Evaluating Attribute Inference Risks in Urban Care Taxi Arrival Time Prediction Models Using Geospatial Data

Yuya Takeuchi[1], Haruki Yonekura[2,3], Kyosuke Yamashita[2], Hirozumi Yamaguchi[2,3]
[1]Toshiba Corporation Corporate Laboratory, Japan
[2]The University of Osaka, Japan
[3]RIKEN Center for Computational Science, Japan
yuya.takeuchi.x89@mail.toshiba, {h-yonekura, yamashita, h-yamagu}@ist.osaka-u.ac.jp

ABSTRACT

In advanced countries, care services for the elderly are becoming essential components of urban infrastructure. Machine learning models that use location information as training input are under consideration for optimizing care taxi dispatch, as seen in programs such as Adult Day Care Transportation in the US and Total Mobility in New Zealand, for example. However, these models often rely on privacy-sensitive data, making intercity data sharing difficult. As a result, approaches such as federated learning have been proposed to protect sensitive information while enabling cross-city collaboration. At the same time, applying an attribute inference attack to a shared model can lead to the leakage of users' personal information. In this study, we build a boarding and alighting time predictor using a dataset of real-world care taxi traces and then apply an attribute inference attack to evaluate how accurately an attacker can infer each user's walking disability in our dataset. To mitigate these privacy risks, we analyze the effects of data processing techniques on attack vulnerability; notably, when class imbalance was handled via SMOTE data augmentation, the attack's accuracy increased from 61.3% to 73.0%. Our findings offer guidelines for designing privacy-preserving machine-learning systems in the context of eldercare transportation.

CCS CONCEPTS

• **Networks** → *Location based services*; • **Security and privacy** → *Privacy Protection*;

KEYWORDS

Privacy-preserving, Attribute Inference Attack, Deep Learning, Care Taxi

ACM Reference Format:
Yuya Takeuchi, Haruki Yonekura, Kyosuke Yamashita, Hirozumi Yamaguchi. 2025. Evaluating Attribute Inference Risks in Urban Care Taxi Arrival Time Prediction Models Using Geospatial Data. In *the 14th International Workshop on Urban Computing (UrbComp'25), held in conjunction with the 31st ACM SIGKDD 2025, August 3rd, 2025, Toronto, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

UrbComp' 25, August 3rd, 2025, Toronto, Canada

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX…$xx.xx
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent decades, population aging has emerged as a defining challenge for many nations, with profound implications for health care systems and social infrastructure. This demographic shift has precipitated an unprecedented surge in demand for long-term care services, even as the supply of qualified care workers dwindles owing to a contracting labor force. Globally, the World Health Organization (WHO) reports that roughly 142 million older people are currently unable to meet their basic needs independently and that two out of three older adults will require some form of long-term care during their lifetime[21]. At the same time, WHO projects a shortfall of approximately 11 million health and care workers by 2030[20], underscoring the severity of the workforce crisis. Against this backdrop, advanced economies are exploring data-driven solutions—such as machine-learning models that leverage real-time location information to optimize care transportation—but these predictive systems often depend on privacy-sensitive inputs, raising critical concerns about data sharing and model vulnerability. In particular, attribute-inference attacks may allow malicious actors to infer individuals' residential addresses or daily routines from ostensibly benign service logs. Thus, there exists an urgent need to develop efficient, privacy-preserving frameworks that can accommodate regional heterogeneity—ranging from densely populated urban centers to depopulated rural areas—while safeguarding sensitive user information and ensuring equitable access to high-quality care.

As a means of knowledge sharing, data sharing among long-term care providers is the most direct approach. However, long-term care data contains highly sensitive personal information, including users' health conditions, medical history, cognitive function, ADL (Activities of Daily Living) assessments, and family structures. This information transcends mere privacy concerns; its leakage could potentially lead to discrimination, prejudice, and even financial or psychological harm. As a countermeasure, applying anonymization processing to the data is considered. However, it is known that statistically demonstrated methods such as k-anonymity and l-diversity can significantly degrade the performance of machine learning models utilizing such data [5]. Especially in the long-term care domain, even a slight decrease in prediction accuracy can directly impact user safety and care quality, making it a challenge to balance anonymization and utility.

In recent years, sharing only machine learning models independently trained by each provider has also been considered. This may appear secure, as no raw data leaves the institution. However, this approach remains vulnerable to Membership Inference Attacks (MIA)[17], a class of privacy attacks that aim to determine

whether a particular individual's record was included in the training dataset. These attacks exploit the fact that machine learning models often behave differently on data they have seen during training versus unseen data, particularly when models are overparameterized, trained on small datasets, or lack proper regularization. Such behavioral discrepancies can be leveraged to infer membership with high confidence, thereby exposing sensitive personal information even without direct access to the original dataset. MIA poses a particularly acute threat in the long-term care domain, where the inclusion of a single record, such as one indicating a rare medical condition or service usage pattern, could reveal highly personal attributes of the individual concerned. Moreover, even aggregated or partially anonymized models can be susceptible if adversaries have auxiliary knowledge or can systematically probe the model through black box queries. These risks are amplified when models are deployed across institutions, shared in collaborative settings, or exposed through APIs. While techniques such as differential privacy or regularization strategies can mitigate some of these vulnerabilities, they often come at the cost of reduced model performance[5]. This trade-off is especially critical in healthcare applications, where predictive accuracy directly affects service quality and user safety. Thus, mitigating the risks associated with MIA requires careful calibration between model utility and privacy protection, along with an understanding of the attacker's capabilities and available background knowledge.

As an alternative to simple model-sharing, Federated Learning (FL) has been proposed to enable collaborative training without exchanging raw data [7]. In the FL paradigm, each provider retains its local dataset and computes updates to model parameters that are then aggregated centrally, thereby reducing direct exposure of sensitive records. Nevertheless, recent work has shown that federated updates can still leak private information through gradient inversion or statistical inference techniques, and that challenges such as client heterogeneity and communication overhead complicate practical deployment [19]. Consequently, even FL requires careful protocol design and threat modeling to ensure that privacy guarantees are upheld in real-world settings[4, 12].

MIA assumes that one possesses all the constituent elements of a single record in the dataset without excess or deficiency. For example, training a model such that the combination of input attributes becomes large could be a partial solution to this vulnerability. Specifically, by increasing the number of data features or using data augmentation techniques, it is possible to increase the difficulty for an attacker to infer the exact record. Building on this perspective, we further consider the risk from Attribute Inference Attacks, in which adversaries attempt to reconstruct unknown features of a user from observed data or model outputs, even without knowing whether the user was part of the training set [25]. These attacks represent a complementary threat model, and together with MIA, they underline the multifaceted risks of sharing models trained on privacy-sensitive data.

In this study, based on care taxi data that include location information provided by a partnered long-term care provider, we discuss the extent to which users' data may be vulnerable. Specifically, we examine the possibility of inferring users' walking disability from care taxi location traces and a prediction model trained on those traces. *Our contributions are summarized as follows: Firstly, to the best of our knowledge, this is the first work to apply an attribute-inference attack to a machine learning model in a taxi-related application. In particular, we conducted an attribute-inference attack on decision trees and evaluated their vulnerabilities by integrating spatial data from urban environments with actual users' sensitive personal information to assess the attack's effectiveness in a realistic scenario. Secondly, we investigated how insights related to data augmentation and class imbalance in the training dataset influence the success of attribute-inference attacks. Thirdly, we partnered with a long-term care provider and performed data collection over a period exceeding ten months.*

2 RELATED WORK

2.1 Privacy Preservation by Adding Noise

Spatio-temporal data, which includes user location information, is crucial for data-driven applications. However, data collection is costly, and it is necessary to consider the leakage of users' private information. Therefore, much existing research has addressed the anonymization and hiding of trajectory data by utilizing concepts of differential privacy and masking processes.

The approaches in [18, 23] propose protecting user privacy by adding dummy location data, not included in the original dataset, to the dataset. These methods are effective if an attacker cannot distinguish the trajectory data as fake. However, there is a risk that fake trajectory data can become unrealistic and be identified by an attacker [8]. Differential privacy is a method that can mathematically guarantee privacy protection up to a certain level by adding appropriate noise to the statistics of a dataset. Privacy protection methods using this technology have also been proposed [11]. For example, CNoise and SDD are examples of the application of differential privacy to trajectory data [11]. While these methods can guarantee a certain level of privacy protection by adding noise to individual trajectories, it is known that the utility of the data in machine learning models and spatio-temporal analysis is reduced.

2.2 Privacy Preservation by Synthetic Data Generation

Methods that achieve privacy protection by replacing the data in the original dataset with synthetic data have also been proposed [1, 16]. Ozeki et.al. [1] propose a method to synthesize the entire dataset's data to achieve k-anonymity in trajectory data, thereby ensuring the privacy of trajectories. K-anonymity is a property of a dataset where there are k or more users with the same characteristics, so even if an attacker tries to identify a user based on specific characteristics, the number of candidates can only be narrowed down to k. As a method to achieve k-anonymity for trajectory data, [1] proposes a method to add uncertainty to location information. However, such simple k-anonymity-based methods are computationally expensive and do not consider the naturalness of the generated trajectories. Therefore, the generated data often becomes unnatural, leading to issues such as reduced utility and the unnaturalness itself increasing the risk of privacy leakage [3]. Furthermore, LSTM-trajGAN [16] is a method that trains a generative adversarial network based on LSTM and replaces the original dataset with the generated synthetic data. While these methods can generate realistic trajectory data, they

do not guarantee the privacy protection or utility of the generated data.

2.3 Attacks on Machine Learning Models

A Membership Inference Attack (MIA) is a method by which an attacker infers whether a specific data point was included in the training data of a machine learning model. Shokri et al. (2017) proposed a framework for MIA assuming black-box access, demonstrating how an attacker can train a model to infer inclusion in training data using the target model's inputs and outputs [17]. Hu et al. (2021) conducted a comprehensive survey on MIA, providing a classification, evaluation, and comparison of various attack techniques and defense strategies [9].

An Attribute Inference Attack (AIA) is a method by which an attacker infers unknown attributes (e.g., age, gender) for a partially known data point. Zhao et al. (2021) examined the feasibility of attribute inference attacks and showed that an attribute inference attack might not succeed even if a membership inference attack is successful. They also demonstrated that approximate attribute inference is possible [25]. Mehnaz et al. (2022) proposed a method to infer unknown attributes with only black-box access, achieving higher accuracy than existing methods [15].

3 ATTRIBUTE INFERENCE ATTACKS

Attribute inference attacks (AIA) are a type of attack against trained machine learning models in which an adversary leverages access to the model to infer individuals' sensitive attributes [25]. Such attacks exploit statistical correlations between observable (non-sensitive) features and unobservable (sensitive) features within the data to reveal sensitive information, even when those attributes are not explicitly included in the model's outputs.

In this study, the data used to train the machine learning models inherently contain sensitive personal information—such as users' age, gender, and mobility capabilities—that, if linked to individuals, may constitute a violation of privacy. Accordingly, we regard any action by a third party (adversary) aimed at identifying personal information from a publicly released machine learning model trained on such data as an attack.

We assume that the adversary possesses, at least partially, a list of users who utilize a facility, or alternatively, a list of households in which elderly individuals reside, and seeks to estimate their gender or mobility capabilities from the machine learning model.

3.1 Definitions

We formalize the problem as follows:

Feature Vector: $\mathbf{x} = (\mathbf{x}_{\text{pub}}, x_{\text{priv}}) \in \mathcal{X} \times \mathcal{A}$, where $\mathbf{x}_{\text{pub}}$ denotes the public (non-sensitive) attributes and x_{priv} denotes the private (sensitive) attribute.

Target Model: $f : \mathcal{X} \to \mathcal{Y}$ is a function trained on a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ that maps input features to predictions

Adversary's Goal: Given access to $\mathbf{x}_{\text{pub}}$ and the model f, the adversary aims to infer x_{priv}.

The adversary models the conditional probability distribution:

$$P\big(x_{\text{priv}} \mid \mathbf{x}_{\text{pub}}, f(\mathbf{x}_{\text{pub}}, x_{\text{priv}})\big). \tag{1}$$

The optimal inference strategy is to choose the value that maximizes this conditional probability:

$$\hat{x}_{\text{priv}} = \arg\max_{a \in \mathcal{A}} P\big(x_{\text{priv}} = a \mid \mathbf{x}_{\text{pub}}, f(\mathbf{x}_{\text{pub}}, a)\big). \tag{2}$$

This approach assumes that the adversary can simulate or approximate the model's behavior for different values of x_{priv}.

3.2 Attack Methods

Several strategies can be employed for attribute inference:

3.2.1 Model Inversion Attack. These attacks reconstruct sensitive attributes by exploiting the model's output confidences. Given $\mathbf{x}_{\text{pub}}$, the adversary searches for the x_{priv} that maximizes the model's output confidence:

$$\hat{x}_{\text{priv}} = \arg\max_{a \in \mathcal{A}} f\big(\mathbf{x}_{\text{pub}}, a\big).$$

This method is particularly effective when the model's outputs are highly sensitive to changes in x_{priv}.

3.2.2 Shadow Model Training. The adversary trains a surrogate model f' on a dataset that approximates the distribution of $\mathcal{D}$. The shadow model approximates the behavior of f, enabling the adversary to learn the mapping from $\mathbf{x}_{\text{pub}}$ to x_{priv}.

3.3 Theoretical Considerations

The success of attribute inference attacks is closely related to the generalization properties of the model. In particular, models with poor generalization (i.e., overfitted models) tend to memorize training data and are thus more vulnerable to such attacks [22]. A formal relationship between generalization error and privacy leakage has been established, suggesting that minimizing overfitting can mitigate the risk of attribute inference.

3.4 Practical Implications

Attribute inference attacks have been demonstrated in various domains. In social networks, sensitive information such as political affiliation or sexual orientation can be inferred from users' public posts and "likes." In the medical field, research has shown that specific health conditions can be predicted even from ostensibly anonymized health records[10]. Moreover, in recommender systems, privacy guarantees are provided for both users' sensitive attributes and the model optimization process by combining the information perturbation mechanism of differential privacy with the recommendation capabilities of graph convolutional networks[24]. These examples underscore the serious potential privacy risks and recent attentions posed by attribute inference attacks, even when sensitive attributes are not explicitly disclosed.

4 ASSUMPTIONS AND ATTACK SCENARIO

In this section, we assume that a care provider publishes a classification-based machine learning model, trained on proprietary data, which estimates the time required for user pick-up and drop-off at one-minute granularity. We further assume that an attacker uses this published model to obtain additional personal information. Figure 1 illustrates the sequence of knowledge sharing and the subsequent attack assumed in this study.

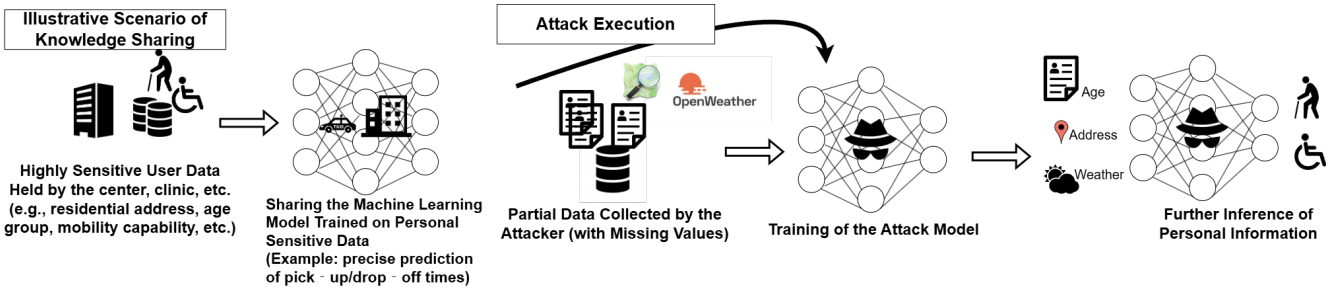


Figure 1: Overview of the Assumed Knowledge Sharing and Attack Flow.

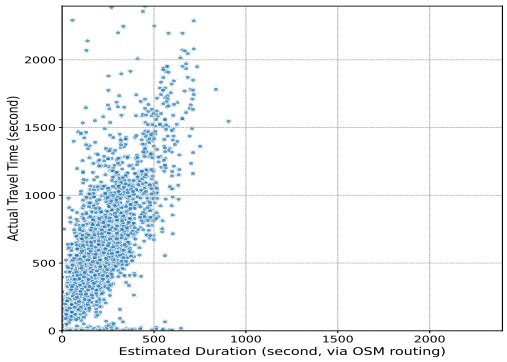


Figure 2: Distribution of Estimated Pick-up/Drop-off Times and Actual Travel Times.

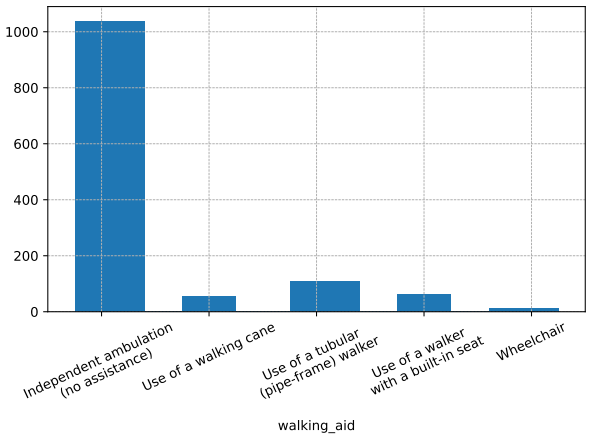


Figure 3: Population Distribution of Walking Aid Usage Categories.

4.1 Pick-up/Drop-off Time Prediction Model

A local facility trains a model to estimate, based on its own data, the time required for a pick-up vehicle to travel from the facility to the user's home or a designated meeting point. This task inherently departs from travel-time estimations computed using platforms such as OpenStreetMap[1], as shown in Figure 2. This discrepancy arises because intrinsic factors—such as the user's preparation behavior and walking ability—often delay actual arrival times beyond the platform's estimates.

As input features, the model receives:

- Departure time of the pick-up taxi,
- Latitude and longitude of the facility,
- Latitude and longitude of the user,
- Gender,
- Age,
- Walking ability,
- Temperature, Wind speed, and Humidity of the day,
- Day of the week,
- Estimated travel time provided by a map application.

Here, walking ability is divided into five levels: "Independent ambulation(no assistance)," "Use of a walking cane," "Use of a tubular(pipe-frame) walker," "Use of a walker with a built-in seat," and "Wheelchair." We assume that walking ability is not directly accessible to the attacker, whereas the remaining attributes, compared to walking ability, are considered obtainable by the attacker. As the machine learning prediction model, we implement a DecisionTree. We configured the decision tree to optimize splits by maximizing Shannon entropy, thereby choosing splits with the highest information gain. We cap the tree depth at sixteen levels and require that any internal node contain at least two samples before splitting, as well as that each terminal leaf contain at least two samples, which ensures that no leaf represents a single instance and reduces sensitivity to noise.

4.2 Attack Model

The attacker, given access to input–output pairs of the published prediction model, seeks to infer the user's walking ability. Figure 3 shows the population distribution across the walking-ability categories. For the attacker, inferring the minority (and thus more sensitive) classes—such as those requiring a walker or cane—carries significant value. In our assumed attack scenario, it is crucial to increase the confidence of predictions indicating that the user uses a walker or cane; hence, reducing False Positive errors becomes essential to achieving the attacker's objective.

To address this, the loss function incorporates, in addition to the standard Focal Loss [13], an extra penalty when a user whose true label indicates no walking-aid requirement (class 0) is misclassified into any other class. First, the target label y (one-dimensional) is

[1] *OpenStreetMap*. https://www.openstreetmap.org, accessed April 15, 2025.

converted to a one-hot representation over C classes to align with the dimensionality of the network's output logits. If a one-hot vector is already provided, it is simply cast to a floating-point type.

For network logits $\mathbf{z}$, the softmax function is applied to compute the probability for each class:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)}. \tag{3}$$

For numerical stability and to prevent division by zero, these probabilities are clipped to the interval $[\epsilon = 10^{-6},\ 1-\epsilon]$. The probability corresponding to the correct class, denoted p_t, is then computed via

$$p_t \ = \ \sum_{i=1}^{C} p_i \, y_i. \tag{4}$$

While the conventional cross-entropy loss is defined as

$$\mathcal{L}_{CE}(p_t) \ = \ -\log(p_t), \tag{5}$$

the Focal Loss is expressed as

$$\mathcal{L}_{FL}(p_t) \ = \ -\alpha_t \, (1-p_t)^{\gamma} \, \log(p_t), \tag{6}$$

where the parameter $\gamma \geq 0$ (the focusing parameter) down-weights the loss contribution of "easy" samples, and α_t is a weighting factor that addresses class imbalance.

In our implementation, for samples whose true label is class 0 (no walking-aid requirement), if the predicted class (determined by the largest softmax output) is not 0, an additional penalty term $-\log(p_0)$ is multiplied, where p_0 is the predicted probability for class 0. We create a mask to extract class 0 samples and, for those misclassified, calculate this penalty. The penalty terms are averaged over the selected samples and then scaled by a constant λ. The overall loss is defined as the sum of the Focal Loss and this additional penalty:

$$\mathcal{L} = \mathcal{L}_{FL} + \lambda \, \mathcal{L}_{penalty}. \tag{7}$$

In this study, we set $\lambda = 8.0$.

This design is particularly effective in scenarios with pronounced class imbalance or in applications where misclassifying a specific class (class 0 in this example) has severe consequences. By augmenting the Focal Loss with an extra term that heightens sensitivity to critical misclassification cases, the model is encouraged to learn more discriminatively for high-risk classes, thereby improving overall classification performance.

We began with the ART library's default configuration for the AttributeInferenceBlackBox attack[2], which uses a simple neural network by default. Next, we made several customizations. First, we replaced the out-of-the-box classifier with a deeper MLP network consisting of four fully connected layers of sizes 512, 256, 128, and 64, followed by a final output layer matching the number of classes. Each hidden layer is followed by a ReLU activation, batch normalization, and dropout. Second, we replaced the original loss function with the one we mentioned in this section. Third, we applied a standard scaler to all input features and to the model's predicted values before feeding them into the attack network. Finally, we adjusted the training loop to run for up to one hundred epochs using the AdamW[14] optimizer with a learning rate of 0.0001, and we implemented early stopping with a patience of five epochs based on a minimum loss improvement threshold of 0.00005. We preserve the model state that achieves the best validation loss. These updates improve robustness when classes are imbalanced and increase the attacker model's confidence when inferring high-sensitivity classes.

[2]https://github.com/Trusted-AI/adversarial-robustness-toolbox

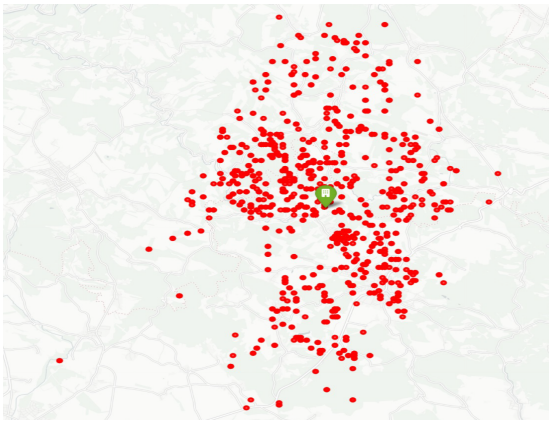


Figure 4: Addresses of the facility and users (the background map has been altered to an unrelated location to protect personal information).

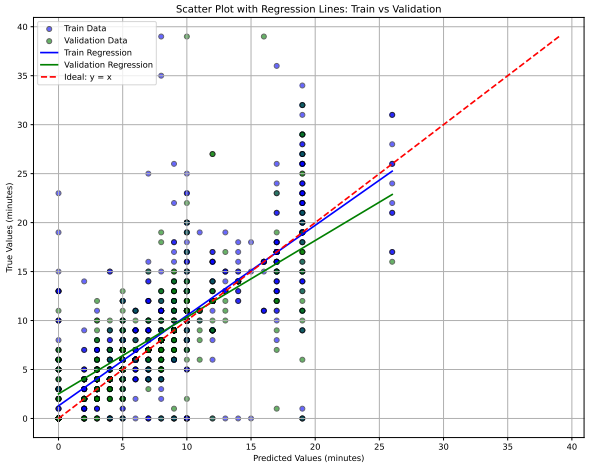
5 EVALUATION

5.1 Dataset

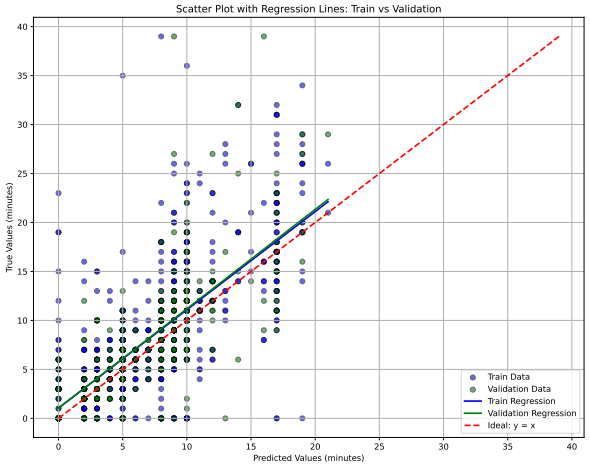
In this study, we validate our approach using real pick-up and drop-off data obtained from a specific care provider between October 1, 2023 and October 31, 2023, yielding 1,930 unique records. The relative locations of the facility and its users are illustrated in Figure 4.

5.2 Evaluation of Pickup Duration Prediction Model

Figure 5 shows, for each sample, the true one-minute–granularity travel time versus the predicted duration by the machine learning model. Blue points correspond to training-data samples, and green points correspond to test-data predictions. For the training data, the mean absolute prediction error was 2.251 minutes; for the test data, the mean absolute prediction error was 3.508 minutes. These results indicate that even the decision tree–based model is somewhat overfitting to the training data. Figure 5b shows the analogous scatter plot and regression lines obtained when the model is trained on data excluding walking aid information. In this case, the mean absolute prediction error rose to 2.853 minutes on the training set and to 3.094 minutes on the validation set. A direct comparison of these errors with those from Figure 5a suggests that including walking aid data does, in fact, improve predictive accuracy to some degree. In other words, the model's performance degrades slightly when walking ability features are omitted, which implies that there is a nontrivial relationship between walking ability and predicted pick-up/drop-off time.



(a) Evaluation of pick-up/drop-off time prediction accuracy by the model trained on data *including* walking aid information.



(b) Evaluation of pick-up/drop-off time prediction accuracy by the model trained on data *excluding* walking aid information.

Figure 5: Comparison of taxi arrival time prediction performance for models trained with or without walking aid data.

5.3 Evaluation of Privacy Risks

In this study, the attacker determines whether an individual's gait capability is impaired, and the attack's success rate is assessed by the proportion of correct determinations. In real-world data, there exists a continuum of impairment severity in walking ability; however, in this work, we disregard that fine-grained variation and focus our evaluation solely on inferring the binary presence or absence of impairment. As an assumption, half of the true data is used to train the black box attribute inference attack model, and the remaining half is used to evaluate attribute inference risk.

Figure 6 and Table 1 show the attribute inference attack result in a confusion matrix. First, as an overall trend, we observed that, due to the original class imbalance, the model's predictions were heavily biased toward class 0 despite the incorporation of a penalty term. Although we introduced Focal Loss and related techniques, this imbalance persisted for the following reasons. In the current task of estimating transportation pickup time, we formulated the problem as a classification problem for each minute. Even though this formulation increased the number of features, we found no sufficiently strong correlation among the walking aid indicator, the estimated pickup time, and the other features that an attribute inference attack could exploit. As a result, the model lacked the information necessary for reliable inference. This conclusion is also supported by the patterns observed in the UMAP visualization[2] of the data in Figure 7. Moreover, it is well known that decision tree-based methods are less prone to overfitting than complex neural network-based methods and that this property makes decision trees more resistant to attribute inference attacks.

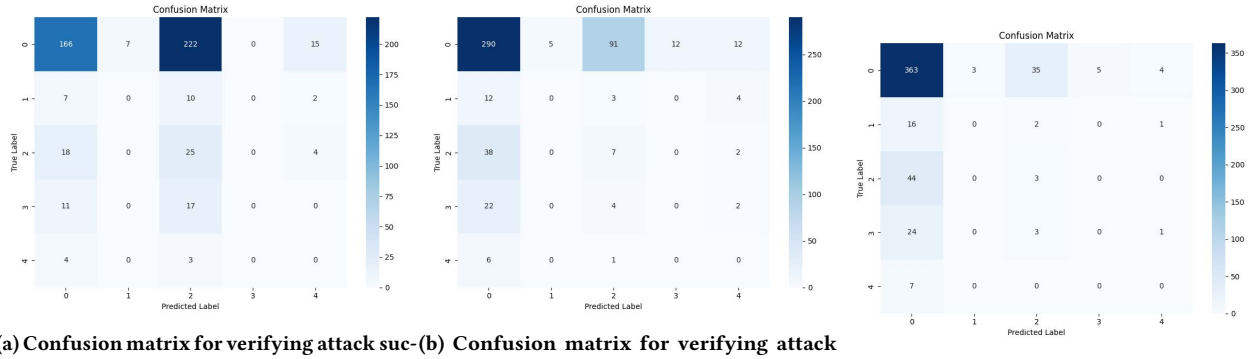
Table 1: Attack Success Performance.

Case	Precision	Recall	Accuracy	F1-score
Figure. 6a	0.2000	0.6040	0.4442	0.3005
Figure. 6b	0.1608	0.2277	0.6125	0.1886
Figure. 6c	0.1754	0.0990	0.7299	0.1265

We also examined scenarios in which the attacker applies data augmentation via SMOTE[6] during the model's training phase. In our dataset, an intriguing observation emerged: when the prediction model for transportation time (the target model) was trained in a manner that corrected for class imbalance by data augmentation, the attack's success rate increased in terms of both precision (from 0.1608 to 0.1754, a relative increase of approximately 9.1 %) and accuracy (from 0.6125 to 0.7299, a relative increase of approximately 19.2 %). This finding indicates that as the target model becomes more generalized via data augmentation, it is more susceptible to successful attacks. In other words, directly eliminating class imbalance through data augmentation can render the model more vulnerable to attribute-inference efforts.

6 CONCLUSION

In this study, we investigated whether users' sensitive information can be inferred from transportation data collected in collaboration with a home-care provider. We trained a decision-tree model to predict transportation duration and then applied attribute-inference attacks to evaluate its vulnerability. The experimental results indicated that, because of the original class imbalance, the attack's overall success rate remained low; however, we also found that a risk of data leakage persists. Moreover, we confirmed that data augmentation can, under certain circumstances, increase the model's susceptibility to such attacks. Future work will examine how different model architectures influence the attack success rate, explore the relationship between overfitting and vulnerability, and assess how varying levels of attacker knowledge affect attack efficacy. We will also evaluate potential defense mechanisms, particularly differential privacy techniques that add calibrated noise during training or in query responses to determine their effectiveness at mitigating attribute-inference risks in eldercare transportation models.



(a) Confusion matrix for verifying attack success when the taxi pickup-time prediction model (Decision Tree) is trained on the original data without SMOTE augmentation, and the attack model is also trained on data without SMOTE augmentation.

(b) Confusion matrix for verifying attack success when the taxi pickup-time prediction model (Decision Tree) is trained on original data without SMOTE augmentation, but the attack model is trained on SMOTE-augmented data.

(c) Confusion matrix for verifying attack success when the taxi pickup-time prediction model (Decision Tree) is trained on SMOTE-augmented data, and the attack model is also trained on SMOTE-augmented data.

Figure 6: Confusion matrices used to verify the success of the attribute inference attack under each scenario.

ACKNOWLEDGMENTS

This work was supported by JST, CREST Grant JPMJCR21M5, Japan, and JST BOOST, Japan Grant Number JPMJBS2402.

REFERENCES

[1] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2008. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In *2008 IEEE 24th International Conference on Data Engineering*. 376–385.

[2] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* 37, 1 (2019), 38–44.

[3] Jinyang Chen, Rangding Wang, Liangxu Liu, and Jiatao Song. 2011. Clustering of trajectories based on Hausdorff distance. In *2011 International Conference on Electronics, Communications and Control (ICECC)*. 1940–1944.

[4] Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. 2022. DisPFL: Towards Communication-Efficient Personalized Federated Learning via Decentralized Sparse Training. In *International Conference on Machine Learning*. PMLR, 4587–4604.

[5] Judith Sáinz-Pardo Díaz and Álvaro López García. 2023. Comparison of machine learning models applied on anonymized data with different techniques. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 618–623.

[6] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research* 61 (2018), 863–905.

[7] István Hegedűs, Gábor Danner, and Márk Jelasity. 2021. Decentralized learning works: An empirical comparison of gossip learning and federated learning. *J. Parallel and Distrib. Comput.* 148 (2021).

[8] D. Hemkumar, S. Ravichandra, and D.V.L.N. Somayajulu. 2020. Impact of prior knowledge on privacy leakage in trajectory data publishing. *Engineering Science and Technology, an International Journal* 23, 6 (2020), 1291–1300.

[9] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[10] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. 2017. Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*. 1561–1569.

[11] Kaifeng Jiang, Dongxu Shao, Stéphane Bressan, Thomas Kister, and Kian-Lee Tan. 2013. Publishing Trajectories with Differential Privacy Guarantees. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM)*. Association for Computing Machinery, New York, NY, USA, Article 12, 12 pages.

[12] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze (Eds.), Vol. 2. 429–450.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[14] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[15] Shagufta Mehnaz, Sayanton V Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino. 2022. Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models. In *31st USENIX Security Symposium (USENIX Security 22)*. 4579–4596.

[16] Ren Ozeki, Haruki Yonekura, Hamada Rizk, and Hirozumi Yamaguchi. 2023. Balancing privacy and utility of spatio-temporal data for taxi-demand prediction. In *2023 24th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 215–220.

[17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[18] Akiyoshi Suzuki, Mayu Iwata, Yuki Arase, Takahiro Hara, Xing Xie, and Shojiro Nishio. 2010. A User Location Anonymization Method for Location Based Services in a Real Environment. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10)*. Association for Computing Machinery, New York, NY, USA, 398–401.

[19] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2512–2520.

[20] World Health Organization. n.d.. Health workforce. https://www.who.int/health-topics/health-workforce. Accessed: May 15, 2025.

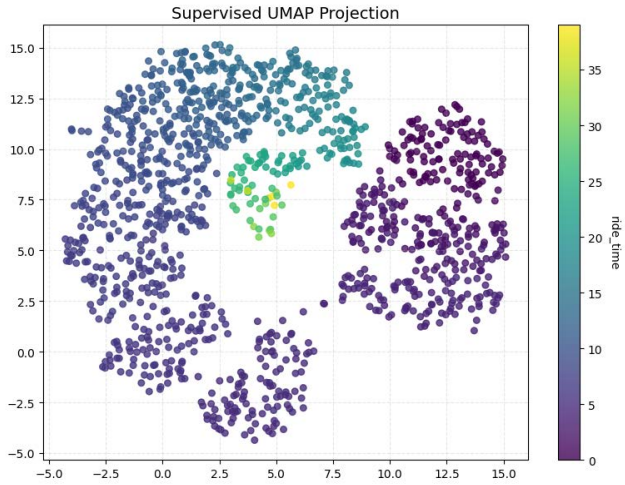
[21] World Health Organization. n.d.. Long-term care. https://www.who.int/europe/news-room/questions-and-answers/item/long-term-care. Accessed: May 15, 2025.

[22] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. 268–282.

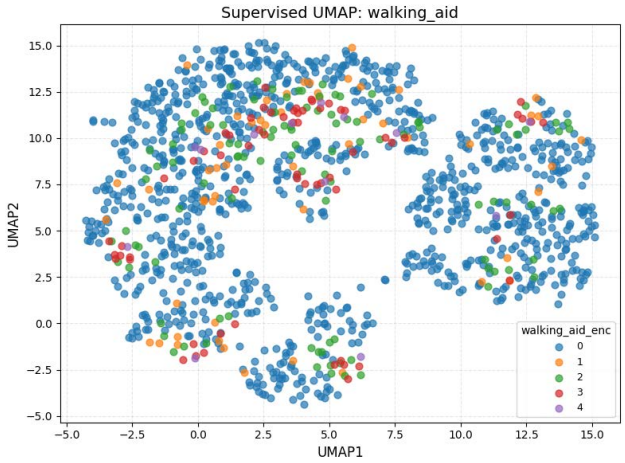
[23] Tun-Hao You, Wen-Chih Peng, and Wang-Chien Lee. 2007. Protecting Moving Trajectories with Dummies. In *2007 International Conference on Mobile Data Management*. 278–282.

[24] Shijie Zhang, Hongzhi Yin, Tong Chen, Zi Huang, Lizhen Cui, and Xiangliang Zhang. 2021. Graph embedding for recommendation against attribute inference attacks. In *Proceedings of the web conference 2021*. 3002–3014.

[25] Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson. 2021. On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models . In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE Computer Society, Los Alamitos, CA, USA, 232–251. https://doi.org/10.1109/EuroSP51992.2021.00025



(a) Supervised UMAP projection of feature vectors, with each point colored by the corresponding ride time in minutes. This visualization reveals how samples cluster according to transit duration.



(b) Supervised UMAP projection of feature vectors, with each point colored by walking aid category (0 = no aid, 1–4 = different types of aid). This plot illustrates how the presence and type of walking aid correlate with the learned embedding space.

Figure 7: UMAP[2] Visualizations.