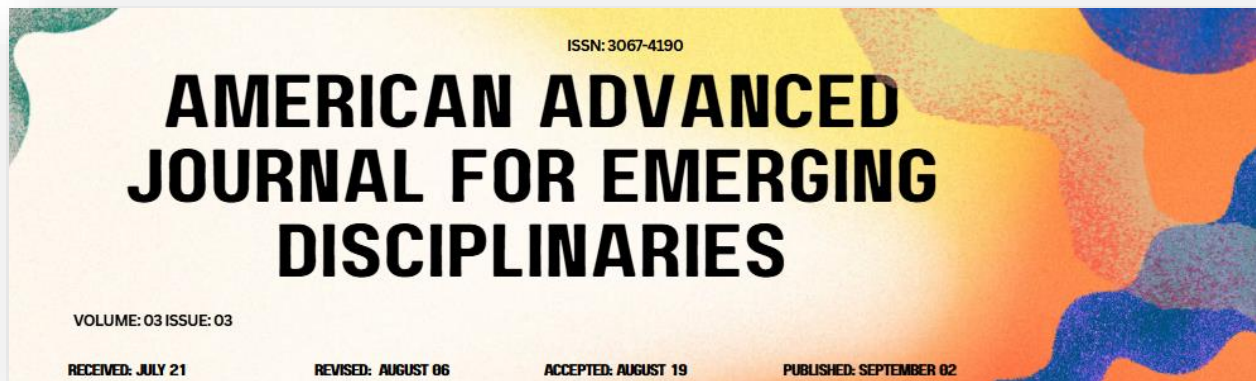




An AI and Data Engineering Perspective

Madhu Sathiri
Independent Researcher
sathirimac@gmail.com

Abstract

The mission of intelligent transportation systems (ITS) is to enhance transportation safety, efficiency, mobility, and sustainability through the integration of data from multiple sources. Despite the celebrated success of artificial intelligence (AI) in development and commercialization, its growing capacities, accessibility, and affordability have not been capitalized on in ITS. This is due partly to the lack of comprehensive and clearly delineated means for infrastructure provision, timeliness, and system continuity—especially in maintenance, operations, and expansion—and partly to the underlying data fundamentals. AI engenders unprecedented opportunities for transportation solutions by providing its own data engineering. Examination of AI-Powered Data Engineering methods reveals the enabling means of AI-based data engineering for ITS.

Data engineering encompasses the core functions for the acquisition, preparation, and deployment of data suitable for analysis and modeling. Like the broader IT domain, AI-Powered Data Engineering for ITS operates on a foundation of data acquisition and ingestion; integration and interoperability; architecture; optimization; security; governance; agency; and deployment. The function for data acquisition and ingestion serves the dual purpose of ingesting system-based information—such as from traffic signals and detection cameras used for system administration—as well as supporting predictive models for demand and congestion forecasting conversation systems for route planning and incident mitigation.

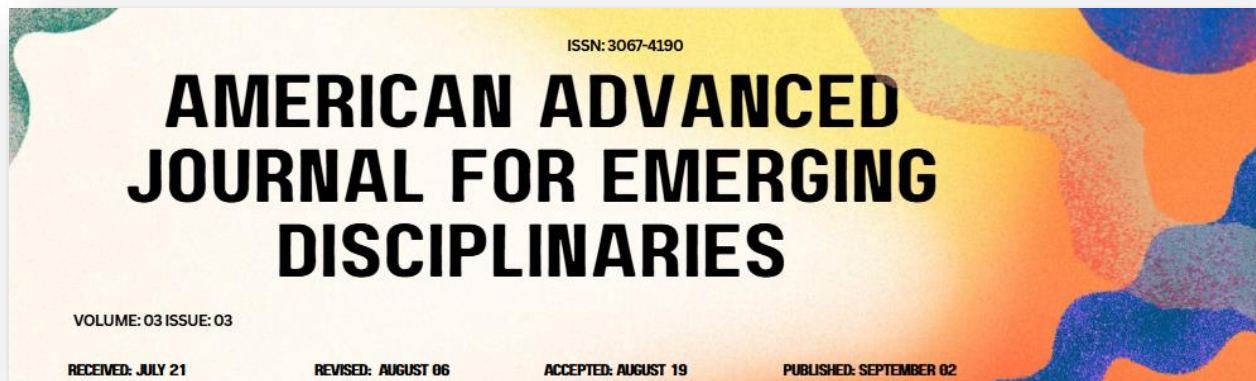
Keywords: Intelligent Transportation Systems, AI-Powered ITS, Transportation Data Integration, AI-Based Data Engineering, Transportation Safety And Efficiency, Mobility And Sustainability, Multi-Source Transportation Data, ITS Infrastructure Provisioning, Real-Time Transportation Analytics, Data Acquisition And Ingestion, Transportation Data Interoperability, ITS Data Architectures, Predictive Traffic Modeling, Demand And Congestion Forecasting, Route Planning Optimization, Incident Detection And Mitigation, Transportation System Operations, ITS Data Governance And Security, AI-Driven Transportation Solutions, Scalable ITS Data Pipelines.

1. Introduction

Having introduced the study, foundations of AI-powered data engineering, and data architectures specifically developed for Intelligent transportation systems (ITS), the chapter now focuses on how AI methods, combining deep learning with data acquired from physical processes, can analyze and utilize the data in transportation optimization tasks like predictive demand modeling, congestion modeling, computer vision for infrastructure and vehicle status monitoring, and sensor fusion, thereby assisting automation in ITS. The chapter further addresses social

aspects of AI data engineering, including data governance, privacy, ethics, bias, and fairness, before investigating lifecycle management of analytical models in ITS using MLOps principles.

AI-powered data engineering can be seen as a set of functions that shape raw input data acquired from various sources into a form that is suitable for model training and inference. These functions are typically not the main business value-creating part of the system but an enabling factor that allows knowledge discovery for the business use case. AI model serving refers to the function that executes trained models with operational data. As in enterprise data



engineering, these functions need to be performed as efficiently, smartly, and cost-effectively as possible.

1.1. Overview of the Study

Innovative digital technologies are radically changing the way in which we live and conduct day-to-day activities, including working, traveling, shopping, obtaining services, and recreational experiences. This new digital society has the potential to enable more efficient, sustainable, and equitable use of resources and services, including the road transportation infrastructure. However, this potential is often hindered because the vast amount of data generated through these activities are employed in an ad hoc manner with little or no global coordination.

Artificial intelligence (AI) has the potential to play a fundamental role in exploiting these data to optimize city life in many different aspects. Whether it is predicting the places and times where taxi or ride-sharing services are needed or the areas where traffic congestion is likely to occur, AI algorithms require a large amount of trustworthy data to be successful. Intelligent transportation systems (ITS) focus on the data generated by transportation activities and how they can be acquired, integrated, and exploited to optimize all kinds of transportation services by both private companies and municipalities. Investment in ITS infrastructure enables a much more accurate and complete set of data to be produced and made accessible for exploitation by traditional AI means.

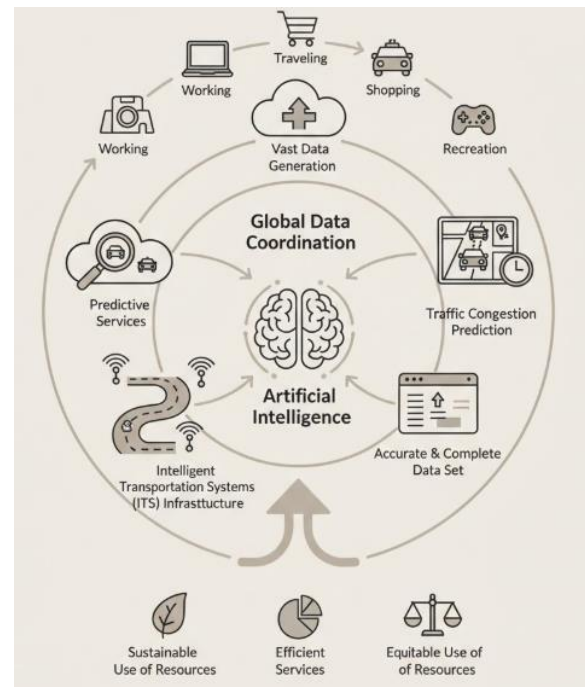
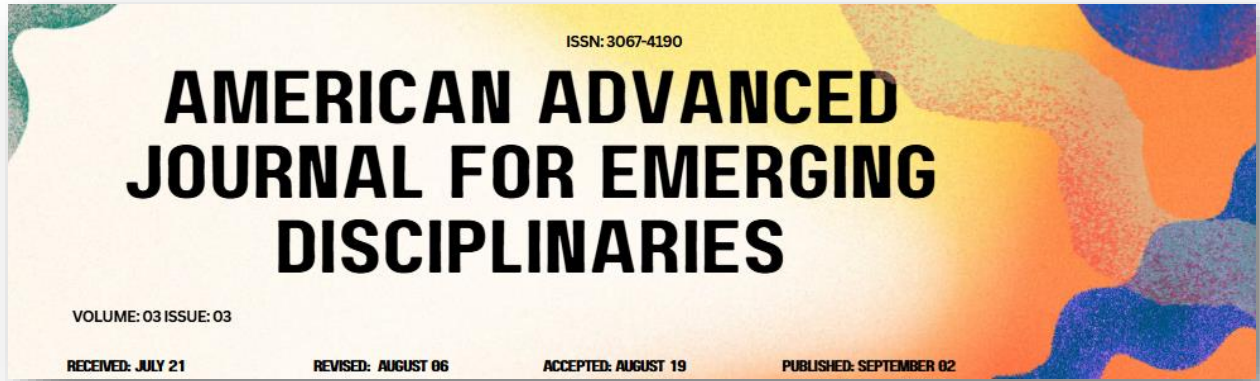


Fig 1: Synergizing Artificial Intelligence and Intelligent Transportation Systems: A Framework for Data-Driven Urban Optimization and Infrastructure Sustainability

2. Foundations of AI-Powered Data Engineering

Quality assurance has often focused on assessing the output of data science and analysis techniques for specific applications rather than on data engineering, even though the success or failure of any AI solution depends largely on the quality of the input data. AI-enabled applications must therefore be complemented by upfront AI-supported data engineering that recognizes the importance of trustworthy data, the costs that need to be incurred to obtain them, and the technologies that make the data preparation processes more efficient.

With their inherent complexity and heterogeneity, the data components required for intelligent transportation systems



(ITS) (i.e., components that address safety, efficiency, sustainability, equity, and/or citizen wellbeing in the management or use of transportation systems) are themselves part of a data engineering challenge. Ample data sources are available for ITS, many of which can have their value enhanced by exploiting information from other sources. Efficient solutions exist to integrate data from different data sources, allowing for responding to queries that require information resident on different databases. A different form of data engineering focuses on accelerating the preparation of the input data streams for any AI pipeline. Most such approaches leverage data engineering techniques that enable distributed disposal of raw information, aggregation with lower requirements for complex queries, improved cost efficiency, or distribution of processing requirements to nodes operating closer to data sources.

timestamp	actual_demand	predicted_demand	event_flag
2026-01-16 00:00:00	130.1	130.1	0.0
2026-01-16 01:00:00	121.1	129.0	0.0
2026-01-16 02:00:00	144.4	124.6	0.0
2026-01-16 03:00:00	141.4	142.5	0.0

Equation 1) Data ingestion and time indexing (streaming ITS data)

The highlights streaming/time-series ingestion for low-latency prediction . A standard formalization is:

1. Define a discrete time index:

$$t \in \{1, 2, \dots, T\}$$

2. Let the raw sensor stream at time t be a vector (speed, counts, weather, incidents, etc.):

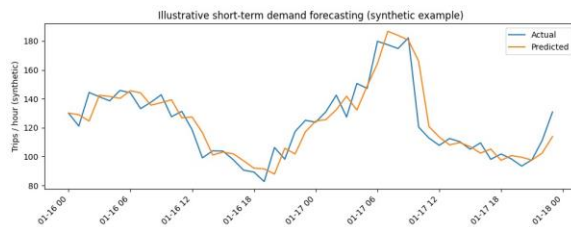
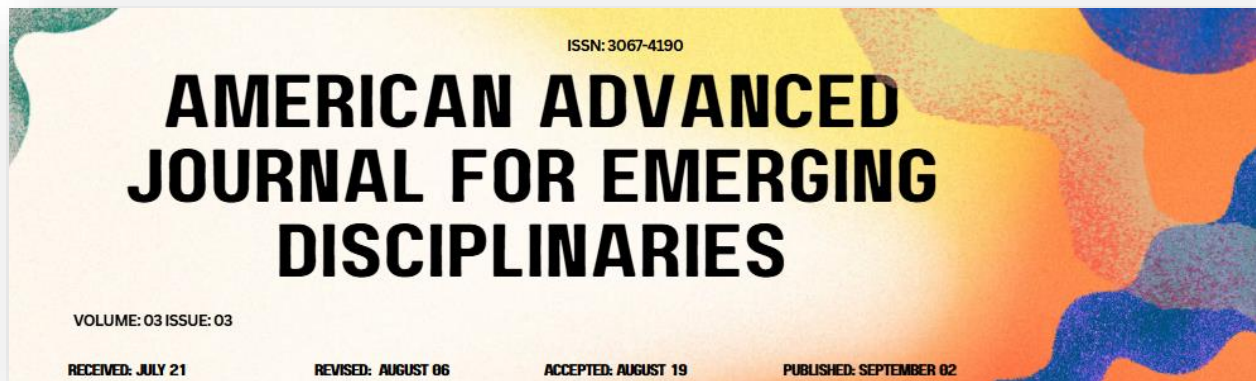
$$\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(d)}]^T$$

3. If you maintain a rolling window of length L for prediction, the model input is:

$$\mathbf{X}_{t-L+1:t} = [\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t]$$

2.1. Data Acquisition and Ingestion in ITS

Data acquisition and ingestion comprise the use of sensors and systems to collect, store, and make available, for real-time and analytics, the data needed for Intelligent Transportation System (ITS) operations. An ITS collects, processes, analyzes, and makes available data for deployment of ITS services, including for their AI methods. Data for these services originate from different sources, including automated traffic surveillance, vehicle positioning, population and vehicle movement models, prediction of events such as accidents, planned maintenance on the roads, public transport schedules, and drivers themselves. These data sources may follow different data standards and formats. Their data must be appropriately integrated, made interoperable, and stored. Real-time data acquisition in an ITS is usually defined in the context of centralized traffic management and relies on detection of vehicle presence and/or classification along the road. However, (a) traffic issues are detected more and more often with the help of drivers, and (b) the definition of real time may differ depending on the adopted ITS services. Emerging vehicle-to-everything technologies open new opportunities to create intelligent transportation systems through merging traffic management services, vehicular ad hoc networks, and intelligent transportation applications. Emerging communication technology allows for transferring telemetry data to the cloud, learning traffic patterns and anomalies to provide predictive insights to fleets and transportation as a service solution providers.



2.2. Data Integration and Interoperability

Achieving data interoperability—understood as the ability of systems to exchange and use data in a meaningful way—is essential for enabling data sharing within multi-party environments. A common approach is to create an interoperable data model that defines an integration schema that can be used to collect and connect data from disparate data sources. Such data models can take the form of ontology-based models, semantic rules, or domain-specific linked open data. In the context of ITS, numerous ontologies have been defined to facilitate data sharing among stakeholders operating in a specific domain (e.g., vehicle sensors, behaviors of nearby vehicles, board-based driver behavior), ensuring semantic alignment during integration.

Using distributed data hubs allows stakeholders in a federated setup to ‘publish’ their data while enabling other stakeholders to discover, access, and use those data sources. When the usage patterns of data are known to some degree, stakeholders can use the federated data source to compute analytics for data aggregation or event correlation before publication with an MDM model. Stakeholders should also create constraints on their published data to avoid the unprincipled use of sensitive information, for instance, to infer the potential movements of nearby vehicles and their drivers. Recent advances in knowledge graphs are also expected to help ITS data integration for service provisioning.

3. Data Architectures for Intelligent Transportation Systems

To reap the full benefits of emerging AI methods, data engineering in ITS should also consider the design and implementation of the underlying data architecture. Therefore, an overview of common data architectures is necessary to identify those whose properties best match the requirements of analytical pipelines for transportation-related data. These requirements are shape of the data (temporal, spatial, modal), freshness and update frequency, processing granularity, query styles, data processing operations and load, and, especially, distribution. Three types of data architecture and processing approach are frequently used in current ITS applications. The first concerns distributed data hubs based on a flight information management system (FIMS) architecture for federated data management, such as the Integrated Transportation Information System in Taipei, Taiwan. The second is a real-time streaming architecture, as in the swarm intelligence model of Beijing’s Traffic Management and Control Center. The third involves linked open data (LOD) or similar approaches, frequently adopted by European cities in their deployment of smart-city services. Data engineering for ITS should further extend beyond these three types of architecture to realize the full potential of its data resources and encompass a wider range of AI methods—from supervised methods all the way to self-supervised and unsupervised learning.

3.1. Distributed Data Hubs and Federated Architectures

Federated Data Hub Architectures are envisioned to enable the deployment of local transportation use cases in a multi-organizational environment without centralized storage of data at a third-party operator location. Local data incorporating sensitive private information can remain within the control of the data owner while still allowing privately sensitive data to remain protected. Instead of centralizing sensitive data for a global model training objective, the proposed approach relies on a two-tier connectivity structure where local models are trained per geographical area. Local models are jointly trained through aggregating the communicated weight updates in a federated learning fashion. The trend toward Electric, Shared, Connected, and Automated Mobility (ESCAM)

further emphasizes the growing need for viewing vertical transportation domains as an integrated system. Alongside accommodating the growing number of end-user units such as electric vehicles or shared fleets, catering for the intelligent logistic segment. This later implies serving logistic end users in a proactive way as opposed to the conventional approach based on re-active approaches such as traffic signal coordination. A Data Hub Architecture is proposed for the Connected and Automated Mobility (CAM) segment with the goal of assimilating vehicle as a sensor paradigms, ideally integrating data from Connected Vehicles (CV), logistic aggregators, Cloud-Based PLC, and Cloud-Based Road-Side Units (RSU). The Data Hub provides an integrated data source for training Computer Vision models at intersections. A real-time streaming architecture for connected transport and logistics safety is also proposed. The architecture aims to continue feeding vehicle as a sensor datasets on-camera detectable events such as accidents, traffic jam, road blockage, road anomaly, and re-routing of vehicles (if AVs, near to decision point). In the long-term, the architecture provides a system at the level of road infrastructure that supports the operation of intelligent traffic systems.

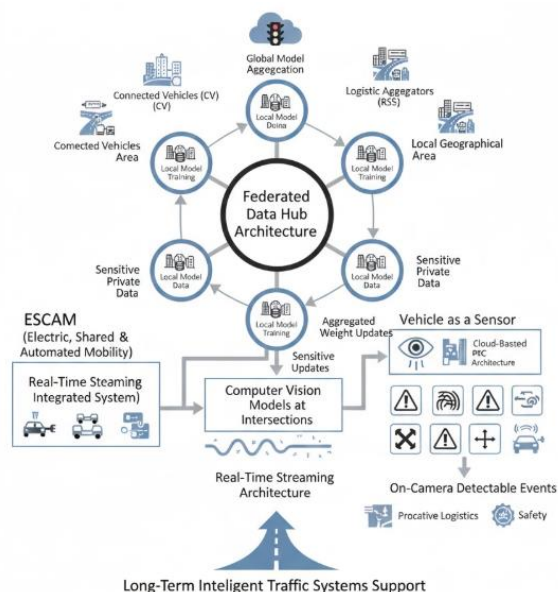


Fig 2: Federated Data Hubs for Connected and Automated Mobility: A Privacy-Preserving Architecture for Integrated Logistic Proactivity and Real-Time Event Detection

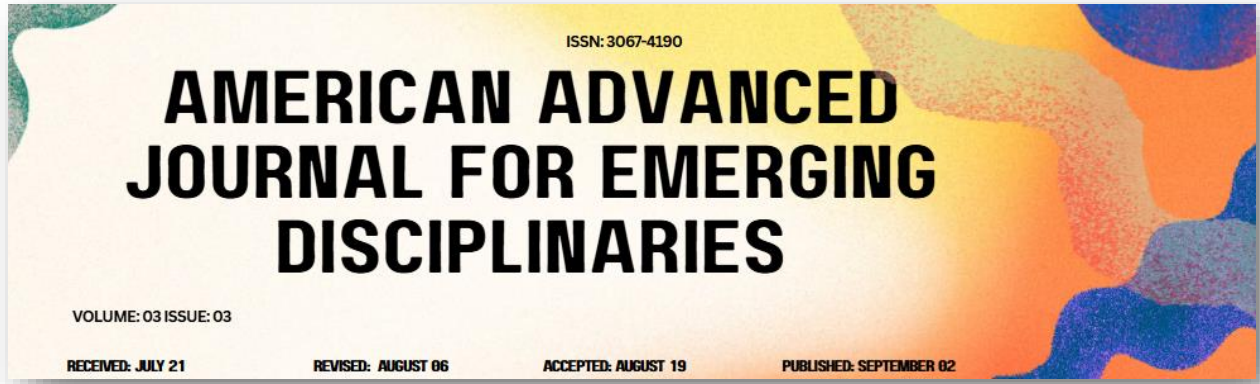
3.2. Real-Time Streaming Architectures

Real-time streaming architectures support data applications that rely on low-latency data ingestion and serve short-term decision-making, including navigational map updates, real-time congestion prediction, and incident detection. These applications require immediate situational awareness, and model predictions are only reliable for a limited time horizon. The incoming data stream typically consists of a continuous flow of time-series data and enables trained machine-learning models to produce predictions in real time.

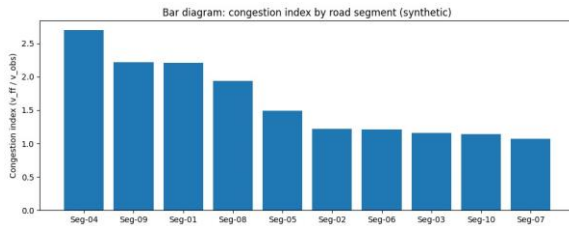
As described in Section 4.1, a dedicated model is not needed for every single prediction, as data from multiple predictions can be processed using the same model. To minimize prediction latency, Seaview uses a one-to-many model for commuting-demand prediction, which serves the multiple short-term predictions of bike-sharing demand over the same prediction horizon and at low latency. The model takes into account predicted traffic incident locations, which allow it to capture demand anymore accurately. In the case of incident detection, it is important to react as quickly as possible when abnormal patterns emerge. For patterns such as major traffic jams, dedicated model ensembles that are trained only for incident detection are applied instead of the original real-time demand-prediction model.

4. AI Methods for Transportation Optimization

A variety of AI methods optimize intelligent transportation systems (ITS) as a service ecosystem through predictive modeling of demand and traffic flow congestion—two key capabilities for effective incident detection—and through computer vision and sensor fusion to monitor vehicles and detect road and infrastructure conditions. New mobility



services, such as ride-hailing and micro-mobility, induce changes in traffic patterns and contribute to urban congestion. City authorities require real-time visibility into these dynamics to take proactive measures; for instance, predictions are needed to place demand-responsive buses and share road-space efficiently with micromobility. Methods benchmarked in the literature include Probabilistic Latent Graph-Trace Model, Tree-LSTM, graph neural networks, isotonic regression, and spatio-temporal convolutional neural networks (STCNN). Understanding traffic flow is crucial to maintaining road network performance. Congestion at intersections causes travel-time delay, dramatically increasing operating costs for freight carriers. Avoiding congestion in high-density areas contributes to a competitive and attractive freight-mode service. Incident detection relies chiefly on historical traffic conditions; an accurate real-time classification enhances performance. Demand forecasting captures short-term demand changes and identifies fluctuation patterns. Variational Recurrent-Dynamic-Factor models are applicable, for example, to plan rescaling of a bike-sharing system to minimize operating costs, boost user satisfaction, and generate revenue.



Equation 2) Short-term demand forecasting (supervised learning)

The discusses demand forecasting and short-term fluctuations .

Goal: predict demand y_{t+h} (e.g., trips next hour) from recent history and context.

Step 1: Define target and features

- Target: y_{t+h}

- Features: $\mathbf{X}_{t-L+1:t}$ and optional context $\mathbf{c}_t$ (events, incidents, weather)

Step 2: Choose a parametric model f_θ

$$\hat{y}_{t+h} = f_\theta(\mathbf{X}_{t-L+1:t}, \mathbf{c}_t)$$

Step 3: Define loss (MSE example)

Over training samples $t \in \mathcal{T}$:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (y_{t+h} - \hat{y}_{t+h})^2$$

Step 4: Optimize

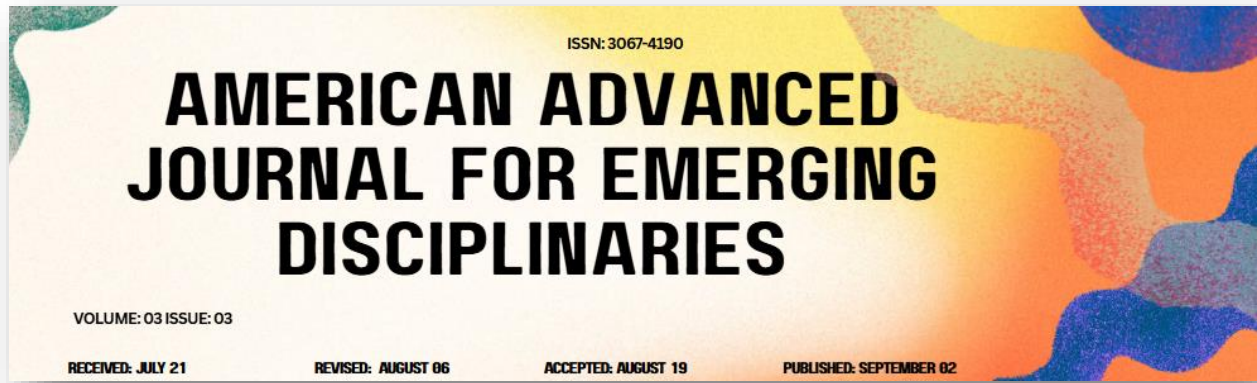
$$\theta^* = \operatorname{argmin}_{\theta} \mathcal{L}(\theta)$$

4.1. Predictive Modeling for Demand and Congestion

Machine learning pipelines trained on historical data help detect anomalous demand and congestion patterns for road agencies and municipalities. Data about vehicle counts, speeds, weather, road closures, and events can help machine learning pipelines recognize patterns in congestion. Forecasting models can predict these congestion hotspots to support early traffic congestion management, allocating resources and adjusting signals as needed to mitigate congestion.

Travel demand and vehicle origin–destination (OD) flows can also be modeled using machine learning. Estimating OD flows helps optimize public transit services, and support money-minding proactive deployments of high-occupancy-vehicle lanes, bus lanes, and more, wherever such resources could bring the maximum possible benefits. Such estimations can also help practice money-minding proactive allocation of parking resources. When combined with a well-trained traffic simulation model, such OD-estimating pipelines can improve money-minded long-term infrastructure planning by helping to identify future-demand evolution patterns.

4.2. Computer Vision and Sensor Fusion for Vehicle and Infrastructure Monitoring



AI-powered camera and sensor systems provide large-scale, cost-effective monitoring of the transportation ecosystem, at both vehicle and infrastructure levels. Such camera and sensing systems can collect high-fidelity data across many spatiotemporal scales, powered by advances in AI methods for image and video analytics. Powered by street-side, vehicle, or UAV-mounted cameras, AI techniques for object detection, classification and counting, and activity recognition are applied to monitor road users, from pedestrians and bi-cycles to cars and trucks. Integrating camera and LiDAR data can further enhance object detection capabilities, while domain adaptation techniques allow networks to be trained with synthetic data, enabling practical use of systems in diverse domains. At an infrastructure level, cameras can monitor the road and surrounding infrastructure. Road conditions (e.g. asphalt distresses), lane occupancy and hopping irregularities can be detected. Major road events such as flooding at intersections or debris can also be detected; spawned events can trigger the deployment of local resources such as tow-in services. Deploying actual vision-based detection of road events and conditions that conduct a direct impact on user sother estimators would also be beneficial. A complementary solution to vision-based camera networks can be a network of battery-less low-cost commercially-available sensors; these sensors offer low monitoring capability yet are present at a dense network while being powered by a central light source. Exploiting redundancy and sensors' own capabilities, fusion at the transportation system level allows the use of data for indirect or event-tied estimations.

metric	value
Demand MAE (trips/hr)	7.92
Worst congestion index	2.7
Final fed loss	0.128
Noise sigma at eps=1	4.84

5. Data Governance, Privacy, and Ethical Considerations

The predictive power of machine learning and deep learning is invaluable for optimizing traffic and transport. However, the usage of external data, including digital traces from social networks and mobile devices, creates substantial privacy risks that need to be mitigated. Privacy-preserving analytics can help to provide valuable insights while protecting sensitive information through spatial and temporal anonymization. New privacy-holed AI algorithms incorporate differential privacy while directly responding to demand-response requests. Furthermore, AI bias caused by under-representation of groups can harm social fairness. The combination of automated transport and ITS can exacerbate existing unfairness in society. Achieving fairness, accountability, and transparency in research, algorithm design, and decision-making in AI-based transport systems will increase public acceptance and thus the benefits of ITS.

Traffic prediction and optimization, and even traffic monitoring, rely on accurate prediction algorithms. The prediction quality can deteriorate due to changes in model boundary conditions, resulting in poor control performance. The predicted traffic situation can also be imprecise and additionally hamper vehicle routing, resulting in increased traffic congestion. The risk of such errors can be reduced by implementing a warning system that notifies operators of potential decreases in prediction quality. Such a system uses indicators that assess the stability of components in the traffic prediction process and the quality of traffic predictions. The indicators are based on historical data and estimate the risk of unstable model boundary conditions and low prediction quality; they do not require assumptions about the underlying prediction algorithms.

5.1. Privacy-Preserving Analytics in Transportation

The ever-increasing volume of potential data sources raises significant privacy concerns. Privacy definitions change according to the context and may also include free will, autonomy, self-esteem, and dignity. The absence of appropriate privacy-preserving policies might lead to a decrease in public trust in Intelligent Transportation Systems (ITS), which jeopardizes the successful

deployment of ITS services. Privacy-preserving information sharing, such as systems that cope with the real-time sharing of private information without exposing sensitive attributes, should be enabled.

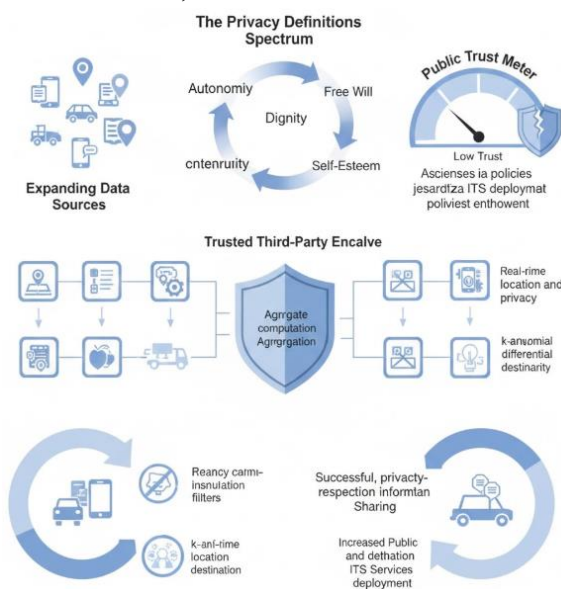


Fig 3: Preserving Public Trust in Intelligent Transportation: A Framework for Privacy-Centric Data Governance and Anonymized Information Sharing

Transport authorities or agencies should be aware of the critical privacy issues and dangers that may arise from collecting and publishing travel information. Existing privacy-preserving databases can either identify the information or allow the identification of information for the various shared groups.

Privacy is a critical consideration in telecommunication and computer networks. Various privacy-preserving techniques have been proposed in the context of transportation systems that allow a third-party trustworthy entity to perform some aggregate computing requests without knowing the individual private data. The entities have little or no knowledge of the systems. They are not allowed to misuse the user's profile. Transportation data are likely to be

sensitive data that can reveal when and where individuals want to move. Users are likely to prefer not to reveal their travel locations and destination information as required privacy needs to be retained.

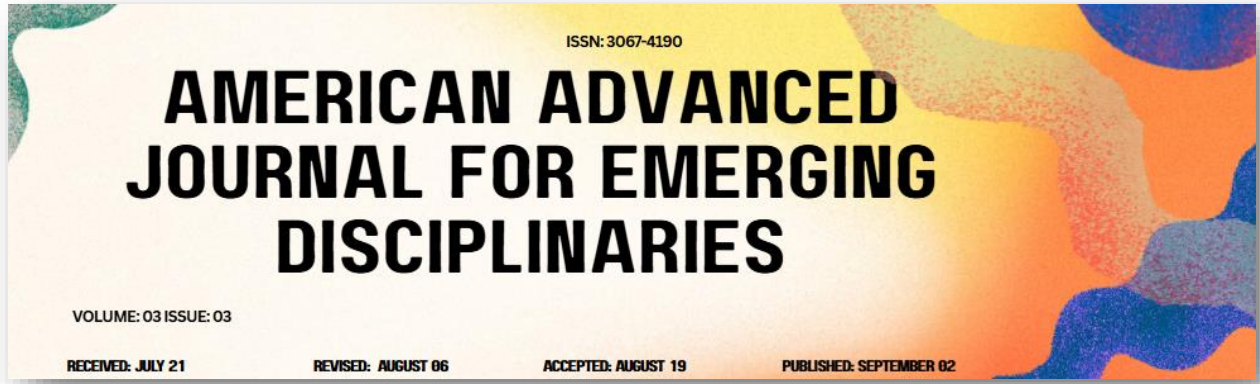
5.2. Bias, Fairness, and Accountability in ITS AI

Many AI methods are not interpretable, which raises concerns for data-driven decision-making. Historically, Intelligent Transportation Systems (ITS) have been a sensitive area for civil rights organizations, given their historical associations with mass surveillance. There are ongoing concerns regarding the balance between safety and civil liberties—e.g., monitoring of traffic for congestion versus monitoring civil rights protestors. When applied to data collection, prediction models, and deployment of autonomous vehicles, the use of AI in ITS raises fairness concerns such as:

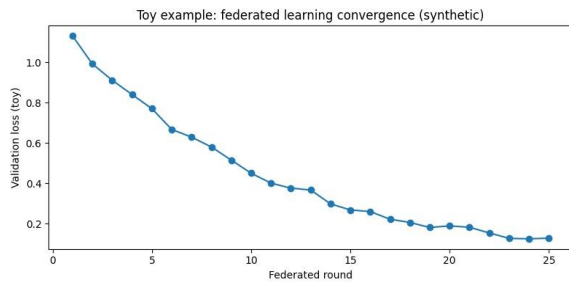
- * Technical biases in the training datasets used for supervised predictive modeling that arise even in the absence of malicious intent. For example, facial recognition technology has been found to be less effective for individuals with darker complexions.
- * Algorithmic bias that may emerge when predictive modeling results are utilized for decision-making—for example, placed police patrol locations causing increased arrest rates at those locations.

Civil rights organizations have raised concerns that anti-traffic deaths proposals could serve as cover for proposals that expand surveillance in communities of color. To mitigate the effects of bias and lack of fairness in ITS, accountability principles should guide the development of scanning and acquisition technologies. Appropriate scanning and acquisition design should include an emphasis on racial data and bias when assessing model performance. Additional techniques such as adversarial modeling may help establish fairness for groups that could be affected by bias in society.

6. Deployment and Operationalization



Deployment of data-centric solutions in Intelligent Transportation Systems (ITS) typically follows the general process of machine learning model deployment and operationalization. This encompasses the steps of model inference, hyper-parameter tuning, retraining, recommender systems setup, configuration dashboard setup, monitoring, logging, and reporting, among others. Specific aspects of these processes, such as managing the data required for inference, supporting hyper-parameter tuning, and providing ML lifecycle management, are thus common to other domains. Some specialized techniques have also been developed, such as monitoring for concept drift in fast-changing environments. However, implementation can vary based on the complexity of maintenance requirements and the demand for providing valuable predictions constantly. The open-source MLOps solution ZenML supports both serving and batch jobs and is tailored to the fast-changing world of Intelligent Transportation Systems, enabling daily retraining of models using MLOps principles. Model lifecycle management and MLOps platforms support all the steps and operations mentioned above, from model development to maintenance and production use. Such infrastructure is essential in domains like Intelligent Transportation Systems, where maintaining highly accurate and relevant predictions is key to providing business value. A proper model lifecycle management and MLOps setup can decrease the overhead of model development and tuning while extending the personalisation, customisation, relevance, and utility of predictions, ultimately improving business outcomes.



Equation 3) Congestion modeling from speed/flow (practical ITS metric)

The emphasizes congestion prediction and hotspot forecasting .

A common congestion index uses free-flow speed v^{ff} and observed speed v_t :

Step 1: Define observed and baseline speeds

- v^{ff} : typical free-flow speed for a segment
- v_t : observed speed at time t

Step 2: Define congestion index

$$CI_t = \frac{v^{ff}}{v_t}$$

- If $CI_t \approx 1$: uncongested
- If $CI_t \gg 1$: congested

Step 3: Congestion classification (thresholding)

$$\text{Congested}_t = \begin{cases} 1, & CI_t \geq \tau \\ 0, & CI_t < \tau \end{cases}$$

6.1. Model Lifecycle Management and MLOps for ITS
Operationalizing ML and DL models—i.e., serving them in production while monitoring performance, retraining them when necessary, and so on—requires a practice referred to as model lifecycle management. It implements management principles and concepts, such as continuous integration (CI), continuous delivery (CD), and infrastructure as code (IaC), tailored to the peculiarities of ML systems. A specialized field called MLOps expands DevOps principles and practices to ML-powered solutions. Learning- and AI-optimized iterations of MLOps have also been proposed. While implementation can be more straightforward than for regular software, specificities of the techniques involved and the associated security, privacy, and ethical aspects warrant a dedicated approach. Generally, models are validated and tested against earlier withheld data before being deployed in production. New and/or faster data—available in streaming architectures supporting demand prediction and congestion mitigation—

are suited for online learning. The most complex models have better accuracy but predict slower. Testing is needed to determine how much they pick (if at all) in real-time setups. Serving parallel branches trained at different cadences or administratively disabled is an alternative.

7. Conclusion

The successful development and operationalization of efficient, adaptive, and trustworthy intelligent transportation systems (ITS) relies heavily on well-architected data environments and AI. This paper provides an overview of foundational, architectural, methodological, ethical, and deployment-related aspects of AI-powered data engineering for ITS. The focus is on the design and implementation of reusable and shareable components and capabilities that support rapid innovation cycles and enable users to employ advanced AI techniques—predictive modeling, computer vision, sensor fusion, and privacy-preserving analytics—on real data and generate repeatable and actionable insights.

Promising directions for future research include the creation of new frameworks, theory, and algorithms for distributed data-hub architectures; the development of efficient incremental-learning and online methods for information extraction from images and videos; the integration of predictive-demand models with traffic-assignment models that simulate the effect of different transit-supply configurations on congestion; and the identification of desirable properties of ethically aligned ITS systems where model predictions might affect real-world decisions.

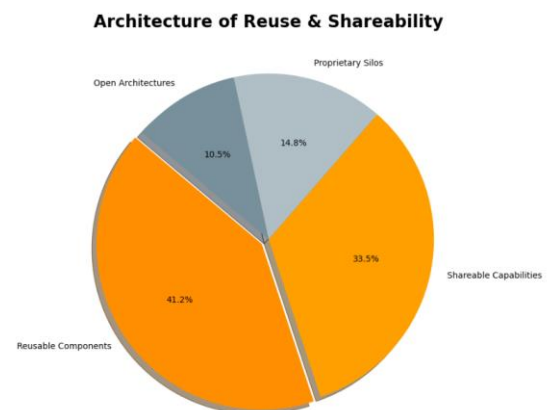
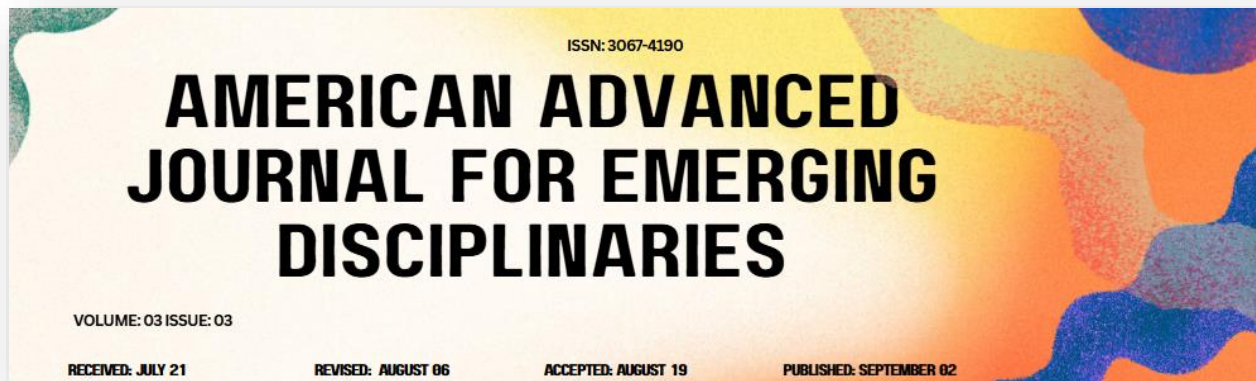


Fig 4: Architecture of Reuse & Shareability

7.1. Final Thoughts and Future Directions

The intelligent transportation system (ITS) of a smart city can be seen as a distributed cyberphysical system that operates constantly in a dynamic environment. A particular feature of this environment is the presence of extreme events, such as traffic jams due to accident or severe weather conditions, terrorist attacks, and so forth. Therefore, the ITS of a smart city needs to be continuously optimized and updated in real time—as every component of any cyberphysical system needs to be monitored. Collector sensors, computer vision, and sensor fusion techniques, among many others, are able to track and collect data about the larger traffic events in real time.

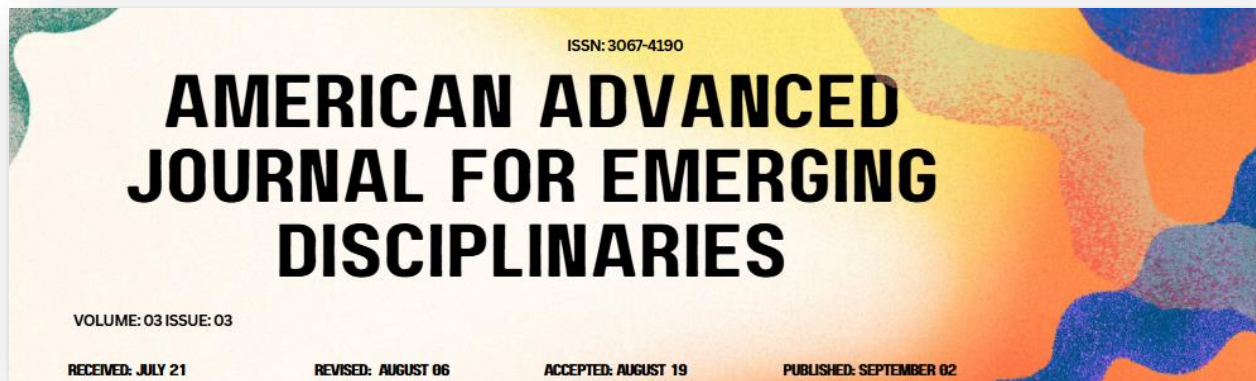
Transport-demand and congestion-prediction models need to be developed, trained, and updated with every regulation change, detective-vessel addition, and model-enhancement activity. In smart-cities working for environmental sustainability, the modeling process extends to environmental aspects that in the past were never included in an ITS-analysis model. All these activities need to be integrated into a single smart-city data engineering hub, able to fully support the development, validation, and continuous updating of any transport-optimization model. Additionally, because ITS data come from different sources and agents, enter the smart-city data engineering hub through different channels, and are processed for different objectives, the final architecture is structured as a data-



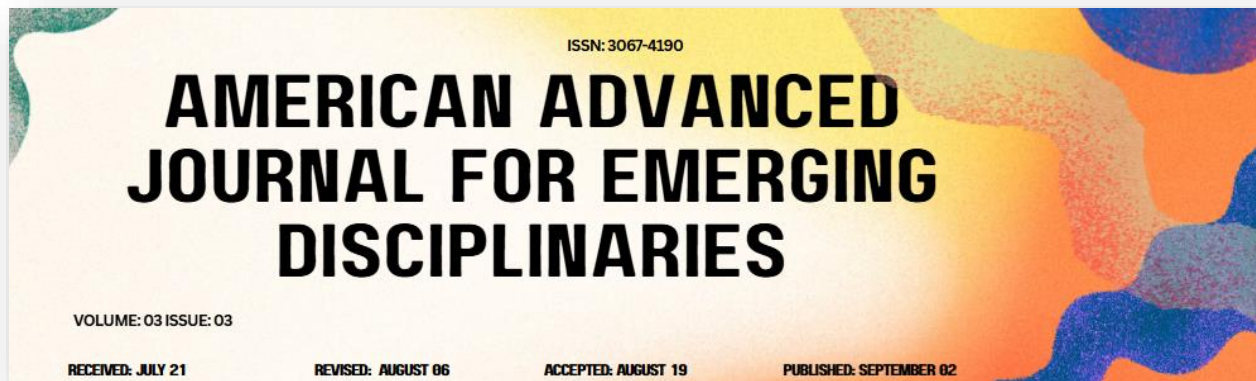
federation hub allowing federated queries on geographically distributed data sources. The architecture is smart in the sense that whenever data for a specific analysis are not locally available, the federated query automatically locates the data source and executes the query.

8. References

- [1] Abdel-Aty, M., Cai, Q., & Yue, L. (2023). Deep learning applications for road safety analysis and crash prediction: A review. *Accident Analysis & Prevention*, 188, 107090.
- [2] Abdulkareem, K. H., Mohammed, M. A., Salim, A., Arif, M., Geman, O., Gupta, D., & Khanna, A. (2021). Real-time big data analytics for intelligent transportation systems: A review. *IEEE Access*, 9, 10155–10178.
- [3] Alghamdi, A., AlOtaibi, R., Alzahrani, A., & Alsubhi, K. (2023). Edge intelligence for connected autonomous vehicles: Architecture, challenges, and future directions. *Future Generation Computer Systems*, 145, 102–118.
- [4] Al-Turjman, F., Ever, E., & Zahmatkesh, H. (2024). AI-driven vehicular edge computing for intelligent transportation: Concepts, taxonomy, and open research problems. *IEEE Communications Surveys & Tutorials*, 26(1), 1–34.
- [5] Anwar, S., Qin, J., & Khan, S. (2023). Vision-based traffic surveillance: Deep learning methods, challenges, and applications. *IEEE Transactions on Intelligent Transportation Systems*, 24(6), 5801–5821.
- [6] Badue, C., Guidolini, R., Carneiro, R. V., et al. (2022). Self-driving cars: A survey. *Expert Systems with Applications*, 165, 113816.
- [7] Bai, L., Yao, L., Li, C., Wang, X., & Wang, C. (2023). Adaptive spatiotemporal graph neural networks for traffic flow forecasting: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(7), 7011–7030.
- [8] Barua, S., & Zou, J. (2024). Federated learning for intelligent transportation systems: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, 25(2), 1320–1344.
- [9] Bhatia, M., Mishra, S., & Sangaiah, A. K. (2023). Explainable AI for intelligent transportation systems: A systematic review. *Information Fusion*, 94, 1–20.
- [10] Bojarski, M., Del Testa, D., Dworakowski, D., et al. (2016). End to end learning for self-driving cars. *arXiv*, 1–9.
- [11] Cao, Z., Yao, Y., Xu, Z., & Wang, T. (2024). Large-scale trajectory data mining for urban mobility analytics: A review. *ACM Computing Surveys*, 56(6), 1–39.
- [12] Chen, C., Liu, Y., Wan, J., Luan, J., & Cui, Q. (2019). Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 20(10), 3726–3737.
- [13] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171–209.



- [14] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
- [15] Das, S., & Kwasinski, A. (2023). Data engineering pipelines for real-time smart mobility: Design patterns and deployment challenges. IEEE Access, 11, 77901–77920.
- [16] Doshi, P., Ranjan, R., & Singh, A. (2022). Edge intelligence for smart transportation: A survey. Future Generation Computer Systems, 128, 390–410.
- [17] Fang, Z., Zhang, Y., & Zhang, J. (2024). Privacy-preserving learning in connected transportation systems: Trends and open challenges. IEEE Network, 38(1), 78–86.
- [18] Gao, J., Li, Y., & Zhang, X. (2023). Multi-source data fusion for urban traffic state estimation: A review. Transportation Research Part C: Emerging Technologies, 148, 104014.
- [19] Guo, J., Huang, W., Williams, B. M., & Wei, Y. (2019). Deep learning for traffic prediction: A review. IEEE Transactions on Intelligent Transportation Systems, 20(12), 4550–4565.
- [20] Han, K., Liu, J., Luo, J., & Wang, C. (2024). Real-time intelligent traffic signal control using deep reinforcement learning: A survey. IEEE Transactions on Intelligent Transportation Systems, 25(4), 3122–3143.
- [21] He, Z., Zhao, J., & Zhang, L. (2023). MLOps for edge AI in intelligent transportation: Requirements, architectures, and deployment strategies. ACM Transactions on Internet Technology, 23(4), 1–27.
- [22] Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends in Computer Graphics and Vision, 12(1–3), 1–308.
- [23] Jiang, W., & Zhang, J. (2023). Graph neural networks for traffic forecasting: A comprehensive review. IEEE Transactions on Intelligent Transportation Systems, 24(8), 8201–8223.
- [24] Khorasani, A., Gupta, V., & Kannan, R. (2024). Secure data engineering for intelligent transportation: Threat models, defenses, and validation. Computers & Security, 136, 103470.
- [25] Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A distributed messaging system for log processing. Proceedings of the NetDB Workshop, 1–7.
- [26] Li, X., Ma, Y., Zhang, Z., & Wang, J. (2023). Streaming analytics for connected vehicle systems: Architectures and performance evaluation. IEEE Internet of Things Journal, 10(8), 6900–6914.
- [27] Liu, Y., Zheng, H., Feng, X., & Chen, Z. (2020). Short-term traffic flow prediction with graph neural networks: A review. IEEE Transactions on Intelligent Transportation Systems, 21(5), 1–14.
- [28] Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). Traffic flow prediction with



big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 16(2), 865–873.

[29] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

[30] NIST. (2023). Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology.

[31] NIST. (2020). Security and privacy controls for information systems and organizations. NIST Special Publication 800-53 (Rev. 5).

[32] Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32, 8024–8035.

[33] Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, Fog et al.: A survey and analysis of security threats and challenges. Future Generation Computer Systems, 78, 680–698.

[34] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

[35] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646.

[36] Wang, X., Li, Y., & Huang, C. (2024). Trustworthy AI for intelligent transportation systems: Fairness, robustness, and accountability. IEEE Transactions on Intelligent Transportation Systems, 25(1), 110–128.

[37] Xu, X., Weber, I., & Staples, M. (2019). Architecture for blockchain applications. Springer.

[38] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 1–7.