https://doi.org/10.21872/2025IISE_6188

Enhancing Human-Robot Interaction through Ensemble Intention Recognition and Trajectory Tracking

Elnaz Alinezhad, Ali Kamali Mohammadzadeh, Sara Masoud

Department of Industrial and Systems Engineering, Wayne State University

Detroit, Michigan, USA

Abstract

Effective intention recognition and trajectory tracking are critical for enabling collaborative robots (cobots) to anticipate and support human actions in Human-Robot Interaction (HRI). This study investigates the application of ensemble deep learning to classify human intentions and track movement trajectories using data collected from Virtual Reality (VR) environments. VR provides a controlled, immersive setting for precise monitoring of human behavior, facilitating robust model training. We develop and evaluate ensemble models combining CNNs, LSTMs, and Transformers, leveraging their complementary strengths. While CNN and CNN-LSTM models achieved high accuracy, they exhibited limitations in distinguishing specific intentions under certain conditions. In contrast, the CNN-Transformer model demonstrated superior precision, recall, and F1-scores in intention classification and exhibited robust trajectory tracking. By integrating multiple architectures, the ensemble approach enhanced predictive performance, improving adaptability to complex human behaviors. These findings highlight the potential of ensemble learning in advancing real-time human intention understanding and motion prediction, fostering more intuitive and effective HRI. The proposed framework contributes to developing intelligent cobots capable of dynamically adapting to human actions, paving the way for safer and more efficient collaborative workspaces.

Keywords

Human Robot Interaction, Ensemble learning, Intention recognition, Trajectory tracking.

1. Introduction

Industry 5.0 represents the next evolution in manufacturing, prioritizing human-robot collaboration (HRI) to enhance safety, efficiency, and productivity. By integrating advanced technologies such as the Internet of Things (IoT), artificial intelligence (AI), digital twins, and extended reality (XR), it establishes intelligent hybrid workspaces where collaborative robots (cobots) operate alongside human workers. In these environments, cobots take on repetitive, hazardous, or physically demanding tasks, enabling humans to focus on cognitively complex activities that require creativity, problem-solving, and decision-making. This paradigm shift not only enables real-time, secure interactions between physical and virtual systems but also facilitates mass customization and drives innovation across industrial sectors [1, 2] . A critical aspect of effective HRI in Industry 5.0 is contextual and situational awareness, which ensures that cobots can dynamically adapt to their environment and human counterparts. Trajectory tracking and intention recognition play a fundamental role in achieving this awareness. Intention recognition enables cobots to infer human goals and anticipate their next steps, allowing for proactive and seamless collaboration. In addition to its capacity to anticipate human goals and respond proactively, intention recognition is pivotal for ensuring safety in HRI. By enabling cobots to predict human movements, it minimizes the risk of collisions and mitigates potential hazards, thereby creating a safer and more reliable interaction framework in dynamic environments [3]. Advanced techniques such as neural networks and deep learning algorithms have been employed to improve accuracy and adaptability in diverse environments [4, 5]. By continuously monitoring human motion and predicting future actions, trajectory tracking enhances safety by preventing collisions and optimizing movement coordination. Together, these capabilities empower cobots to respond intelligently to changing task demands, environmental conditions, and human behaviors, leading to more adaptive, efficient, and intuitive interactions within industrial settings [6].

Recent advancements in indoor trajectory tracking leverage a combination of sensor fusion, machine learning, and probabilistic modeling to enhance accuracy and robustness [7]. Traditional approaches rely on inertial measurement units (IMUs), radio frequency (RF)-based methods (e.g., Wi-Fi, Bluetooth, UWB), and computer vision techniques for localization and motion estimation [8]. Deep learning models, such as recurrent neural networks (RNNs) and

https://doi.org/10.21872/2025IISE_6188

transformer-based architectures, have improved trajectory prediction by capturing complex spatiotemporal dependencies [9]. Probabilistic filtering techniques, including Kalman and particle filters, remain widely used for real-time tracking and noise reduction [10]. Despite these advancements, challenges such as occlusions, dynamic environments, and real-time processing constraints continue to drive research toward more efficient and scalable solutions [7-9]. [7, 9] have conducted a comprehensive review of current trends and practices in indoor tracking, covering theoretical foundations, methodologies, and enabling technologies. Building on this foundation and following the current trends as reported by [7-9] , the present work focuses on the role of ensembled deep learning models in enhancing indoor trajectory tracking in HRI. As for intention recognition, we have published a comprehensive survey on human behavior modeling in HRI [10]. State-of-the-art intention recognition in HRI relies on machine learning and/or probabilistic modeling to enable cobots to infer human goals and respond proactively [11]. Recent approaches leverage vision-based methods [12], such as pose estimation and gaze tracking, alongside physiological signals (e.g., EEG, EMG) and speech processing to enhance recognition accuracy. Deep learning models, including convolutional neural networks (CNNs) and transformers, have significantly improved the ability to extract meaningful patterns from high-dimensional data [11-14]. Probabilistic frameworks, such as hidden Markov models (HMMs) [13] are widely used for real-time intention inference under uncertainty. Recent studies have introduced methods for early human intention prediction. For instance, one approach involves extracting features from human motion trajectories and employing a Hidden Markov Model (HMM) to identify state transitions. While, using Transformer and Bidirectional Long Short-Term Memory (Bi-LSTM) models to classify motion intentions [15].

However, while intention recognition and trajectory tracking individually play crucial roles in enhancing HRI in manufacturing, their combined potential remains largely unexplored in the literature. Integrating these two concepts allows cobots to not only infer human intentions but also anticipate and adapt to their movements in real time, fostering truly dynamic and adaptive HRI systems. One of the key contributions of this paper is the application of algorithms that unify these approaches to optimize their collective impact. Building on recent advancements, our research aims to push these boundaries further by developing and training ensemble deep learning models [16, 17] that leverage the strengths of CNNs, LSTM networks, and Transformers. By employing an ensemble learning framework, we seek to overcome the limitations of individual models, enhancing robustness, adaptability, and real-time responsiveness in complex HRI scenarios. Ultimately, our goal is to enable cobots to better perceive, predict, and respond to human intentions and actions, facilitating more intuitive, seamless, and effective collaboration in industrial settings.

2. Methodology

As detailed in Fig.1, we propose a framework to enable parallel trajectory tracking and intention recognition. The framework comprises three fundamental components: data acquisition, data processing, and ensembled learning.

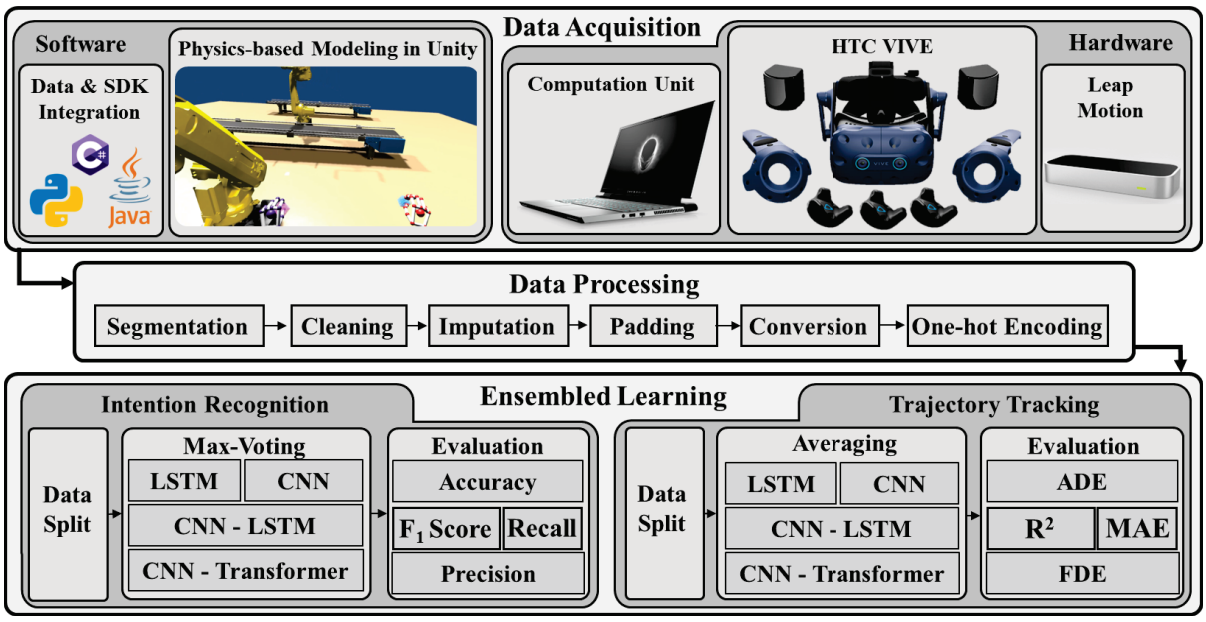

Figure 1. The proposed ensembled learning framework

https://doi.org/10.21872/2025IISE_6188

2.1. Data Acquisition

This component creates a holistic representation of human actions in the virtual environment and tracks human motion patterns and decision-making in a collaborative industrial setting via using Unity3D, HTC VIVE Pro Eye, and a leap motion controller. Unity3D, a physics-based game engine, serves as the foundation for creating immersive virtual environments, seamlessly integrating various software development kits (SDKs) and data processing pipelines using C#, Java, and Python. This enables the synchronization of multiple hardware components, facilitating real-time tracking and analysis of human motion within the virtual workspace. By leveraging Unity3D, the system ensures a high-fidelity simulation that accurately represents real-world industrial scenarios, providing an interactive platform for studying HRI. The HTC VIVE Pro Eye system captures users' spatial orientation and movement patterns. The head-mounted display (HMD) provides high-resolution visualization while simultaneously tracking head movements and gaze direction. Additionally, HTC VIVE trackers, placed on the chest and elbows, record spatiotemporal body trajectories, enabling precise motion analysis. Complementing this system, the Leap Motion controller enhances the fidelity of hand tracking by capturing detailed joint movements and finger gestures. Unlike traditional motion-tracking systems that rely on external markers or gloves, the Leap Motion device uses infrared sensors to model hand movements with high precision, enabling in-depth analysis of manual interactions and realistic interaction with virtual objects in VR. This is particularly useful for understanding how users manipulate objects, perform fine-motor tasks, and interact with cobots.

2.2. Data Processing

The data undergoes a structured processing phase, beginning with segmentation via change point analysis. Given the continuous nature of sensory data collection, change point analysis facilitates the identification of shifts in intention or trajectory. Segmentation is conducted using the Pruned Exact Linear Time (PELT) algorithm. Next, data cleaning is performed to address inconsistencies. Observations with more than 30% missing values are removed, while the remaining observations use iterative imputation for addressing missing values. Upon completion of normalization, the data is restructured into a sequential time series format appropriate for the study. Padding is then applied to standardize sequence lengths across all samples, ensuring uniformity for model training. If required, categorical labels are encoded using one-hot encoding to enhance compatibility with deep learning algorithms.

2.3. Ensemble Learning

After processing, the dataset is divided into 80% for training and 20% for validation. Model training is conducted with parameter adjustments guided by validation set performance. The predictive framework in this study integrates neural network architectures designed to effectively model the complexities of human movement trajectory and intentions in manufacturing settings. The model architectures are adjusted accordingly to accommodates the needs for intention recognition and trajectory tracking. Given the spatiotemporal nature of the dataset, the selected models include CNN, LSTM, CNN-LSTM, and CNN-Transformer. CNNs excel at capturing spatial features, while LSTMs are proficient in capturing long-term dependencies, making them highly effective for time series analysis with temporal variations. CNN-LSTM seamlessly combines spatial analysis with temporal sequence learning. Lastly, CNN-Transformers integrate the spatial feature extraction capabilities of CNNs with the sequence processing and attention mechanisms of Transformers. To further enhance model performance, ensemble learning techniques such as maximum voting and averaging are applied to deliver robust predictions for tracking human intentions and locations in HRI.

2.3.1. Intention Recognition

In our intention recognition study, seven general classes of activities are defined: Idle (sanding in place), Bending (bending down), Sitting (sitting on the ground/chair), Moving/walking (walking around), Relocating (putting parts in a box/table), Grabbing with one hand (grabbing a component in either right or left hand) and Grabbing with two hands (grabbing a component with both hands). Training CNN, LSTM, CNN-LSTM, and CNN-Transformer models for intention recognition over seven classes requires careful selection of model architectures and optimization strategies.

CNNs excel at spatial feature extraction, making them particularly effective for handling skeletal pose data. By applying multiple convolutional layers, CNNs detect movement patterns such as posture shifts and hand positioning. These extracted spatial features are then classified into one of the seven intention categories through fully connected layers. To enhance generalization, batch normalization and dropout layers reduce overfitting, while the Adam optimizer ensures stable convergence. LSTM networks, in contrast, are designed to capture temporal dependencies in sequential motion data, making them well-suited for tracking continuous human movements. The LSTM model processes time-series inputs through memory cells that retain long-term dependencies, enabling it to differentiate between actions such as bending, walking, and relocating. Multiple LSTM layers improve the model’s ability to track

evolving motion patterns, while dropout regularization mitigates overfitting. The Adam optimizer with gradient clipping stabilizes training, preventing issues like the exploding gradient problem common in deep recurrent networks.

A CNN-LSTM hybrid model integrates the strengths of both architectures, making it particularly effective for recognizing actions that require both spatial and temporal understanding. The CNN component identifies key postural cues, while LSTM layers model temporal dependencies to distinguish between similar but sequentially different actions. To enhance performance, the model employs Adam optimizer, and weight decay for regularization. The CNN-Transformer model further improves sequential modeling by capturing long-range dependencies in motion data. The CNN component extracts spatial features, which are then processed through pooling layers to reduce dimensionality while preserving key movement characteristics. Transformers use self-attention mechanisms to dynamically weigh the importance of different motion frames to better differentiate between similar movements with different intent, such as grabbing with one hand versus relocating an object. The multi-head self-attention mechanism enhances feature representation, while positional encoding preserves temporal order information.

Each model demonstrates strong accuracy in predicting human intentions. To further enhance predictive performance, ensemble learning is employed using maximum voting among the models, given the classification nature of the task (e.g., identifying behaviors such as standing, bending, or grabbing). This method is especially advantageous when dealing with large datasets, as it improves robustness and reliability in the predictions. Metrics such as accuracy, precision, recall, and F1-score are used to evaluate the performance of the trained classifiers. While accuracy provides a general assessment of model performance, precision indicates the reliability of positive classifications and recall reflects the model's ability to detect positive instances. Finally, the F1-score, the harmonic mean of precision and recall, balances these two metrics, making it particularly useful for handling class imbalances.

2.3.2. Trajectory Tracking

Trajectory tracking involves predicting the future positions of a moving human based on past movement data. Since trajectory data is inherently sequential, models must be capable of processing time-series data while extracting meaningful movement patterns. In this portion of our study, we incorporate six common walking behaviors to model human motion: Straight Line (linear movement without direction changes), Zigzag (sharp left-right turns in sequence), Circular (repetitive curved loops), Random (unpredictable path changes), S-Shaped (smooth curved turns in alternating directions), and Stopping-Starting (pauses at workstations or checkpoints). These patterns reflect diverse scenarios in manufacturing, such as task-driven pauses, obstacle navigation, and structured workflows. Variations account for individual differences and environmental factors, ensuring adaptable prediction for HRI.

As CNNs are most effective when trajectory data is represented as spatial heatmaps, the raw coordinate data are transformed into a spatial representation first. Then, convolutional layers are applied to extract spatial features, which are later passed to fully connected layers for trajectory prediction. As LSTM networks excel at modeling sequential dependencies, LSTM learns temporal patterns such as velocity changes, acceleration trends, and movement transitions given a sequence of past positions for each body part. Passing the sequential motion data through multiple LSTM layers to let the model capture long-term dependencies and movement trends, while minimizing Mean Squared Error (MSE) loss to optimize trajectory predictions. A hybrid CNN-LSTM model combines the strengths of CNNs for spatial feature extraction and LSTMs for temporal sequence modeling. In this architecture, CNN layers extract meaningful spatial representations of body part trajectories, which are passed into an LSTM layer, to model the evolution of movement over time. Whereas the CNN-Transformer model takes trajectory tracking a step further by incorporating self-attention mechanisms to model global temporal dependencies in human movement. In this architecture, CNN layers first extract spatial features from trajectory representations, similar to the CNN-LSTM model. These features are then fed into a Transformer encoder, which applies multi-head self-attention to identify important motion sequences. Unlike LSTMs, which process data sequentially, Transformers can analyze all trajectory points simultaneously, allowing for better understanding of movement relationships across different body parts.

To further enhance predictive performance, ensemble learning is employed using averaging among the models, as trajectory prediction is a regression problem. This method is particularly effective when working with large datasets, as it improves the precision and reliability of the trajectory predictions. To assess the performance of trajectory prediction models, four key metrics are employed: Mean Absolute Error (MAE), R^2, Average Displacement Error (ADE), and Final Displacement Error (FDE). While MAE offers a robust and intuitive measure of prediction error and R^2 measures model's goodness-of-fit, ADE and FDE reflect the model's ability to accurately capture the entire trajectory and ensure good precision in predicting the destination.

https://doi.org/10.21872/2025IISE_6188

3. Results and Discussion

Fig.2. displays the performance of the four models for intention recognition. The CNN-Transformer model demonstrates exceptional performance, achieving near-perfect precision, recall, and F1-scores across all activity classes, with an overall accuracy of 99.42%. In comparison, the LSTM model exhibits weaker performance, struggling with the classification of 'sitting' and 'standing' activities. The superior ability of the Transformer architecture to handle temporal dependencies and capture intricate patterns highlights its advantage in human intention recognition.

Figure 2. The comparison of four models based on the specific metrics.

Figure 3 details the performance the trained models in trajectory tracking. As shown in Figure 3, the CNN-Transformer model demonstrates superior performance across most metrics. The CNN-LSTM model also performs well, as its sequential memory capabilities excel in capturing frequent directional changes. In contrast, the baseline CNN model lags, as it lacks the temporal awareness needed to model dynamic movements effectively. Overall, the CNN-Transformer and CNN-LSTM models consistently outperform the other models, with the choice depending on the specific trajectory characteristics—structured patterns favor CNN-LSTM, while complex, less predictable trajectories benefit from the CNN-Transformer's attention mechanisms.

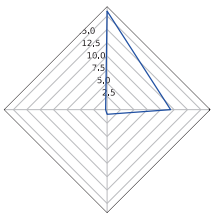

Figure 3. The comparison of four models based on the specific metrics.

To improve the performance of the trained models, ensemble learning has been employed using maximum voting for intention recognition and averaging for trajectory tracking. This approach leverages the complementary strengths of individual models, enhancing overall robustness. For intention recognition, the ensemble model achieves an accuracy of 99.14%, which is slightly lower than that of the standalone CNN-Transformer model. This discrepancy arises because maximum voting assigns equal weight to all predictions, allowing weaker models—such as LSTM—to influence the final outcome. For trajectory tracking, ensemble learning yields performance scores of 10.07 for ADE, 0.27 for FDE, 0.95 for R^2, and 6.50 for MAE. Notably, the averaging strategy mitigates the impact of biases or errors from individual models by smoothing out prediction variances. This enhances generalization and stability, making ensemble learning particularly well-suited for complex human-robot interaction (HRI) scenarios.

4. Conclusion

This study demonstrates the effectiveness of ensemble learning in enhancing HRI by integrating intention recognition and trajectory tracking within a unified framework. By employing maximum voting for intention recognition and

averaging for trajectory tracking, we leverage the complementary strengths of CNNs, LSTMs, and Transformers while mitigating the limitations of individual models. The results indicate that ensemble learning significantly improves robustness, particularly in capturing complex and dynamic human behaviors within VR-generated environments. For intention recognition, the ensemble framework achieved higher precision and recall by combining the distinct capabilities of its constituent models. Likewise, for trajectory tracking, the ensemble approach enhanced prediction reliability by balancing the sequential modeling strengths of CNN-LSTM with the contextual adaptability of CNN-Transformer. These findings highlight the versatility and scalability of ensemble learning in both classification and regression tasks within HRI, providing a robust solution for real-time decision-making in collaborative settings.

References

[1] M. Dhanda, B. A. Rogers, S. Hall, E. Dekoninck, and V. Dhokia, "Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0," *Robot Comput Integr Manuf*, vol. 93, p. 102937, 2025.

[2] R. R, R. R. Sathya, V. V, B. S, and J. L. N, "Industry 5.0: Enhancing Human-Robot Collaboration through Collaborative Robots – A Review," in *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2023, pp. 1–6. doi: 10.1109/ICAECA56562.2023.10201120.

[3] A. Pilacinski *et al.*, "Human in the collaborative loop: a strategy for integrating human activity recognition and non-invasive brain-machine interfaces to control collaborative robots," *Front Neurorobot*, vol. 18, 2024, doi: 10.3389/fnbot.2024.1383089.

[4] F. Formica, S. Vaghi, N. Lucci, and A. M. Zanchettin, "Neural Networks based Human Intent Prediction for Collaborative Robotics Applications," in *2021 20th International Conference on Advanced Robotics (ICAR)*, 2021, pp. 1018–1023. doi: 10.1109/ICAR53236.2021.9659328.

[5] R. Zhong, B. Hu, Z. Hong, Z. Zhang, Y. Feng, and J. Tan, "A Human Digital Twin Based Framework for Human–Robot Hand-Over Task Intention Recognition," in *ICMD: International Conference on Mechanical Design*, Springer, 2023, pp. 283–295.

[6] J. Palmieri, P. Di Lillo, M. Lippi, S. Chiaverini, and A. Marino, "A Control Architecture for Safe Trajectory Generation in Human–Robot Collaborative Settings," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 365–380, 2025, doi: 10.1109/TASE.2024.3350976.

[7] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[8] B. D. B. Chowdhury, S. Masoud, Y.-J. Son, C. Kubota, and R. Tronstad, "A dynamic HMM-based real-time location tracking system utilizing UHF passive RFID," *IEEE Journal of Radio Frequency Identification*, vol. 6, pp. 41–53, 2021.

[9] L. Qi, Y. Liu, Y. Yu, L. Chen, and R. Chen, "Current Status and Future Trends of Meter-Level Indoor Positioning Technology: A Review," *Remote Sens (Basel)*, vol. 16, no. 2, p. 398, 2024.

[10] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke, "Robotic vision for human-robot interaction and collaboration: A survey and systematic review," *ACM Trans Hum Robot Interact*, vol. 12, no. 1, 1–66, 2023.

[11] Kamali Mohammadzadeh, A., Allen, C. L., & Masoud, S. (2023, June). VR Driven Unsupervised Classification for Context Aware Human Robot Collaboration. In *International Conference on Flexible Automation and Intelligent Manufacturing* (pp. 3-11). Cham: Springer Nature Switzerland.

[11] R. Jahanmahin, S. Masoud, J. Rickli, and A. Djuric, "Human-robot interactions in manufacturing: A survey of human behavior modeling," *Robot Comput Integr Manuf*, vol. 78, p. 102404, 2022.

[12] D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley, "A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction," *Appl Mech Rev*, vol. 70, no. 1, p. 010804, 2018.

[13] L. Yan, X. Gao, X. Zhang, and S. Chang, "Human-Robot Collaboration by Intention Recognition using Deep LSTM Neural Network," in *2019 IEEE 8th International Conference on Fluid Power and Mechatronics (FPM)*, 2019, pp. 1390–1396. doi: 10.1109/FPM45753.2019.9035907.

[14] X. Zhang, S. Tian, X. Liang, M. Zheng, and S. Behdad, "Early Prediction of Human Intention for Human–Robot Collaboration Using Transformer Network," *J Comput Inf Sci Eng*, vol. 24, no. 5, 2024.

[15] Z. Li *et al.*, "An Ensemble Learning Framework for Vehicle Trajectory Prediction in Interactive Scenarios," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 51–57. doi: 10.1109/IV51971.2022.9827070.

[16] B. Geng, J. Ma, and S. Zhang, "Ensemble deep learning-based lane-changing behavior prediction of manually driven vehicles in mixed traffic environments.," *Electronic Research Archive*, vol. 31, no. 10, 2023.

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.