress towards an intelligent society.

## 1.8 Thesis Structure

The focus of this research is to propose solutions to various challenges in IoV. Some public IoV datasets are used for experiments to verify our proposed framework. This thesis is divided into five chapters, with the following specific contents.

The first chapter introduces the thesis's research background and significance, content, scope and organization. The second chapter conducts a literature review from multiple perspectives, including IoV, federated learning and incremental learning. The improvements and innovations of our framework are described in the third chapter. The experimental results are explained in detail in the fourth chapter. Lastly, the fifth chapter is the summary and outlook.

## 1.9 Chapter Summary

This chapter has laid the foundational principles of this research, and offered a comprehensive overview as well as delineated the basic principles that guide our investigation. We have articulated the research objectives, questions, and goals that form the cornerstone of this thesis, structuring the inquiry and hypotheses that drive our study forward. These elements collectively underscore the focus of our work,





establishing a clear path toward addressing the challenges identified within the context of the IoV.

The justification for this research has been substantiated based on its relevance and significance in the current landscape. By situating our study within the broader context of IoV, we have demonstrated the pressing need for innovative solutions to noise and incremental data processing challenges that this domain faces. A brief discussion on the conceptual framework was presented, setting the stage for a deeper exploration of the key technologies employed in our research. These include federated learning, machine learning algorithms, incremental learning, and noise data filtration techniques, which are pivotal for advancing IoV data processing capabilities.



Moreover, this study's formation process was elucidated, emphasizing the iterative approach taken from conceptualization to execution. This process was underpinned by a thorough review of contemporary literature relevant to our research domain. Such a review not only enriched our understanding of the state-of-the-art in IoV data processing but also highlighted gaps and opportunities for innovation that our study aims to address. By critically engaging with existing scholarly work, we have positioned our research to contribute meaningful advancements to the field of IoV.

In concluding this chapter, it is evident that our research stands on a robust foundation of scholarly inquiry and technological exploration. The objectives, questions, and goals outlined here reflect a strategic approach to tackling the complexities of data processing within the IoV. As we progress, the methodologies and analyses detailed in the subsequent chapters are geared towards validating our





hypotheses and achieving the research outcomes we have set forth. Ultimately, this study aims not only to advance the academic discourse on IoV but also to offer practical solutions that can be implemented to improve the efficiency, reliability, and security of IoV systems, paving the way for a more connected and intelligent transportation ecosystem.

