symmetry

MDPI

Article

A Comprehensive Literature Review on Artificial Dataset Generation for Repositioning Challenges in Shared Electric Automated and Connected Mobility †

Antoine Kazadi Kayisu [1,2], Witesyavwirwa Vianney Kambale [3,4], Taha Benarbia [3,5], Pitshou Ntambu Bokoro [2] and Kyandoghere Kyamakya [1,3,*]

[1] Faculté Polytechnique, Université de Kinshasa (UNIKIN), Kinshasa 15373, Democratic Republic of the Congo; antoine.kayisu@unikin.ac.cd
[2] Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg 2006, South Africa; pitshoub@uj.ac.za
[3] Institute for Smart Systems Technologies, Universitaet Klagenfurt, 9020 Klagenfurt, Austria; witesyavwirwa.kambale@aau.at (W.V.K.)
[4] Faculty of Information and Communication Technology, Tshwane University of Technology, Private Bag x680, Pretoria 0001, South Africa
[5] Department of Industrial engineering, Institute of Maintenance and Industrial Security, University of Oran 2, Oran 31000, Algeria
* Correspondence: kyandoghere.kyamakya@aau.at
† This paper is an extended version of our paper published in the Circuits, Systems, Communications and Computers (CSCC-2023) Conference, Rhodes Island (Rodos Island), Greece.

Abstract: In the near future, the incorporation of shared electric automated and connected mobility (SEACM) technologies will significantly transform the landscape of transportation into a sustainable and efficient mobility ecosystem. However, these technological advances raise complex scientific challenges. Problems related to safety, energy efficiency, and route optimization in dynamic urban environments are major issues to be resolved. In addition, the unavailability of realistic and various data of such systems makes their deployment, design, and performance evaluation very challenging. As a result, to avoid the constraints of real data collection, using generated artificial datasets is crucial for simulation to test and validate algorithms and models under various scenarios. These artificial datasets are used for the training of ML (Machine Learning) models, allowing researchers and operators to evaluate performance and predict system behavior under various conditions. To generate artificial datasets, numerous elements such as user behavior, vehicle dynamics, charging infrastructure, and environmental conditions must be considered. In all these elements, symmetry is a core concern; in some cases, asymmetry is more realistic; however, in others, reaching/maintaining as much symmetry as possible is a core requirement. This review paper provides a comprehensive literature survey of the most relevant techniques generating synthetic datasets in the literature, with a particular focus on the shared electric automated and connected mobility context. Furthermore, this paper also investigates central issues of these complex and dynamic systems regarding how artificial datasets could be used in the training of ML models to address the repositioning problem. Hereby, symmetry is undoubtedly a crucial consideration for ML models. In the case of datasets, it is imperative that they accurately emulate the symmetry or asymmetry observed in real-world scenarios to be effectively represented by the generated datasets. Then, this paper investigates the current challenges and limitations of synthetic datasets, such as the reliability of simulations to the real world, and the validation of generative models. Additionally, it explores how ML-based algorithms can be used to optimize vehicle routing, charging infrastructure usage, demand forecasting, and other important operational elements. In conclusion, this paper outlines a series of interesting new research avenues concerning the generation of artificial data for SEACM systems.

Citation: Kayisu, A.K.; Kambale, W.V.; Benarbia, T.; Bokoro, P.N.; Kyamakya, K. A Comprehensive Literature Review on Artificial Dataset Generation for Repositioning Challenges in Shared Electric Automated and Connected Mobility. *Symmetry* **2024**, *16*, 128. https://doi.org/10.3390/sym16010128

Academic Editor: Nikos Mastorakis

Received: 18 November 2023
Revised: 15 January 2024
Accepted: 18 January 2024
Published: 21 January 2024

Keywords: artificial data; shared electric vehicles; autonomous repositioning; automated and connected mobility

1. Introduction

1.1. Background and Motivation

In recent years, there has been a consistent and notable rising trend in the adoption of automated, connected, and electric driving. These fundamental components of modern mobility have experienced substantial growth, with the expansion further propelled by the surge of IoT and AI within the automotive sector. "Autonomous driving" refers to a vehicle's ability to operate and navigate independently, without requiring direct human intervention. Conversely, "connectivity" involves the integration of communication technologies to enable data exchange between vehicles, infrastructure, and other devices for enhanced safety, efficiency, and accessibility. "Electrification", in the context of electric vehicles, signifies the process of replacing conventional internal combustion engines with electric systems powered with batteries or alternative electric sources. According to experts [1],

- Autonomous driving technology (minimum level 2) will be integrated in 70% of new vehicles by 2030 around the world.
- A total of 96% of new vehicles worldwide would be equipped with integrated connectivity.
- A total of 24% of new cars will be electric by 2030.

The core technologies detailed above (see Figure 1) will change the current landscape of mobility to a fully automated and interconnected system. Furthermore, shared automated and connected electric systems (also called robotaxis) will have the ability to optimize resource utilization, reduce congestion, and enhance sustainability. Nevertheless, the successful implementation and assessment of these complex systems require access to various datasets.

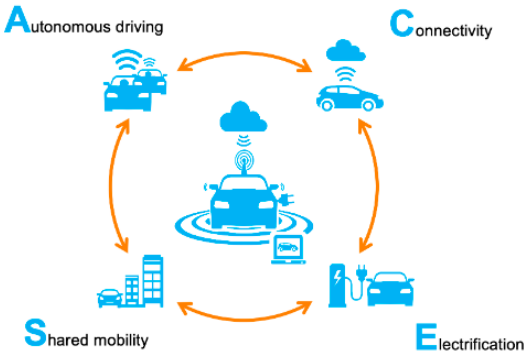

Figure 1. Core technologies of shared electric automated and connected mobility systems.

These systems are designed to tackle challenges in conventional transportation arising from urbanization and population growth, resulting in congestion, pollution, and inefficiencies. SEACM systems, employing innovative approaches such as autonomous repositioning, optimize the use of vehicles and infrastructure. However, several challenges associated with shared transportation have emerged for SEACM, potentially impeding the advancement of these systems. These challenges encompass safety concerns, traffic obstacles, accessibility issues, and particularly, social exclusion frequently reported by users of car-sharing services. The literature on shared autonomous electric mobility systems has addressed various operational research issues and topics, including vehicle assignment, vehicle repositioning, fleet dimensioning, energy and battery capacity, station location, and cybersecurity attacks [2]. To design and assess these systems effectively, comprehensive

datasets are indispensable. These datasets should encompass vehicle trajectories, user behaviors, charging patterns, and infrastructure availability. They play a crucial role in facilitating the development and validation of algorithms, optimization models, and simulation frameworks, ultimately enhancing system performance and user experience.

The analysis of these datasets allows policy makers, urban planners, and system operators to understand how users interact with the system, how they travel, and how the system's parameters affect them. This information can then be used to inform system design and fleet management, as well as infrastructure planning.

In recent years, several studies have addressed the use of datasets in shared electric automated and connected mobility (SEACM) systems. For instance, Li et al. [3] analyzed how vehicle trajectory data could be used to improve vehicle distribution strategies, and Zhang et al. [4] looked at patterns and preferences of user behaviors. These studies show how important datasets are in research and innovation. Having access to diverse datasets is essential for their successful implementation and evaluation. Datasets allow for the development of cutting-edge algorithms and insights into user behavior. By continuing to collect and analyze data, the field will move forward and provide more efficient and sustainable solutions.

A dataset generated for modeling and simulating SEACM systems with autonomous repositioning typically encompasses diverse data types and categories, as illustrated in Figure 2 [5,6]. Here are some specific examples:

1. Geospatial data: These comprise geographical information, including latitude, longitude, and elevation, representing the locations of vehicles, charging stations, and other infrastructure components within the mobility system [5].
2. Vehicle data: These encompass vehicle details like unique identifiers, types (e.g., electric, autonomous), current and past locations, battery status, and operational parameters [5].
3. Vehicle availability: These data serve to formulate effective redistribution strategies for maximizing demand coverage. They include information such as electric autonomous vehicle (EAV) fleet size, vehicle availability, EAV locations, and charging status [5].
4. Demand data: Capturing transportation service demand, these data incorporate origins, destinations, duration, distance, and travel preferences, along with reservation specifics, pick-up and drop-off points, distance, and travel time. They are essential for machine learning model training and optimizing SEACM's vehicle redistribution efficiency [5].
5. Infrastructure data: This category involves the location, capacity, availability, and connectivity of charging stations in SEACM systems, along with information on road traffic and network configuration.
6. Environmental data: Essential for evaluating the SEACM system's performance, these data include weather conditions (temperature, rain, wind, and snow), and seasonal variations.
7. Communication and connectivity data: Encompassing details about communication networks, these include links between vehicles (V2V), communication between vehicles and infrastructure (V2I), as well as vehicle communication with pedestrian centers.
8. Sensor data: This category incorporates information from various sensors in vehicles or infrastructure, such as lidar, battery charge levels, temperature sensors, and camera data for perception.
9. Simulation parameters: These parameters control the simulation process, including simulation time, duration, traffic density, and repositioning strategies.

Furthermore, in the intricate process of generating datasets tailored for the modeling and simulation of SEACM systems with autonomous repositioning, meticulous attention must be given to symmetry considerations. This entails not merely meeting the requisite balance in representation but also delving into the nuanced aspects of asymmetry. Achieving symmetry involves ensuring an equitable and well-distributed portrayal of various features and classes within the dataset. However, the complexity of real-world

scenarios often demands a more nuanced approach. Delving into asymmetry, the generated datasets should not only mirror the overall equilibrium but also encapsulate the disparities and irregularities inherent in the dynamic landscape of autonomous systems. This dual perspective not only upholds the essential balance in the dataset but also captures the richness of asymmetrical scenarios. By doing so, the dataset becomes a comprehensive and versatile tool for modeling SEACM systems, encompassing both the structured symmetry required for fundamental representation and the subtle asymmetry essential for addressing real-world intricacies.

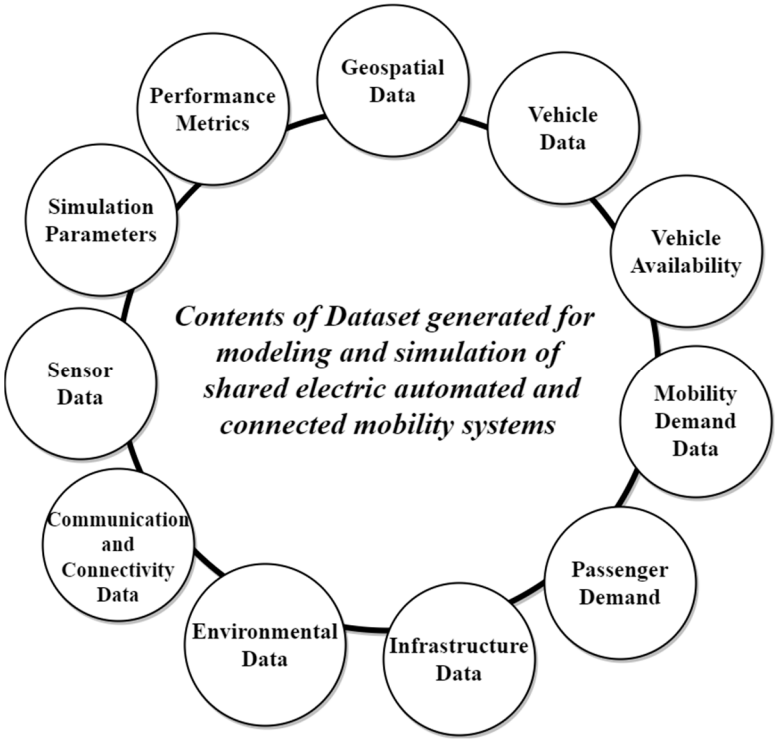

Figure 2. Contents of a dataset generated for the modeling and simulation of SEADM systems.

1.2. Research Objectives

This review paper seeks to investigate the capability of synthetic data in the modeling and simulation of SEACM systems. It will also discuss various methodologies for generating synthetic data, highlighting their significance in performance evaluation and their impact on deployment strategies. To achieve these objectives, the following steps will be undertaken:

(1) Literature review: A comprehensive examination of the recent literature on artificial dataset generation for SEACM systems will be conducted. This review aims to identify both the strengths and gaps in proposed approaches, outlining diverse techniques, methods, and challenges associated with the generation of artificial datasets.
(2) Analysis of synthetic dataset generation techniques: This section will scrutinize different techniques employed for generating synthetic datasets, including data augmentation, synthetic data generation, and the amalgamation of real and synthetic data. The evaluation will delve into the advantages, limitations, and relevance of each technique to SEACM systems.
(3) Relevance of synthetic datasets in performance evaluation: The effectiveness of synthetic datasets in assessing the performance of SEACM systems will be examined. Emphasis will be placed on highlighting the importance of employing realistic and diverse artificial datasets and analyzing how various dataset characteristics impact system performance metrics.

(4) Impact of synthetic datasets on deployment strategies: This segment will investigate the influence of synthetic datasets on the deployment strategy of shared electric automated and connected mobility systems. Furthermore, it will explore how the utilization of artificial datasets can contribute to decision-making processes, such as determining fleet size, planning infrastructure, and optimizing the system, through the examination of case studies and practical examples.

To strengthen our findings and offer a thorough perspective, we will reference pertinent sources, including [7–12].

In addition, one of the key challenges in the dynamics of shared electric, automated, and networked transportation is optimizing vehicle repositioning. To address the relocation issue, this research conducts an extensive literature review and investigates the complexity of artificial dataset generation. The main objective is to provide a comprehensive review of the techniques used in generating artificial dataset to enhance automated shared mobility system optimization.

By synthesizing insights from existing studies, this review aims to contribute to the literature on SEACM, highlighting novel strategies and identifying promising directions for further investigation. Furthermore, through this investigation on the state of the art, we also aim to give readers a concise picture of the repositioning problems in shared electric automated and connected mobility, as well as the critical role that artificial dataset generation plays in solving these problems.

Once our research purposes are achieved, we intend to add to the existing literature on generating artificial datasets for SEACM systems. The information gathered in this survey paper will be useful for researchers, practitioners, and policymakers engaged in the design, assessment, and implementation of these future transportation systems.

Many research questions could be raised in this topic that may directly affect the quality of synthetic data. How to ensure that the generated data are of high quality and relevant to the original data and the modeling task? How to avoid over-synthesizing the data, as this can cause overfitting or the loss of information? How to create enough diversity and balance in the data? Do data generation techniques raise any ethical and privacy issues, such as consent, ownership, and transparency? What is the level of risk that synthetic techniques can modify the original data, this being particularly dangerous for personal or confidential data, such as health, biometric, or financial data? How to ensure that the generated data are reliable in terms of cyber security, as that affects the security effectiveness of connected mobility? How to ensure that the synthetic data contain enough diversity and balance? How to interpret and evaluate the data and the model outcomes to assess the quality, validity, and reliability of the data and the model? Poor quality or irrelevant data can introduce noise, bias, or inconsistency to the model, leading to inaccurate or misleading predictions.

1.3. Structure of the Paper

The paper is organized as follows: Section 2 gives an overview of techniques for generating artificial datasets; Section 3 delves into the importance of artificial datasets in shared mobility systems; Section 4 explores machine learning model training for performance evaluation; Section 5 addresses deployment considerations; Section 6 showcases case studies and experiments; and Section 7 provides a discussion on future directions. Finally, Section 8 concludes the paper.

2. Artificial Dataset Generation Techniques

Through this literature review, a historical overview of data generation techniques leading up to AI-driven methods will be presented. In the late 20th century, Monte Carlo simulations introduced sophisticated methods for modeling complex phenomena through random sampling, while data augmentation emerged to enhance limited datasets with diverse variations. The 21st century marked a central shift with the rise of AI. Generative

adversarial networks (GANs), founded by Ian Goodfellow in 2014, transformed data generation by employing two neural networks to produce remarkably accurate synthetic data.

2.1. Monte Carlo Simulations

The Monte Carlo technique employs randomness to solve complex problems or estimate values. It involves generating numerous random samples or simulations to approximate results, making it useful for tasks involving uncertainty, numerical integration, and optimization. Several works on transportation and shared mobility employed this technique for extending data [13]. By using Monte Carlo, Axhausen and Martin Frick [14] have generated synthetic populations of Swiss PUS data (1970, 1980, and 1990) and Swiss micro census. The data sources were fused together using the Monte Carlo technique to estimate the joint distribution of demographic characteristics such as age, sex, car ownership, driver's license ownership, vehicle miles, car availability, etc.

2.2. Bayesian Network

A Bayesian network is a graphical model representing probabilistic relationships among variables. It enables the generation of synthetic data for public release while preserving confidentiality, allowing external analysts to study attribute associations. Given the high cost to collect real data about the entire population in large cities, population synthesis is an unavoidable way for analysis. In this respect, Anugrah Ilahi and Axhausen [15] have proposed a Bayesian network model for population synthesis to construct the population for Greater Jakarta, Indonesia, which consists of 30 million inhabitants.

2.3. Synthetic Data Generation

Synthetic data generation involves creating artificial datasets using generative models to mimic real data's statistical characteristics and patterns. GANs, VAEs, and DBNs are powerful techniques in this regard, closely resembling real-world distributions.

Generative adversarial networks (GANs) use a game-theoretic structure, with a discriminator differentiating real from fake samples and a generator producing synthetic instances. GANs have shown impressive results in various applications like picture synthesis, text generation, and generating data for SEACM systems [16].

Variational autoencoders (VAEs) employ hidden variable models to generate synthetic data by reconstructing data from the hidden space. VAEs are suitable for SEACM systems, capturing their complex nature and producing data with controllable characteristics [17].

Deep belief networks (DBNs) use multiple layers of hidden variables to generate hierarchical representations of data, playing a significant role in generating synthetic data for complex systems, such as recognizing the structure of AMS circuits in shared automated and connected electric transportation systems [18].

The ability to generate datasets of varying sizes allows researchers to conduct in-depth experiments, validate models, and assess system performance. Using generative models, researchers can develop effective deployment methods, better understand SEACM system behavior, and evaluate the impact of various elements.

2.4. Data Augmentation

Data augmentation has been a fundamental aspect of machine learning for an extended period, as evidenced by its application in pioneering ML models such as AlexNet [10]. This technique involves making diverse changes and adjustments to real datasets to expand and diversify them, playing a crucial role in enhancing the performance of machine learning models [19]. In situations where the dataset volume is limited, extension becomes a significant approach to generate additional samples that replicate the original characteristics of the real dataset. This extension technique holds potential for overcoming the challenge posed by the constraints of real datasets. Consequently, machine learning models can be trained on a large and varied set of data samples, leading to the development of reliable and generalized models capable of performing well under various scenarios. Moreover,

data augmentation has the capability to preserve the original form and characteristics of the data, facilitating the effective training of machine learning models for accurate predictions with new data. Several works in the literature provide comprehensive information on the issue of data augmentation [19–21]. These references explore various data augmentation techniques applicable to the training of deep learning models.

2.5. Transfer Learning

Transfer learning is a machine learning technique that allows a model to be learned for one task to be used for another activity that is similar to it. Using this method can significantly reduce the time and effort required for updating and retraining models compared to starting from scratch [22]. With transfer learning, pre-trained models can be used on real datasets in general and refined on more manageable synthetic datasets. With this method, machine learning models can make use of the unique patterns in the synthetic data as well as the rich features of the trained models [23]. Transfer learning has potential applications in a wide range of fields, including natural language processing and computer vision.

By employing transfer learning to create synthetic datasets, we can benefit from the representations acquired by models trained on copious quantities of real data. It will be feasible to mimic SEACM systems with autonomous repositioning more precisely and successfully by utilizing this experience. The general traits and patterns required will also be captured by the pre-trained models. The model can then be refined on the synthetic dataset to understand the unique traits and subtleties of the target domain.

Transfer learning has been effectively used in computer vision for applications like semantic segmentation, object detection, and image categorization. Convolutional neural network (CNN)-based pre-trained models, such VGGNet and ResNet, have demonstrated exceptional feature transferability across many datasets and domains [24]. Similar to this, transfer learning has proven useful in natural language processing for applications including machine translation, named entity recognition, and sentiment analysis. Pre-training language models on large-scale text corpora, including BERT and GPT, has greatly enhanced performance in downstream tasks [24].

Enhancing the modeling and simulation of SEACM systems can be made possible by utilizing transfer learning in the context of artificial dataset generation [25].

2.6. Other Techniques

There are numerous alternative approaches that can be effectively used to generate artificial datasets in addition to the previously described techniques of transfer learning, data augmentation, and synthetic data generation. In addition to increasing the overall quality and diversity of the generated data, these strategies extend the range of possible outcomes. Two approaches—data sampling strategies and ensemble learning—will be discussed in this section.

Data sampling methods: By constructing representative subsets from bigger datasets, data sampling techniques guarantee that the generated fake datasets have all the important components of the original data. Stratified sampling is a widely used technique to guarantee the proportionate representation of various classes or categories within the dataset [26]. This is an extremely helpful tactic for uneven datasets when some classes are underrepresented. Including stratified samples in the artificial dataset improves its ability to replicate the distribution and class proportions of the real data, which is useful for subsequent analyses and modeling.

Random sampling is another frequently used technique for building artificial datasets. To present a varied image of the underlying data distribution, samples are chosen at random from the original dataset. Random sampling is frequently used in conjunction with other techniques to achieve unpredictability and provide a comprehensive representation of the data.

Ensemble learning: Bootstrap aggregation (also known as bagging) and other ensemble learning techniques facilitate the creation of artificial datasets. In order to create several

subsets of the original dataset, bagging entails sampling with replacement [27]. Subsequently, every subgroup is trained separately, and the combined forecasts of all models yield the ultimate prognosis. By training models on various subsets of data and aggregating their output, ensemble learning can be used to construct artificial datasets. As a result, the artificial dataset is more diverse, generalization is encouraged, and overfitting is decreased.

Additionally, data sampling strategies and ensemble learning approaches can be used by researchers to enhance the body of literature on the production of artificial datasets. By ensuring that the datasets produced are richer, more diversified, and representative, these techniques contribute to more accurate and comprehensive SEACM system analysis, modeling, and evaluation. Along with synthetic data synthesis, data augmentation, and transfer learning, techniques like data sampling and ensemble learning are crucial in creating fake datasets. Table 1 provides an overview of different synthetic dataset techniques, highlighting their strengths, weaknesses, some applications in transportation, and the theory behind them. In order to make sure that the generated datasets accurately reflect the characters and behaviors of the real system, further tasks such as validation and verification are still required.

Furthermore, it is necessary to confirm the consistency and dependability of the synthetic data using various experiments and simulations. There are many ways to do this, such as comparing the generated data to previous or expected results, or the data may be assessed for consistency through time or across different data sources. Several approach developed in literature that can be used to find a qualitative and quantitative understanding of a large dataset [28].

Table 1. Literature overview—synthetic data generation techniques.

Dataset Generation Technique	Strengths	Limitations	Application in Transportation	Mathematical Theory/ Architecture
Data Augmentation	• Easily applicable to existing datasets. • Can increase dataset size. • Preserves original data distribution.	• Limited in generating entirely new scenarios. • High risk of generating unrealistic or contradictory data.	Contributes in augmenting real traffic data to improve the accuracy of traffic flow prediction models.	Applies mathematical functions (e.g., Markov chain process, kernel classifier, Gaussian noise, and affine transformations)
Generative Adversarial Network (GAN)	• Can create realistic and diverse datasets. • Captures complex relationships between variables.	• GAN training can be unstable and require careful tuning. • Mode collapse can lead to limited diversity in generated data.	Might be used in generating synthetic urban movement patterns for vehicles, pedestrians, and cyclists to assist in urban mobility planning.	Applies mathematical functions (e.g., Markov chain process, kernel classifier, Gaussian noise, and affine transformations)
VAE (Variational Autoencoder)	• Captures latent data distribution.	• Can generate blurry images.	Can contribute to improve traffic flow reconstruction and contributes to improve traffic management.	Encoder-Decoder architecture
Deep Belief Network	• Can capture complex and hierarchical data patterns. • Generates high-dimensional and realistic data.	• Training can be computationally intensive. • Requires a large amount of training data.	Can provide valuable insights for designing traffic signal timing and strategies that minimize congestion.	Constructs multilayered generative models with latent variables, using deep neural networks to model
Transfer Learning	• Leverages knowledge from source domains. • Accelerates dataset generation for the target domain. • Can capture high-level patterns and relationships.	• Requires the availability of suitable source datasets. • May require fine-tuning for domain adaptation.	Can be used in vehicle trajectory prediction and public transit demand forecasting.	Fine-tunes pre-trained models using target domain data

Table 1. *Cont.*

Dataset Generation Technique	Strengths	Limitations	Application in Transportation	Mathematical Theory/ Architecture
Bayesian Networks	• Represents uncertainty and probabilistic relationships. • Incorporates expert knowledge. • Handles missing data effectively.	• Assumes independence between non-adjacent nodes. • Complex models can be hard to interpret.	Might be used in modeling probabilistic relationships between traffic flow and road conditions to optimize traffic signal timing and reduce congestion.	Probabilistic graphical model represents the conditional probability for the corresponding random variables [9]
Monte Carlo Simulation	• Generates a diverse set of scenarios. • Enables the analysis of rare events.	• Computationally intensive for complex models. • Requires accurate input distributions.	Used in traffic impact assessment	Employs random sampling from defined input distributions, using statistical methods

3. Relevance of Artificial Datasets in Shared Electric Automated and Connected Mobility Systems

Accurate modeling and simulation are essential for shared mobility systems' performance assessment and operational optimization. Artificial datasets offer scalable and realistic simulations [29]. These databases can reproduce a number of features of automated shared mobility systems, including location strategies, vehicle dynamics, charging infrastructure, and user behavior. Researchers can assess system performance in diverse conditions, test the efficacy of algorithms, and investigate the impact of numerous factors on system performance by utilizing artificial datasets.

3.1. Data Requirements for Modeling and Simulation in Shared Electric Automated and Connected Mobility Systems

Extensive datasets covering key components are vital for accurately modeling and simulating shared electric automated and connected mobility systems. These datasets should encompass diverse data types, including vehicle trajectories, charging infrastructure details, traffic patterns, user behavior, and environmental conditions. However, due to challenges like privacy concerns, data unavailability, and the dynamic nature of these systems, collecting real-world data can be challenging [29]. In such cases, artificial datasets serve as a valuable alternative, offering researchers controlled and customizable data that closely mirror real-world characteristics.

To ensure accuracy in the modeling and simulation process, the following data requirements must be considered [2,30]:

1. Vehicle trajectories: Essential for dynamic modeling, this involves collecting data on the position, speed, acceleration, and direction of vehicles over time.
2. Charging infrastructure data: This category includes details on the location, capacity, availability, and usage of charging stations, contributing to evaluating the effectiveness and efficiency of the dynamic system.
3. Traffic patterns: Crucial for simulating shared mobility systems, traffic models aid in understanding traffic patterns, the impact of data on traffic levels, congestion, and the configuration of the road network.
4. User behavior: These data encompass variables influencing system demand and operation, such as trip start and end points, travel preferences, and transportation mode choices.
5. Environmental factors: Environmental conditions, including weather and road conditions, significantly impact system efficiency and user behavior. Additionally, these data contribute to designing more accurate and comprehensive models for SEACM systems, deepening the understanding of system dynamics, facilitating performance evaluation, and formulating effective deployment strategies.

3.2. Challenges in Data Collection from Real-World Scenarios

The process of gathering real datasets for SEACM systems faces numerous challenges, prompting researchers to devise solutions to overcome these hurdles. These challenges span a range of issues, including securing exclusive data from service providers, ensuring data privacy and security, managing the complexity and diversity of the data, and collecting data from diverse operational settings and geographic locations. Additionally, the data gathering process often involves high costs and logistical constraints, impeding the accessibility and comprehensiveness of real-world statistics.

To address these challenges, one effective strategy involves turning to synthetic datasets. Artificial datasets offer a solution by providing researchers with immediately accessible datasets that assist in achieving specific study objectives [31]. These datasets are intentionally crafted to replicate real-world scenarios and mimic the behavior of SEACM systems, achieved through techniques like data augmentation or data synthesis. The use of artificial datasets allows researchers to bypass limitations associated with real-world

data collection, enabling extensive analyses and simulations within a controlled and reproducible environment [32].

Overall, the challenges linked to collecting data from real-world scenarios highlight the necessity of exploring alternative approaches, such as leveraging artificial datasets. Through the incorporation of artificial datasets and the consultation of the relevant literature, researchers can effectively tackle the obstacles associated with real-world data collection, paving the way for advancements in the field of SEACM systems.

3.3. Benefits of Artificial Datasets in Simulation Studies

Employing artificial datasets in simulation studies for SEACM systems offers a multitude of advantages, significantly enhancing the efficacy and reliability of research findings. Expanding on these key advantages, we emphasize the following [32–35]:

(1) Enhanced control and systematic exploration: The utilization of artificial datasets provides researchers with the capability to operate within a controlled environment. This control facilitates the systematic exploration of various scenarios and parameters. Through extensive sensitivity analyses and performance evaluations, researchers can gain valuable insights into the behavior and performance of these intricate systems.

(2) Coverage of diverse situations and edge cases: Artificial datasets empower the generation of data that span a broad spectrum of situations, capturing rare or challenging-to-obtain edge cases. In the real world, encountering these specific scenarios can be challenging or infrequent. Artificial datasets address this limitation by offering researchers the means to simulate and study these situations in a controlled manner. This comprehensive coverage contributes to a more thorough understanding of system behavior and performance. Synthetic datasets provide researchers with the tools needed to model and meticulously investigate these scenarios, which can be difficult to encounter in the actual world.

(3) Reproducibility and comparability: The use of generated datasets brings about several advantages, including the ease with which research results can be replicated and compared. Researchers can share the datasets they develop, enabling others to use the same data for evaluation and analysis. This fosters transparency and collaboration within the scientific community. Additionally, researchers employing comparable artificial datasets can compare results, facilitating accurate comparisons and meta-analyses.

These advantages underscore the pivotal role of artificial datasets in simulating the stochastic behavior of shared automated and networked electric mobility systems. Leveraging these datasets deepens the understanding of these systems' behavior, ultimately enhancing the optimization and deployment of such systems.

4. Training ML Models for Performance Evaluation

Machine learning techniques are typically employed for numerous tasks, such as demand forecasting, trajectory optimization, and vehicle repositioning, when developing machine learning models to assess the effectiveness of SEACM systems [30]. Because artificial datasets offer a variety of representative datasets, they are essential for training these algorithms. Machine learning models can be effectively trained on synthetic datasets to assess their performance and enhance their accuracy, generalization skills, and precision. This section will cover a variety of machine learning (ML) approaches that are frequently employed in shared mobility systems, as well as how they are trained using synthetic datasets.

4.1. ML Models in Shared Mobility Systems

Machine learning (ML) models, which enable the prediction of numerous critical system performance aspects, including user behavior, charging station usage, and vehicle demand, are a much-appreciated tool for enhancing the efficacy of SEACM systems [30]. To accomplish these purposes, numerous machine learning models are typically employed,

such as clustering techniques, reinforcement learning strategies, regression models, and classification models. The goal and the properties of the data are used to determine which machine learning models are best.

Regression models are also helpful in quantitative variable forecasting [36]. These models can provide valuable insights into variables like time, weather, and local events that affect the demand for vehicles. By analyzing historical data, these models are able to predict trends in future demand. Operators may use these data to better allocate resources and develop fleet management plans, which makes them really important. Additionally, Classification techniques play a significant role in enhancing the understanding of user preferences and behaviors within shared mobility systems [37]. By using a diversity of datasets, including demographic data, geographic data, and trip-related characteristics, these models can categorize people into distinct groups. This important information is utilized to create tailored recommendations, improve customer service, and create focused marketing efforts, among other things. To assess data from shared mobility systems in the literature, clustering techniques have been used in numerous research studies. By combining related users or places, they facilitate the identification of trends and groups. The total customer experience is enhanced by this grouping, which makes it easier to provide personalized services, pricing schemes, and operational procedures [38]. They can help identify the best places to install charging stations and provide further details on traffic patterns. These data can be used to accurately make judgments about infrastructure design, leading to the creation of a shared mobility system that works well.

Furthermore, reinforcement learning techniques are increasingly being used in SEACM modeling to boost system performance [39]. These models employ a trial-and-error methodology to learn from interactions with the environment and make well-informed decisions. Furthermore, to increase customer happiness, lower operational costs, and boost system efficiency, dynamic pricing, fleet management, and route optimization issues can all be resolved with reinforcement learning.

In this sense, ML models can be used by researchers and industry professionals to make data-driven decisions, gain a valuable understanding of the dynamics of shared mobility networks, and help in the creation of more efficient and environmentally friendly transportation options.

4.2. Supervised and Unsupervised Learning Approaches

Supervised and unsupervised learning are the two main methods for training machine learning (ML) models to measure the effectiveness of shared electric automated and connected mobility (SEACM) systems. These approaches are extremely significant, with each providing special benefits that are essential to the optimization of SEACM systems. For the purpose of making predictions or classifying data according to target variables, supervised learning models are trained on labeled datasets. Conversely, unsupervised learning locates inherent patterns and correlations in data without the need for intentional labeling. There are benefits and drawbacks to each technique, and the best approach depends on the particular research goals and the availability of data.

A. Supervised Learning:

Supervised learning involves training models on labeled datasets to make predictions or classify data based on target variables. In the context of SEACM, where labeled datasets are often available, supervised learning becomes a powerful tool for recognizing patterns from historical data [39]. Notably, the work of Chang et al. [40] utilizes supervised learning approaches with long short-term memory networks to predict the charging demand of electric vehicles. Using real-world datasets measured from fast-charging stations in Jeju Island, South Korea, the predicting performance of the proposed model is evaluated by comparison with existing deep learning techniques. The results have demonstrated that the suggested model performs better in predicting the power consumption for fast charging when data from several charging stations are combined. Furthermore, the prediction

of future charging demand and better planning of the charging infrastructure are made possible by their trained machine learning algorithms using labeled charging data.

B. Unsupervised Learning:

In contrast, unsupervised learning can be applied to study and identify underlying structures and patterns in data. Since this method does not require labeled data, it is useful when the dataset is unstructured or does not have preset categories. Several features of electric vehicles can be mentioned here, such as the driving cycle, the batteries that are utilized, and the charging stations. These features were examined in the work of Nazari, Hussain, and Musilek [41], utilizing unsupervised learning techniques. Using clustering-based algorithms, they were able to replicate the behavior of EV users, the cycle of EV driving, the classification of EV batteries, and EV charging stations. In the same way, to anticipate EV parking and load the day before, previous charging records are fed into an unsupervised clustering algorithm and multilayer perceptron [42]. Cross-validation findings from experiments demonstrate that the model can efficiently predict online EV load and schedule charging control.

On the other hand, while selecting a learning technique, it is crucial to take the study objective, the data that are accessible, and the computational resources into account. While unsupervised learning can find hidden structures and patterns in unlabeled data, supervised learning can produce precise predictions or classifications when there is a sufficient amount of labeled data available. Therefore, hybrid approaches that integrate supervised and unsupervised learning can also be used to study electric and automated shared transportation systems.

The comprehension of system dynamics and performance is enhanced by these learning strategies, which are essential to SEACM research. Selecting a learning technique requires the careful consideration of the available data, computational resources, and research objectives. Supervised learning produces precise predictions or classifications when there is an adequate amount of labeled data; unsupervised learning, on the other hand, finds patterns and hidden structures in unlabeled data. Furthermore, the analytical capacities of automated and electric shared mobility systems may be further improved by the application of hybrid methodologies that include supervised and unsupervised learning.

4.3. Feature Engineering and Selection

For the optimal evaluation of machine learning (ML) models and the development of features, feature engineering and selection play key roles. The objective of feature engineering is to enhance the model's predictive capability by transforming the data and creating additional features [43]. Factors like time, environmental conditions, traffic conditions, and the availability of charging stations are taken into account. Incorporating these variables during feature engineering allows the model to identify meaningful patterns in the data, resulting in improved performance.

Additionally, the identification of key features that significantly enhance model performance is crucial. Employing feature selection methods enables the extraction of the most discriminative features, exerting the greatest influence on model accuracy. Common approaches for feature selection include correlation analysis, and regularization methods are frequently used to simplify models and enhance interpretability. Integrating both feature development and selection approaches enhances the capability of machine learning models to predict the future behavior of SEACM networks [43].

4.4. Training on Artificial Datasets

The lack of real data often forces machine learning models for SEACM systems to be trained on synthetic datasets. It is possible to replicate the true nature and measure the essential features of the system by creating clearly synthetic datasets [44]. By training machine learning models using simulated datasets, it could be feasible to assess the stability and performance of the models, offering valuable information for system implementation. The ability to simulate various scenarios and setups is another benefit of using synthetic

datasets for machine learning model training, which is sometimes difficult to accomplish with real data [45]. When numerous scenarios are shaped with the aid of synthetic data, it becomes easier to analyze the behavior and performance of machine learning models in a variety of contexts and assess how adaptable they are.

Additionally, ML models may be built to analyze traffic flow patterns and enhance the placement of charging stations by generating synthetic datasets that may imitate particular properties of SEACM systems [46]. It is imperative to guarantee that the generated data accurately depict the actual behavior of the system and accurately replicate its operational mechanisms. In this context, the synthetic dataset should take into account several aspects of shared electric mobility, including vehicle trajectory, user behavior, charging patterns, and environmental variables [45]. It is essential to simulate the trained machine learning models on real system data in order to improve the dependability of the model outputs. Additionally, one can assess the generalization abilities of the models by comparing their performance on synthetic datasets.

5. Deployment Considerations

For SEACM systems with autonomous repositioning to be successfully deployed, deployment planning and decision-making procedures must be carried out using artificial datasets. The benefits of this approach are that decision makers may analyze several deployment scenarios, assess the influence of system characteristics, and optimize resource allocation by using synthetic datasets to imitate system behavior. This section will discuss the various uses of synthetic datasets in deployment studies, including policy formulation and sensitivity analysis.

5.1. Real-World Challenges and Limitations

The primary difficulties and restrictions encountered by ML models trained on artificial datasets of SEACM systems are highlighted in this section [45].

5.1.1. Uncertainties in Data Quality and Availability

Real data availability and data source reliability are important factors that need to be carefully considered as they can impact the quality and effectiveness of the machine learning models that have been constructed. However, locating accurate and detailed data from real systems to train machine learning models is quite difficult. ML models' dependability and performance can be impacted by a number of issues, including noise, incoherence, and missing data.

5.1.2. System Dynamics and Evolution

Numerous aspects, including random demand patterns and stochastic user preferences and behaviors, make SEACM systems complex, dynamic, and stochastic. Because of this system complexity, ML models that have been trained on static synthetic datasets find it difficult to accurately capture the dynamic nature of these systems. To guarantee the relevance and effectiveness of ML models, it is crucial to consider these dynamic elements and retrain them.

5.1.3. Impact of External Factors

The operating mechanism of SEACM systems is further made more complex and stochastic by a number of external elements, including stochastic traffic patterns, a complex surroundings, and the stochastic behavior of nearby road users. The system's behavior and performance may be significantly impacted by these external factors. ML models that are solely trained on synthetic datasets, on the other hand, may perform worse since they are unable to account for the effects of these external variables. As such, when training and implementing ML models, it is imperative to incorporate real data and external influences.

5.1.4. Transferability to Real Scenarios

Ensuring the possible deployment of machine learning models trained on synthetic datasets in real-world situations requires conducting thorough transferability analyses. However, ML models trained on artificial datasets may be less effective due to the complexity and randomness of real systems. Consequently, testing and validating machine learning models can help to assess their robustness and transferability, ensuring their dependability and application.

It is essential to look at novel approaches and strategies, such as domain adaptation [47,48], reinforcement learning [39], and integrating real data into the training process [49], in order to go beyond these obstacles and limits in SEACM systems.

5.2. Generalization of ML Models Trained on Artificial Datasets

Generalization is the primary step of ML model training using generated datasets. As a result, evaluating and improving ML models' generalization capacity takes into account a variety of parameters, including dataset structure, model complexity, and feature engineering approaches [35]. There are several strategies provided in the literature for testing generalization, including cross-validation, which divides data into training and validation sets and tests the model on fresh and unlabeled datasets. Independent test sets can also offer a more precise estimate of generalizability. Another useful strategy is transfer learning, which uses knowledge from one dataset to improve the model's performance on another [22,50].

Furthermore, the applicability of generalization methodologies in the context of shared mobility systems has also been studied. Hua et al. [51] conducted an investigation into several machine learning models and transfer learning techniques related to cross-modal demand forecasting. Their findings made clear how well transfer learning works to improve forecast accuracy. Similarly, Huang et al.'s [52] experimental investigation on transfer learning for traffic forecasting demonstrated how transfer learning may be used to improve traffic speed prediction models' performance in data-poor places by utilizing large amounts of data from other cities.

By taking these characteristics into account and applying the appropriate approaches, machine learning models that have been trained on synthetic datasets may be effectively assessed and enhanced to achieve optimal generalization on actual data. For SEACM systems to be deployed and operated effectively, this is essential.

5.3. Ethical and Privacy Concerns

The use of machine learning (ML) models in shared mobility systems raises important ethical and privacy issues that need to be addressed. Nevertheless, biases or discriminatory behaviors that may be present in the training data may be amplified when ML models are trained on synthetic datasets. To reduce these risks, it is crucial to guarantee equity, transparency, and accountability in the ML model's implementation.

On the other hand, many additional studies using novel fairness-aware methodologies for training and performance evaluation should be carried out to address ethical difficulties [53]. Fairness-aware models seek to reduce bias and ensure the equal treatment of persons from various demographic groups [54]. We can achieve more equal outcomes in shared mobility systems by taking race, gender, and socioeconomic position into account during model building.

When using personal data for training and assessment purposes, privacy is an important factor to take into account in addition to fairness. Enforcing privacy legislation and putting strong data anonymization and protection methods in place are crucial. The illegal use or disclosure of sensitive information can be avoided by implementing privacy-preserving strategies like federated learning or differential privacy [55]. In addition, the repositioning algorithm must consider ethical considerations in its decision-making process. Ensuring fair treatment across various demographic groups during vehicle repositioning is essential to avoid unfair outcomes. Furthermore, the repositioning problem is not protected

against cyber security threats. Integrating robust cybersecurity measures in the design and deployment of repositioning algorithms is crucial to maintain the integrity of the system. For instance, consider a machine learning model trained for image recognition, such as a system used in autonomous vehicles to identify traffic signs. Cyber attackers, with knowledge of the model's architecture and access to the training data, can modify images to deceive the model. In another scenario, an attacker may manipulate an image to make the model specifically recognize a stop sign as a yield sign. This could be used maliciously to deceive autonomous vehicles or other systems relying on accurate image recognition. Clear policies and agreements that specify how data will be gathered, stored, and used should also regulate the sharing and usage of data in shared mobility systems. It is imperative to acquire informed consent from consumers for the gathering of their data and to maintain transparency regarding the intent and extent of data usage. We should endeavor to create reliable shared mobility systems that put user privacy, equality, and society well-being first by taking these ethical and privacy considerations into account.

6. How to Design/Conceive/Manage Case Studies and Experiments

In this section, case studies and experiments from real-world applications of artificial datasets for modeling and simulating SEACM systems with autonomous repositioning are presented. These case studies will highlight the particular issues raised, the datasets that were employed, the procedures that were adhered to, and the lessons that were obtained from the trials. Examples could be analyzing the effects of various car repositioning algorithms, gauging the operation of the system under various demand patterns, or maximizing the location of charging infrastructure.

6.1. Description of the Case Studies

To fully evaluate the efficacy and applicability of artificial datasets, we present a number of case studies in this section that have been or are currently being carried out in the context of SEACM systems (as shown in Figure 3). These case studies offer ideas to improve machine learning models trained on artificial datasets by addressing particular scenarios and operational factors.

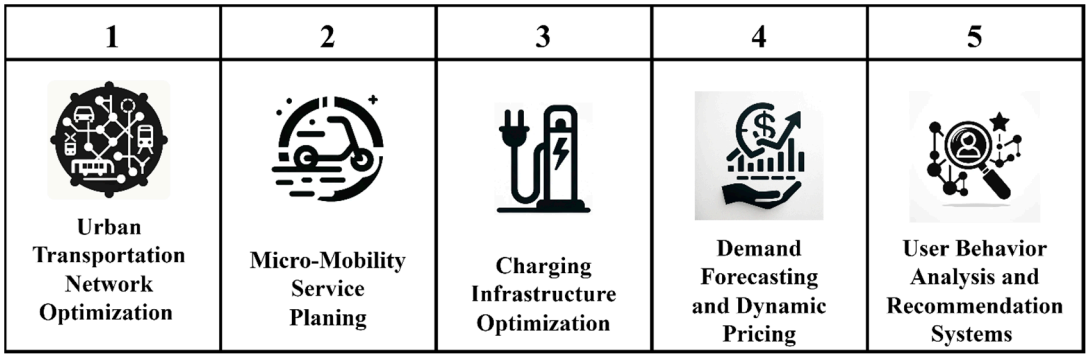

Figure 3. Real world scenarios where artificial datasets are needed.

6.1.1. Urban Transportation Network Optimization

This case study aims to show the SEACM system's performance in a complex metropolitan transportation network [56]. It seeks to optimize truck routes, reduce energy consumption, and boost overall system efficiency using machine learning models trained on synthetic datasets. This is accomplished by modeling various situations, such as gridlock, fluctuations in user demand, and the presence of charging stations.

6.1.2. Micro-Mobility Service Planning

This case study aims to assess the impact of artificial datasets in planning and improving micro-mobility services [57]. By considering user preferences, service coverage, fleet

management, and infrastructure distribution, it evaluates ML model performance using artificial datasets to forecast demand patterns, optimize vehicle distribution, and improve the accessibility and availability of micro-mobility choices.

6.1.3. Charging Infrastructure Optimization

This case study addresses the optimization issue of charging infrastructure for SEACM network [39]. The study suggests an approach to evaluate the accuracy and efficiency of ML models in forecasting charging demand and identifying optimal charging station locations. Furthermore, this work develops strategies to optimize the use of charging resources by using artificial datasets. The developed models can simulate different charging scenarios, including different charging station capacities, locations, and charging protocols.

6.1.4. Demand Forecasting and Dynamic Pricing

This case study focuses on predicting user demand patterns and developing dynamic pricing approaches for shared automated and connected electric transportation systems [35]. The study's goal is to evaluate how accurately ML models forecast demand, adapt pricing strategies in real-time, and maximize mobility system employment and profitability. This is accomplished by using synthetic datasets that represent different user behaviors, environmental factors, and pricing approaches.

6.1.5. User Behavior Analysis and Recommendation Systems

Using synthetic datasets, this SEACM case study seeks to provide important insights into how ML model performance and efficacy in SEACM systems may be enhanced [58]. Furthermore, by combining the features of various datasets, decision-makers are able to create well-informed choices for the deployment, optimization, and design of systems.

6.2. Experimental Setup

When assessing the dependability of case studies, the experimental component is crucial. To help readers better comprehend the research findings, this part provides a detailed explanation of all the important elements and variables used in the trials.

6.2.1. Simulation Platform

Choosing the right simulation platform is essential for accurate and realistic investigations. A strong simulation framework that can handle complexity, combine modules, and successfully use synthetic datasets is needed for the case study in order to model SEACM systems.

6.2.2. Artificial Datasets

It is unavoidable to use synthetic datasets to assess the performance of these stochastic systems due to the difficulty of genuine data for SEACM systems being unavailable. Consequently, it is necessary to create artificial datasets that incorporate multiple configurations and accurately reflect the real nature of systems. Furthermore, it is imperative to accurately determine the dimensions of the datasets to guarantee appropriate system behavior depiction. In order to improve the realism and diversity of the datasets, it is also necessary to apply modern data augmentation and synthesis techniques during the generation procedures.

6.2.3. ML Models and Algorithms

Machine learning methods and algorithms can be used to assess the SEACM network's performance. They are able to use or incorporate methods like data-driven strategies, demand-responsive optimization, and reinforcement learning. These machine learning models can be evaluated for their effectiveness in streamlining decision-making processes, allocating resources, and improving system performance by comparing them to baseline techniques.

A thorough explanation of the experimental setup is intended to guarantee complete transparency and the reproducibility of the study carried out. Researchers and practitioners can replicate and expand upon the findings by using the information provided here to better understand the approaches that will be used.

6.3. Results and Analysis

In this section, we describe the methodical processing of the outcomes derived from a carefully carried-out case study and offer the structure of a subsequent in-depth examination of the research conclusions. A wide range of performance metrics, such as scalability, accuracy, and efficiency, will be included in the thorough evaluation. These metrics are crucial for determining how well machine learning (ML) models that have been trained on synthetic datasets are working. Beyond simple numerical measurements, a thorough study explores the advantages and disadvantages of the machine learning models used in the research. Furthermore, the purpose of this review study is to provide insights into the advantages and disadvantages of employing generated datasets in shared mobility systems. For validation, it recommends contrasting case study findings with actual data or relevant research. The following are some of the several aspects that will be brought to light by the study of the mentioned results.

6.3.1. Performance Metrics

The performance measures for machine learning models trained on synthetic datasets, such as predictive accuracy, training efficiency, and inference procedures, are examined and discussed in this paper. The performance of the model and its suitability for shared mobility systems may also be fully understood by analyzing these measures.

6.3.2. Strengths and Limitations

To evaluate the strengths and limitations of the used machine learning models, a review will be carried out. Aspects including the model's ability to handle complex scenarios and adapt to changing contexts will be considered in this evaluation.

6.3.3. Potential Benefits and Challenges

In SEACM systems, the employment of artificial datasets can enhance resource allocation, system efficiency, and forecast accuracy. Furthermore, the difficulties posed by artificial datasets are covered, including the necessity for representative and varied data, the possibility of biases arising from data production, and the computing demands involved in training machine learning models on artificial datasets of considerable size.

6.3.4. Contextualization of Findings

One might compare the collected results with real-world data or cite other studies that have looked into related topics in order to establish a wider context for the conclusions drawn from a particular case study. An enhanced comprehension of the consequences and significance of the research undertaken in relation to the case study is made possible by this contextualization, which also serves to validate the current findings.

This research paper's results and analysis section essentially explains how to perform an in-depth analysis of the results derived from a particular case study. Several performance measures are investigated through a thorough study, along with the advantages and disadvantages of ML models, possible advantages and difficulties, and pertinent references are used to contextualize the results. This section will provide insights that will further our understanding of artificial dataset generation for shared mobility systems.

6.4. Case Study Specification of SEACM Systems, Categories of Data Needed to Address Repositioning and Assignement Issues

Based on our previous work on SEACM addressing the repositioning issue [2], we formulated a conceptual framework describing the mechanism process of this such systems

(see Figure 4). The conceptual framework of a system might be considered as a spatial queuing model within stochastic service. Customers have the option to book or reserve electric autonomous vehicles (EAVs) online from the nearest station. If a vehicle is available, customers wait; otherwise, they exit the system. Upon availability, the EAV reaches the customer for pick-up. Following drop-off at the destination, vehicles initiate the next service, park, head to a charging station for battery recharge, or reposition themselves to balance the network across different areas. In contrast, by introducing a rebalancing strategy, the vehicles need to have larger batteries or need to recharge more frequently. The proposed concept considers the crucial constraint of the battery charging level.

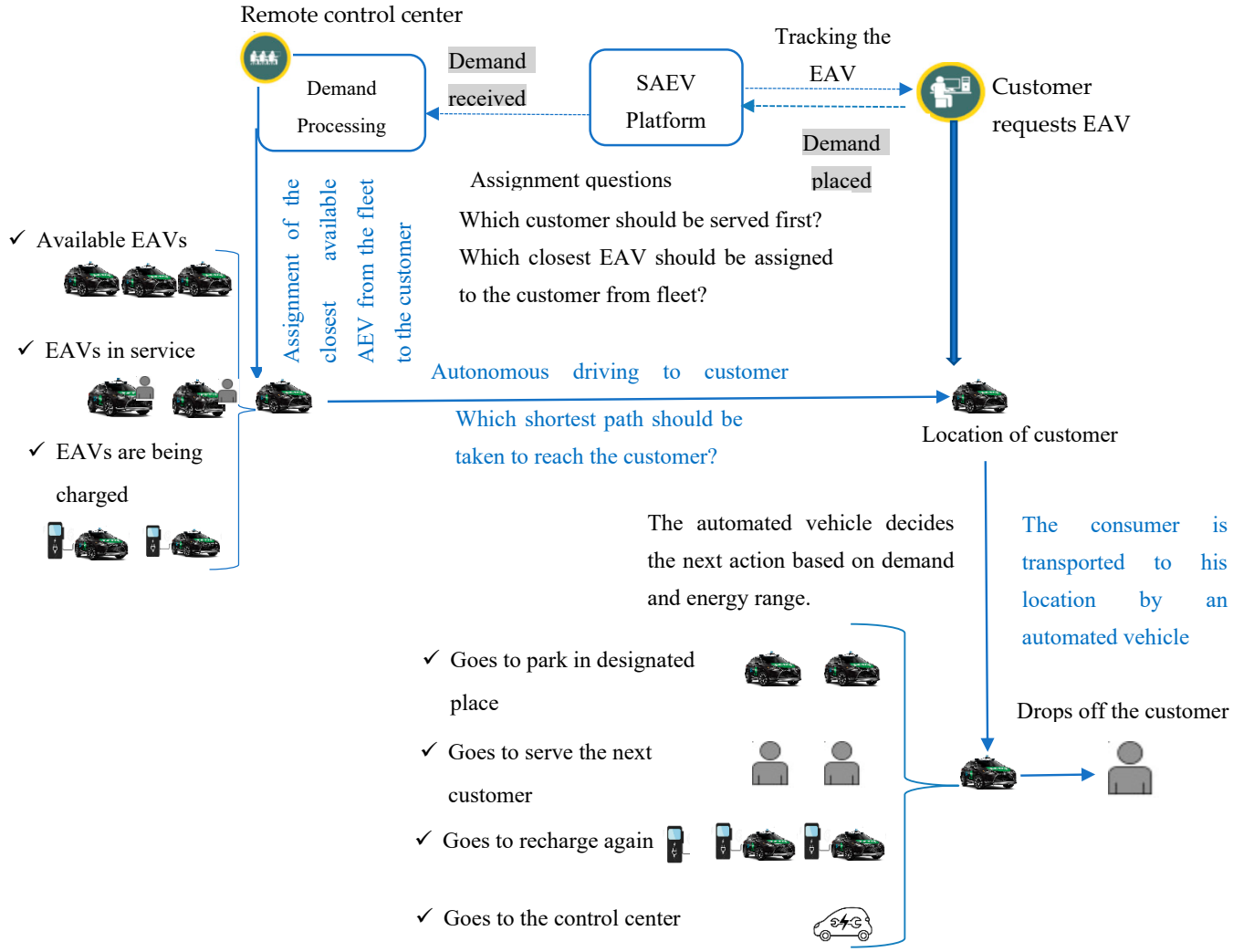

Figure 4. Shared electric automated and connected mobility concept.

The literature presents various operational concepts for shared electric automated mobility, and we adopted a simplified approach [5,6,59,60]. The concept framework includes five EAV operational statuses: parking, reserved, occupied, navigating to a charging station, and charging. When a user requests a trip, an electric autonomous vehicle (EAV) is assigned from the fleet in real-time based on demand and the EAV's battery range (including inactive, in-service, or charging station AVs), where the status of the EAV changes from parked to reserved. Once the EAV picks up the customer, its status changes to occupied. Following the trip, the EAV makes a decision among heading to a charging station (based on its battery range), moving to the next user, going to a designated waiting spot for the next request, or returning to the control center. As a basic strategy, users can reserve the electric vehicle only if it is fully charged. The available EAVs include the ones that are

parkedand the ones that are charging, provided that have a sufficient state of charge. In other cases, customers can opt to reserve an available electric autonomous vehicle (EAV) provided that the battery's charging level (L) is equal or superior to the proposed availability threshold ($L \geq s$). Suppose that during the ride, the EAV might stop at multiple locations for customer pick-ups or drop-offs. We made an initial assumption of situating an *Nsch* charging station at each area and considered the system subject to customer requests randomly, following a probabilistic process. In contrast, shared mobility systems face with repositioning challenges. We illustrate the network as unbalanced when there is an unequal distribution of vehicles expected in the middle term. This is different from that of conventional sharing mobility, since automated vehicles can automatically balance the network. This means that certain regions might experience either a shortage of vehicles (with almost none present) or an excessive number of vehicles. Consequently, automated repositioning mechanisms are crucial for shared vehicles to tackle imbalances and enhance network efficiency, mitigating issues in congested or underserved areas. The repositioning strategy involves the autonomous movement of vehicles from the congested area *Ai* (with a high number of EAVs) to the underserved area *Aj* (with rare vehicles). This approach takes into account the battery range of each vehicle, ensuring that only charged vehicles are relocated. Many constraints can affect the model such as customer waiting time, pick-up travel time, taxi–customer distance, and battery range. To optimize the use of e-chargers, we assume that when an EAV is fully charged (100%), it is going to free up the charger in the charging station. It can then either find a parking location or respond to the next demand with priority. This approach ensures charger availability throughout the day, considering the limited number of chargers across the city.

Regarding the assignment policy, we assume that the employed assignment policy in the model prioritizes assigning a fully charged vehicle to fulfill the initial demand. In cases where no fully charged vehicle is accessible, the system prioritizes vehicles with a 50% charging level. However, if all vehicles at the charging station have less than 50% charge, the assignment process designates a parked vehicle within area Ai for customer pick-up. Otherwise, the system suggests vehicles in ride with a customer as a final option (see Figure 5).

Once a ride is completed and the customer is dropped off, the automated vehicle is capable of autonomously performing the following actions:

1. Navigate to serve the next request as long as no available EAV is parked in the area, provided that its battery range is sufficiently charged for the next trip.
2. Navigate to the charging station to achieve either a full charge or a 50% charge level.
3. Navigate to park somewhere if no charging points are accessible.
4. Proceed to the control center if any technical issues are identified, necessitating further diagnostics and resolution.

In SEACM, data distribution involves sharing information within and between vehicles. These data include details about the vehicle's surroundings, like the positions of other vehicles, obstacles, traffic signals, road signs, and road conditions. Additionally, it includes data concerning the vehicle's own status, speed, acceleration, and direction [61].

Based on the provided description of the shared electric automated and connected systems, our goal is to use machine learning to predict future behavior and determine the optimal timing for initiating the repositioning process before the network becomes unbalanced. To achieve this, the ML model would require various categories of data to create an effective predictive model [6]. In this section, we detail the data categories we might need to train the ML model, along with examples of specific data within each category.

Demand data: these data can include, firstly, the trip requests data providing a great deal of information regarding user trip requests. For instance, the pick-up time from EAV to customer, the drop-off time, the number of customers in the trip, the booking time, as well as the pick-up and drop-off locations, ride durations, the pick-up distance from EAV to customer, and trip distances [5,6,62]. For instance, a trip request was made on 1 August 2023, at 08:30 AM, starting from location A and finishing at location B, with a ride

duration of 15 min. Secondly, the user preferences data offer insights into user behavior and preferences. These could include their favorite travel times, frequently selected routes, and a history of bookings.

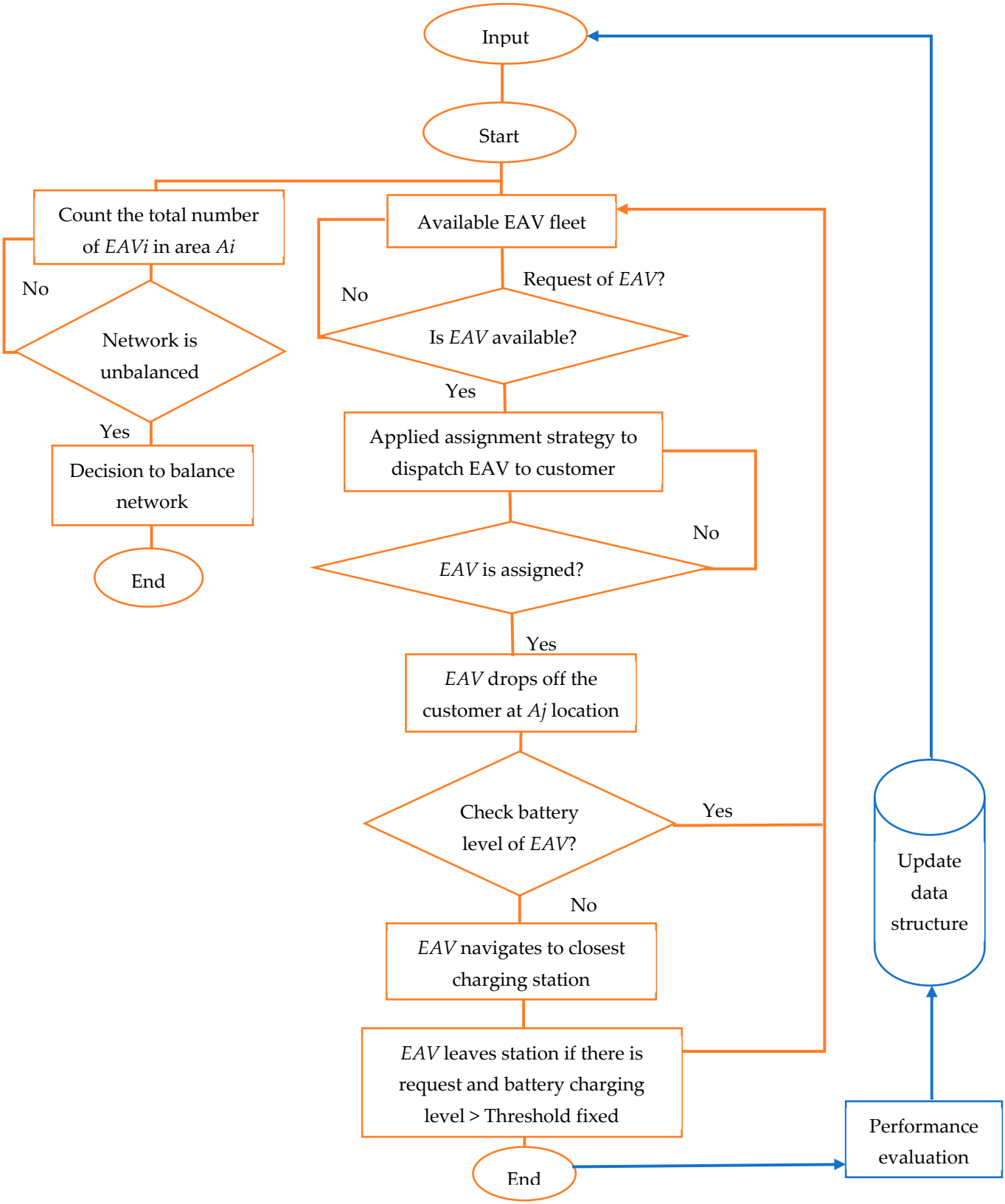

Figure 5. Flowchart illustrating the assignment and repositioning mechanism concept.

Vehicle data: the vehicle data category includes essential information related to the EAVs such as vehicle location within the SEACM network. In addition, it includes battery status data of the EAV such as current charging levels, remaining battery capacity of vehicle, battery condition, and historical charging data [61]. For instance, EAV1—Battery level: 80%, last charged: 30 July 2023. Furthermore, it includes the operational status of vehicle, which is also crucial information, indicating whether an EAV is inactive, in-service, or at a charging situation.

Charging infrastructure data: these data represent information about the charging stations within the SEACM network. These data include accurate geographic coordinates (latitude and longitude) for each station, helping efficient navigation. Additionally, they provide details on charging capacities, indicating the maximum number of vehicles in the station that can be charged at the same time, such as "number of chargers 10 in station S1." Occupancy rates are also important; they indicate the current charging EAV in relation to the total capacity of the station, for instance, "Occupancy: 6/10" means that six out of ten available chargers are currently in use at a specific station.

Geographic information: Geospatial data can be obtained from various sources, including public datasets, commercial providers, and government agencies. Many platforms allow developers to access geospatial data programmatically. For instance, Google Maps, Mapbox, and OpenStreetMap (OSM). OpenStreetMap is a community-driven mapping project that provides free geospatial data. In the context of SEACM, most studies on shared mobility request geographic information regarding spatial data, including the coordinates of areas *Ai* and *Aj*, and local traffic patterns. They also request route data: information about routes and distances between various locations within the city [5]. For instance, Area *Ai*—Latitude: 42.1234, Longitude: −71.5678, High traffic during peak hours.

Historical repositioning data: these include specific information about the repositioning timing, locations, and success rates of previous repositioning actions. The launching time of repositioning indicates when repositioning actions were initiated, providing a chronological history of repositioning activities. Locations represent the specific areas or regions involved in each repositioning action. This includes the origin (Area Ai) and destination (Area *Aj*) locations, offering insights into the movement of vehicles across the network. Success rates represents the effectiveness of the repositioning strategy and indicate the number of repositioning actions that achieved their proposed goals. For instance, repositioning 1 has been executed on 20 July 2023 from Area *Ai* to Area *Aj*, with a success rate of 90%.

Assignment data: these data provide a history of vehicle assignments, subject to constraints such as charging levels of EAV and waiting time of pick-up. For instance, at such and such time EAV-10- at such and such location and associated with such and such charging level has been dispatched to pick up customer X-2- at such and such location. These data contribute to the assessment of assignment policies, enhancing the overall user experience within the shared electric automated and connected mobility system.

Vehicle condition data: these data include technical problem reports and the history of maintenance, which might give important information about vehicle condition and repair. For instance, EAV3, which has had a brake problem on 25 July 2023, was then successfully repaired and fixed by 28 July 2023. This dataset can give more information about the average EAV moving to control per day or week, which can help in the design of effective repositioning and assignment strategies.

Network balancing metrics: many metrics could be evaluated in the context of repositioning, such as EAV distribution metrics, i.e., the number of EAVs per area, and trends over time. For instance, Area Ai—Number of EAVs: 15, Area *Aj*—Number of EAVs: 5. These metrics can provide valuable information on the distribution of vehicles within the network, and help the formulation of effective repositioning strategies to balance the SEACM system efficiently.

Employing the data described above will help in the training of predictive models that forecast demand, and make informed decisions about repositioning strategies to achieve a balanced network of SEACM.

6.5. Real World Application in the Context of Shared Electric Automated and Connected Mobility

In this section, we explore a real-world case study based on the shared autonomous electric vehicle (SAEV) system framework conducted by Kim et al. [63]. This comprehensive framework, as depicted in Figure 6, comprises a fleet operation model, charging station model, and SAEV design, providing a comprehensive approach to optimize the operation of shared autonomous electric vehicles. The fleet operation model uses road circumstances for the shortest time-path search and determines the optimal fleet assignment by considering the destination selection of inactive vehicles and the charging schedule. Simultaneously, the charging station model determines optimal locations for charging stations and the total number of chargers, optimizing their placement based on demand patterns and operational considerations.

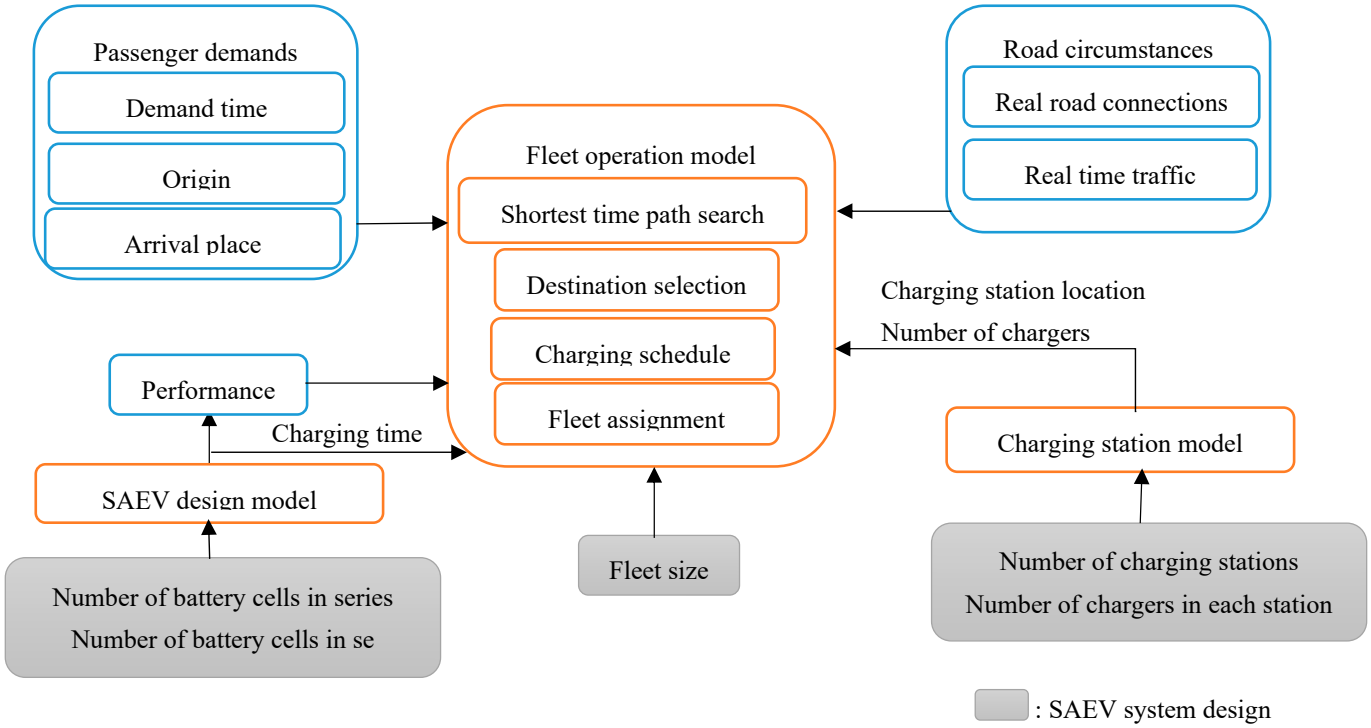

Figure 6. Shared automated and electric vehicle system framework.

Kim et al. specifically addressed the relocation issue in shared automated and electric vehicles by deploying deep learning techniques. They introduced a machine learning (ML) framework capable of forecasting the optimal repositioning of inactive vehicles within the network under different traffic conditions. This repositioning mechanism comprises three key models, including a user demand prediction model, an inactive vehicle optimization repositioning model, and a deep learning model for parameter estimation. The user demand prediction model predicts the location and frequency of user demand, using actual taxi call data provided by the Seoul Metropolitan Government (SEO-Taxi 2), collected over 15 months—from August 2015 to October 2016 [64]. The inactive vehicle optimization repositioning model, based on predicted demands, efficiently decides inactive vehicle relocation, considering factors such as passenger demand, vehicle status, and charging station states. The deep learning model further refines this process by estimating optimal parameter values in real time, adapting to changing conditions within the shared autonomous electric vehicle system. Overall, three factors are used as input: (1) location

and frequency of user demands 30 min later, based on the historical passenger demand location and frequency data; (2) charging-station congestion, indicating charger occupancy and charging end times; and (3) status of vehicles, including current location, remaining battery capacity, and service status.

Based on simulations of Seoul City's SAEV operation, that study aims to determine optimal configurations for minimizing operation costs and reducing waiting times. Seongsin Kim's work provided a comprehensive framework and a practical solution to address the complex challenges of fleet management, charging station optimization, and electric vehicle design in the context of shared autonomous electric mobility. Key outcomes include an optimal SAEV fleet size, battery design, charging station locations, and the number of chargers at each station. That study is considered as a valuable real-world example for addressing repositioning issues in shared electric automated and connected mobility.

Nevertheless, despite the fact that the study conducted by Kim et al. demonstrated the effectiveness of their SAEV framework using real-world data, it is important to consider the potential challenges and biases that might be present in the dataset. For instance, changes in city infrastructure, regulations, or the introduction of new transportation services could impact the model's applicability to more recent or different urban environments. Additionally, ensuring data privacy and ethical considerations related to the use of real-world taxi service data should be addressed in the paper. Considering the complexities of dynamic urban environments, the incorporation of generated synthetic data can be used to augment the real data, may help to represent diverse scenarios, and reinforce the model's adaptability and predictive capabilities. Moreover, generated data can simulate diverse conditions and variations in demand or future conditions that may not be present in the historical data, helping the model generalize better to unexpected circumstances or changes in the urban environment. In addition, in cases where privacy concerns or ethical considerations limit the use of real data, artificial data can play a crucial role to create representative scenarios without the violation of individual privacy.

In addition, to boost the SAEV model proposed by Kim et al., involving the generated data in training of the model may increase the prediction efficiency of the model in determining the optimal repositioning of inactive vehicles. This augmented dataset may comprises diverse traffic conditions, and empowers the machine learning model to make informed decisions about the future destination of inactive vehicles. Furthermore, the generated data, carefully calibrated to replicate real-world complexities, enable the model to make strategic decisions. They reinforce the model not only to predict the optimal relocation path but also to proactively address potential network imbalances. By providing the model with a variety of synthetic scenarios, we enhance its ability to avoid unbalanced network situations, ensuring a more robust, adaptive, and effective relocation strategy for inactive vehicles.

7. Discussion and Future Directions

7.1. Evaluation of Artificial Dataset Generation Techniques

It is essential to assess synthetic dataset generation methods in order to guarantee that the resulting data adequately capture the essential features of SEACM systems and successfully replicate the real nature of the system. Conducting thorough comparison studies to assess the capabilities, drawbacks, and performance of various data production strategies is essential to enhancing the efficacy of such systems.

These techniques involve the creation of synthetic data, data augmentation, and transfer learning [23,25,44–47]. We can learn more about how these methods successfully reproduce real-world settings in generated datasets by carrying out an in-depth analysis. This facilitates the process of choosing appropriate approaches and producing datasets of superior quality. In comparison research, dataset efficiency is guaranteed by using metrics such as accuracy and mean square error. Additionally, it is essential to assess the created datasets' effectiveness by examining how they perform under various conditions, such as changes in data distribution, size, and complexity [65]. Simplifying the assessment of

synthetic dataset generation techniques requires the creation of benchmark datasets and consistent evaluation procedures. Furthermore, the establishment of open platforms for the sharing of created datasets and evaluation results might promote collaboration and expedite the advancement of this area of research.

7.2. Integration of Real and Artificial Data

The merging of fake and real data is essential for improving the reliability and realism of simulation studies in the field of modeling and simulating shared electric automated and linked mobility systems with autonomous repositioning [66]. Artificial datasets may not have the dynamic quality and complexity of real-world data, despite having several benefits including scalability and control over data parameters. For this reason, it is essential to combine artificial datasets with real data—that is, genuine user behavior patterns, traffic patterns, and infrastructure characteristics—in order to obtain more realistic representations of the system dynamics.

Combining artificial and real data presents potential as well as obstacles. On the one hand, gathering data from the real world can be difficult and time-consuming [45]. It necessitates gathering data from multiple sources, including user surveys, GPS devices, and sensors. Furthermore, protecting the security and privacy of data is crucial [55]. However, by validating and calibrating their simulation models against actual situations, researchers can increase the validity and applicability of their findings through the integration process.

Investigating approaches and strategies that successfully integrate synthetic and real data in simulation studies should be the main emphasis of future study. It is necessary to do this by generating strong data fusion algorithms that can manage diverse data sources and take care of problems with data quality. The effect of including real data on the precision and dependability of simulation findings should also be measured. Comparison studies that examine how well simulations running with artificial data alone perform against those running with a mix of artificial and actual data can help achieve this.

7.3. Potential Research Directions

The use of artificial datasets in modeling and simulating shared electric automated and connected mobility systems can be further advanced in a number of potential approaches. Among them are the following:

(1) Investigating advanced machine learning methods: To create more varied and life-like simulated datasets, future research could focus on applying advanced machine learning methods, such as generative models and deep learning [16,17,19,67]. These methods could improve the accuracy of simulations of real-world situations and boost the dependability of generated datasets.
(2) Reviewing how data generation methods affect machine learning models: Investigating the effects of various data production methods on the functionality of machine learning models in simulation experiments is essential. Selecting the best methods to produce useful synthetic datasets for the modeling and assessment of the SEACM network will be made easier by being aware of its advantages and disadvantages.
(3) In order to address issues with data lack and privacy, approaches for maintaining privacy can be developed during the creation of synthetic datasets. While protecting the privacy of sensitive data, these techniques must retain the statistical properties of actual data. Researcher access to realistic datasets for simulation studies can be expanded and data restrictions can be solved by addressing these issues.

7.4. Discussion of Research Outcomes on Automated Shared Mobility Systems in the Context of Data

Certain works have employed surveys to estimate mobility demand or have used historical data to generate synthetic datasets (e.g., [62]). In the same way, some works [60,68] introduced an agent-based simulation model. This simulation model was driven by synthetic data generated through survey. The researchers conducted a series of simulations

across various fleet sizes and capacities, analyzing scenarios both with and without considering relocation strategies.

Liang and Jing [6] applied real-world data to assess their dispatching strategy proposed of E-automated Taxis. They introduced an artificial neural network (ANN)-model for redistribution that has learned optimal dispatch strategies generated from an optimization model. Similarly, Wang and Guo [61] have addressed the repositioning and dispatching issues of shared electric and automated vehicles. They developed a decision-making framework of multi-task dynamic relocation of shared electric and automated vehicles, combined with a theoretical optimization model based on deep reinforcement learning, bipartite graphs, and network maximum flow methods. The simulations were run using real-world data from the Didi platform that includes an operational dataset of ride-hailing in Chengdu, Sichuan Province, China (including the real-time position of vehicles, origin and destination information of the daily trip order, charging station locations, and the sub-regions). In addition, Turoń [69] has tackled various challenges of shared mobility systems in smart cities, employing a machine learning model trained on pre-existing data from previous studies. The developed model is used to determine and assess the accuracy of trips undertaken by users of shared mobility systems.

Indeed, the extensive deployment of fully shared autonomous taxi remains limited, resulting in a lack of real-world operational data. Very few studies used real data to train ML models on SEACM systems [6,61]. In addition to the modeling complexities involved in designing extensive shared electric automated systems, most research studies face a significant obstacle in sourcing real data relevant to shared electric automated and connected mobility systems [5,12,13,60]. This includes critical data components such as battery range, assignment of Vehicle Vi to a customer Cj, electric autonomous vehicle (EAV) locations, EAV pick-up trip data, and charging capacities of stations. Acquiring real data from automotive manufacturers and obtaining accurate information about customer locations and their historical demands is a considerable challenge. Consequently, the integration of real data, specifically from automated taxi operators such as UBER, and EAV manufacturers like TESLA, Nio, and Toyota, can result in powerful strategic decisions about the deployment feasibility of the SEACM network.

8. Conclusions

In the framework of modeling and simulation with autonomous repositioning of the SEACM system, this literature review provided a comprehensive investigation into the creation of synthetic datasets. Moreover, this paper explores several advanced methods for generating artificial datasets, including transfer learning, data augmentation, synthetic data creation, and others. In addition to improving the diversity and realism of datasets used in the modeling of these intricate systems, these methods provide vital solutions to the issue of the scarcity of actual data.

This paper also discussed the importance of synthetic datasets, focusing on the data needed for precise modeling and simulation, the difficulties in collecting real datasets, and the advantages of using synthetic datasets. Fundamentally, without being limited by a lack of real-world data, synthetic datasets may be employed to conduct comprehensive testing, evaluate various system configurations, and measure performance under various circumstances.

Furthermore, a few fundamental research concerns that are relevant to these new automated E-mobility systems were brought up in this review paper and could potentially be technically addressed in our upcoming studies, as follows:

- In what ways can the methods for creating synthetic datasets capture the essential features of these dynamic and stochastic systems while maintaining their realistic nature?
- Which strategies are used to generate synthetic datasets of SEACM systems, and what are their strengths and limitations in terms of performance?
- How is the SEACM network's simulation more realistic and reliable when it combines actual and artificial datasets?

- How might machine learning model performance and generalization be impacted by the creation of diverse datasets in SEACM system simulations?
- Can real-world scenarios be effectively used with machine learning models that were trained on artificial datasets?
- How may machine learning approaches advance to improve the variety and realism of synthetic datasets?

Finally, the significance of artificial datasets in SEACM system modeling and performance assessment was highlighted in this literature review. To evaluate the effectiveness of artificial dataset generation techniques, integrate real and artificial data, and solve privacy and data lack concerns, more research is necessary.

In the future work, after collecting/generating the required data illustrated above, we can create a set of diverse features that would be crucial for constructing an accurate machine learning model. These samples are designed to involve the comprehensive complicated mechanism of system operation and multiform characteristics that can influence the dynamics of the system. Through these diverse data, we will be able to capture invaluable information on the behavior of the shared automated and connected electric mobility system. Good employment of this information could play a crucial role in the development of predictive models capable of predicting user demand accurately. In addition, using these datasets could help to construct predictive models that can forecast future situations of the SEACM network (congested area, deficit area), and make decisions about repositioning strategies to achieve a balanced network and optimal performance.

Author Contributions: Conceptualization, K.K., A.K.K. and T.B.; methodology, K.K., A.K.K., T.B. and W.V.K.; validation, K.K., W.V.K. and T.B.; writing—original draft preparation, T.B., K.K. and A.K.K.; writing—review and editing, T.B., P.N.B. and W.V.K.; visualization, P.N.B. and T.B.; supervision, K.K. and P.N.B.; formal analysis, T.B. and A.K.K.; project administration, W.V.K., T.B. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Afshar, V. The Car of the Future Is Connected, Autonomous, Shared, and Electric. Available online: https://www.zdnet.com/article/the-car-of-the-future-is-connected-autonomous-shared-and-electric/ (accessed on 8 December 2020).
2. Benarbia, T.; Kyamakya, K.; Al Machot, F.; Kambale, W.V. Modeling and Simulation of Shared Electric Automated and Connected Mobility Systems with Autonomous Repositioning: Performance Evaluation and Deployment. *Sustainability* **2023**, *15*, 881. [CrossRef]
3. Li, L.; Pantelidis, T.; Chow, J.Y.; Jabari, S.E. A real-time dispatching strategy for shared automated electric vehicles with performance guarantees. *Transp. Res. Part E Logist. Transp. Rev.* **2021**, *152*, 102392. [CrossRef]
4. Zhang, W.; Wang, K.; Wang, S.; Jiang, Z.; Mondschein, A.; Noland, R.B. Synthesizing neighborhood preferences for automated vehicles. *Transp. Res. Part C Emerg. Technol.* **2020**, *120*, 102774. [CrossRef]
5. Sanchez, N.C.; Martinez, I.; Pastor, L.A.; Larson, K. On the simulation of shared autonomous micro-mobility. *Commun. Transp. Res.* **2022**, *2*, 100065. [CrossRef]
6. Hu, L.; Dong, J. An Artificial-Neural-Network-Based Model for Real-Time Dispatching of Electric Autonomous Taxis. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 1519–1528. [CrossRef]
7. Yuan, Y.; Cheng, H.; Sester, M. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3054–3061. [CrossRef]
8. Patella, S.M.; Scrucca, F.; Asdrubali, F.; Carrese, S. Carbon Footprint of autonomous vehicles at the urban mobility system level: A traffic simulation-based approach. *Transp. Res. Part D Transp. Environ.* **2019**, *74*, 189–200. [CrossRef]
9. Rath, S.; Liu, B.; Yoon, G.; Chow, J.Y. Microtransit deployment portfolio management using simulation-based scenario data upscaling. *Transp. Res. Part A Policy Pract.* **2023**, *169*, 103584. [CrossRef]
10. Wang, X.; Mavromatis, I.; Tassi, A.; Santos-Rodriguez, R.; Piechocki, R.J. Location anomalies detection for connected and autonomous vehicles. In Proceedings of the 2019 IEEE 2nd Connected and Automated Vehicles Symposium (CAVS), Honolulu, HI, USA, 22–23 September 2019; pp. 1–5.

11. Muthurajan, S.; Loganathan, R.; Hemamalini, R.R. Deep Reinforcement Learning Algorithm based PMSM Motor Control for Energy Management of Hybrid Electric Vehicles. *WSEAS Trans. Power Syst.* **2023**, *18*, 18–25. [CrossRef]
12. Karandinou, A.A.; Kanellos, F.D. A Method for the Assessment of Multi-objective Optimal Charging of Plug-in Electric Vehicles at Power System Level. *WSEAS Trans. Syst. Control* **2022**, *17*, 314–323. [CrossRef]
13. Miok, K.; Nguyen-Doan, D.; Zaharie, D. Generating Data using Monte Carlo Dropout. *arXiv* **2019**, arXiv:1909.05755v2.
14. Frick, M.; Axhausen, K.W. Generating Synthetic Populations using Iterative Proportional Fitting (IPF) and Monte Carlo Techniques. In Proceedings of the 3rd Swiss Transport Research Conference (STRC 2003), Ascona, Switzerland, 19–21 March 2003.
15. Ilahi, A.; Axhausen, K.W. Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta. *Reg. Stud. Reg. Sci.* **2019**, *6*, 623–636. [CrossRef]
16. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
17. Islam, Z.; Abdel-Aty, M.; Cai, Q.; Yuan, J. Crash data augmentation using variational autoencoder. *Accid. Anal. Prev.* **2021**, *151*, 105950. [CrossRef] [PubMed]
18. Hinton, G.E. Deep belief networks. *Scholarpedia* **2009**, *4*, 5947. [CrossRef]
19. Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **2022**, *16*, 100258. [CrossRef]
20. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
21. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
22. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
23. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **2014**, *3320–3328*, 3320–3328.
24. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-trained models: Past, present and future. *AI Open* **2021**, *2*, 225–250. [CrossRef]
25. Chiba, S.; Sasaoka, H. Basic study for transfer learning for autonomous driving in car race of model car. In Proceedings of the 2021 6th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 20–21 May 2021; pp. 138–141.
26. Liberty, E.; Lang, K.; Shmakov, K. Stratified sampling meets machine learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2320–2329.
27. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
28. Cyril, P.; Jürg, S.; Faez, A. Guidelines for Creating Synthetic Datasets for Engineering Design Applications. *arXiv* **2023**, arXiv:2305.09018v1.
29. El Emam, K.; Mosquera, L.; Hoptroff, R. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*; O'Reilly Media: Sebastopol, CA, USA, 2020.
30. Narayanan, S.; Chaniotakis, E.; Antoniou, C. Shared autonomous vehicle services: A comprehensive review. *Transp. Res. Part C Emerg. Technol.* **2020**, *111*, 255–293. [CrossRef]
31. Cai, J.; Deng, W.; Guang, H.; Wang, Y.; Li, J.; Ding, J. A survey on data-driven scenario generation for automated vehicle testing. *Machines* **2022**, *10*, 1101. [CrossRef]
32. Tang, S.; Zhang, Z.; Zhang, Y.; Zhou, J.; Guo, Y.; Liu, S.; Guo, S.; Li, Y.-F.; Ma, L.; Xue, Y.; et al. A Survey on Automated Driving System Testing: Landscapes and Trends. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 1–62. [CrossRef]
33. Huang, Z.; Hale, D.K.; Shladover, S.E.; Lu, X.Y.; Liu, H.; Li, Q.; Li, X.; Mahmassani, H.; Talebpour, A.; Hosseini, M.; et al. *Developing Analysis, Modeling, and Simulation Tools for Connected and Automated Vehicle Applications*; No. FHWA-HRT-21-077; Federal Highway Administration, Office of Operations Research and Development: McLean, VA, USA, 2021.
34. Wang, C.; Xie, Y.; Huang, H.; Liu, P. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. *Accid. Anal. Prev.* **2021**, *157*, 106157. [CrossRef] [PubMed]
35. Zhong, Z.; Tang, Y.; Zhou, Y.; Neves, V.D.O.; Liu, Y.; Ray, B. A survey on scenario-based testing for automated driving systems in high-fidelity simulation. *arXiv* **2021**, arXiv:2112.00964.
36. Yu, L.; Feng, T.; Li, T.; Cheng, L. Demand prediction and optimal allocation of shared bikes around urban rail transit stations. *Urban Rail Transit* **2023**, *9*, 57–71. [CrossRef]
37. Abouelela, M.; Lyu, C.; Antoniou, C. Exploring the Potentials of Open-Source Big Data and Machine Learning in Shared Mobility Fleet Utilization Prediction. *Data Sci. Transp.* **2023**, *5*, 5. [CrossRef]
38. Fauser, J.; Hertweck, D. Identifying e-scooter sharing customer segments using clustering. In Proceedings of the 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), Stuttgart, Germany, 17–20 June 2018; pp. 1–8.
39. Liang, Y.; Ding, Z.; Ding, T.; Lee, W.J. Mobility-aware charging scheduling for shared on-demand electric vehicle fleet using deep reinforcement learning. *IEEE Trans. Smart Grid* **2020**, *12*, 1380–1393. [CrossRef]
40. Chang, M.; Bae, S.; Cha, G.; Yoo, J. Aggregated electric vehicle fast-charging power demand analysis and forecast based on LSTM neural network. *Sustainability* **2021**, *13*, 13783. [CrossRef]
41. Nazari, M.; Hussain, A.; Musilek, P. Applications of Clustering Methods for Different Aspects of Electric Vehicles. *Electronics* **2023**, *12*, 790. [CrossRef]
42. Xiong, Y.; Wang, B.; Chu, C.-C.; Gadh, R. Electric Vehicle Driver Clustering using Statistical Model and Machine Learning. In Proceedings of the 2018 IEEE Power & Energy Society General Meeting (PESGM), Portland, OR, USA, 5–10 August 2018; pp. 1–5. [CrossRef]

43. Orzechowski, A.; Lugosch, L.; Shu, H.; Yang, R.; Li, W.; Meyer, B.H. A data-driven framework for medium-term electric vehicle charging demand forecasting. *Energy AI* **2023**, *14*, 100267. [CrossRef]
44. Lucini, F. The real deal about synthetic data. *MIT Sloan Manag. Rev.* **2022**, *63*, 11–13.
45. Lu, Y.; Wang, H.; Wei, W. Machine Learning for Synthetic Data Generation: A Review. *arXiv* **2023**, arXiv:2302.04062.
46. Kar, A.; Prakash, A.; Liu, M.Y.; Cameracci, E.; Yuan, J.; Rusiniak, M.; Fidler, S. Meta-sim: Learning to generate synthetic datasets. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4551–4560.
47. Li, P.; Liang, X.; Jia, D.; Xing, E.P. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv* **2018**, arXiv:1801.01726.
48. Prakash, A.; Boochoon, S.; Brophy, M.; Acuna, D.; Cameracci, E.; State, G.; Shapira, O.; Birchfield, S. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7249–7255.
49. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 23–30.
50. Song, Q.C.; Tang, C.; Wee, S. Making sense of model generalizability: A tutorial on cross-validation in R and Shiny. *Adv. Methods Pract. Psychol. Sci.* **2021**, *4*, 2515245920947067. [CrossRef]
51. Hua, M.; Pereira, F.C.; Jiang, Y.; Chen, X. Transfer learning for cross-modal demand prediction of bike-share and public transit. *arXiv* **2022**, arXiv:2203.09279.
52. Huang, Y.; Song, X.; Zhang, S.; James, J.Q. Transfer learning in traffic prediction with graph neural networks. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3732–3737.
53. Bonnefon, J.F.; Černy, D.; Danaher, J.; Devillier, N.; Johansson, V.; Kovacikova, T.; Martens, M.; Mladenovic, M.; Palade, P.; Reed, N.; et al. *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*; Directorate-General for Research and Innovation (European Commission): Brussels, Belgium, 2020.
54. Wang, G.; Zhong, S.; Wang, S.; Miao, F.; Dong, Z.; Zhang, D. Data-driven fairness-aware vehicle displacement for large-scale electric taxi fleets. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; pp. 1200–1211.
55. Hahn, D.; Munir, A.; Behzadan, V. Security and privacy issues in intelligent transportation systems: Classification and challenges. *IEEE Intell. Transp. Syst. Mag.* **2019**, *13*, 181–196. [CrossRef]
56. Zhao, D.; Li, X.; Cui, J. A simulation-based optimization model for infrastructure planning for electric autonomous vehicle sharing. *Comput. Aided Civ. Infrastruct. Eng.* **2021**, *36*, 858–876. [CrossRef]
57. Comi, A.; Polimeni, A.; Nuzzolo, A. An innovative methodology for micro-mobility network planning. *Transp. Res. Procedia* **2022**, *60*, 20–27. [CrossRef]
58. Wang, G.; Zhang, Y.; Fang, Z.; Wang, S.; Zhang, F.; Zhang, D. FairCharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 1–25. [CrossRef]
59. Zhao, L.; Malikopoulos, A.A. Enhanced Mobility with Connectivity and Automation: A Review of Shared Autonomous Vehicle Systems. *IEEE Intell. Transp. Syst. Mag.* **2019**, *14*, 87–102. [CrossRef]
60. Vosooghi, R. Shared Autonomous Vehicle Service Design, Modeling, and Simulation. Ph.D. Thesis, l'Université Paris-Saclay préparée à CentraleSupélec, Paris, France, 2019.
61. Wang, N.; Guo, J. Multi-task dispatch of shared autonomous electric vehicles for Mobility-on-Demand services—Combination of deep reinforcement learning and combinatorial optimization method. *Heliyon* **2022**, *8*, 11. [CrossRef]
62. Meneses-Cime, K.; Aksun Guvenc, B.; Guvenc, L. Optimization of On-Demand Shared Autonomous Vehicle Deployments Utilizing Reinforcement Learning. *Sensors* **2022**, *22*, 8317. [CrossRef]
63. Kim, S.; Lee, U.; Lee, I.; Kang, N. Idle Vehicle Relocation Strategy through Deep Learning for Shared Autonomous Electric Vehicle System Optimization. *J. Clean. Prod.* **2022**, *333*, 130055. [CrossRef]
64. Donovan, B.; Work, D. *New York City Taxi Trip Data (2010–2013)*; The University of Illinois Urbana-Champaign: Champaign, IL, USA, 2016. [CrossRef]
65. Song, Z.; He, Z.; Li, X.; Ma, Q.; Ming, R.; Mao, Z.; Pei, H.; Peng, L.; Hu, J.; Yao, D.; et al. Synthetic Datasets for Autonomous Driving: A Survey. *arXiv* **2023**, arXiv:2304.12205. [CrossRef]
66. Mütsch, F.; Gremmelmaier, H.; Becker, N.; Bogdoll, D.; Zofka, M.R.; Zöllner, J.M. From Model-Based to Data-Driven Simulation: Challenges and Trends in Autonomous Driving. *arXiv* **2023**, arXiv:2305.13960.
67. Suo, S.; Regalado, S.; Casas, S.; Urtasun, R. Trafficsim: Learning to simulate realistic multi-agent behaviors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10400–10409.

68. Vosooghia, R.; Puchingera, J.; Jankovicb, M.; Vouillon, A. Shared autonomous vehicle simulation and service design. *Transp. Res. Part C* **2019**, *107*, 15–33. [CrossRef]
69. Turoń, K.; Kubik, A.; Chen, F. Operational Aspects of Electric Vehicles from Car-Sharing Systems. *Energies* **2019**, *12*, 4614. [CrossRef]