



A two-stage context-aware meta-ensemble framework for collision risk prediction in Internet of Vehicles

Tamara Zhukabayeva^{1,2,3}, Zulfiqar Ahmad^{2,4}, Nurdaulet Karabayev², Dilaram Baumuratova^{2,5}, Elhadj Benkhelifa^{6,7}

Keywords:

Internet of Vehicles, collision risk prediction, cyber-physical systems, feature-level fusion, deep learning, explainable artificial intelligence, robustness analysis

Citation: Zhukabayeva, T.; Ahmad, Z.; Karabayev, N.; Baumuratova, D.; Benkhelifa, E. A two-stage context-aware meta-ensemble framework for collision risk prediction in Internet of Vehicles. *Intell. Robot.* 2026, 6(3), 643-77. <https://dx.doi.org/10.20517/ir.2026.30>

Received: 15 Jun 2026
First Decision: 4 Aug 2026
Revised: 20 Aug 2026
Accepted: 3 Sep 2026
Published: 24 Sep 2026

Academic Editor:

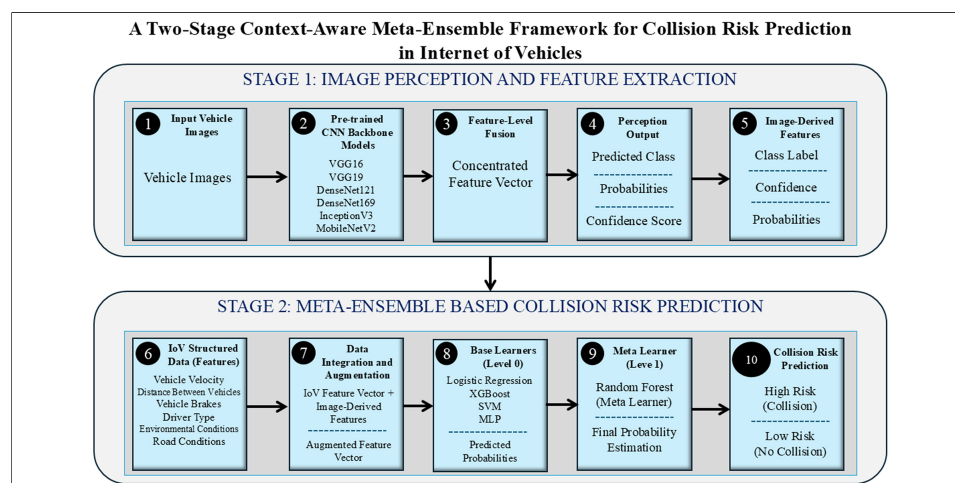
Lei Lei

Copy Editor:

Pei-Yun Wang

Production Editor:

Pei-Yun Wang



Abstract

A key challenge in Internet of Vehicles (IoV) environments is accurately predicting collision risks, given the dynamic nature of traffic systems and the variety of available data sources. Existing approaches often rely on structured IoV data, while visual contextual information is not fully incorporated into the prediction process. This limitation motivates a unified framework that integrates environmental perception with sensor-derived information to enable more accurate collision-risk prediction. In this study, we propose a two-stage context-aware meta-ensemble framework that combines image-based perception and IoV-based predictive analysis for collision-risk prediction. The first stage uses multiple pre-trained convolutional neural networks to obtain rich visual representations from vehicle images through feature-level fusion. The second stage employs a hybrid stacked ensemble architecture that combines the probabilistic outputs of heterogeneous machine learning models with IoV and image-derived features, which a meta-learner then processes for final

¹School of AI+X, Qazaq AI Research University, Astana 010000, Kazakhstan.

²Department of Cryptology, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan.

³Department of Software Engineering, Astana IT University, Astana 010000, Kazakhstan.

⁴Department of Computer Science and Information Technology, Hazara University, Mansehra 21300, Pakistan.

⁵Department of Education, Astana International University, Astana 010000, Kazakhstan.

⁶Smart Systems, AI and Cybersecurity Research Centre, University of Staffordshire, Stoke-on-Trent ST4 2DE, UK.

⁷School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool L3 3AF, UK.

Correspondence to: Dr. Zulfiqar Ahmad, Nurdaulet Karabayev, Department of Cryptology, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan. E-mail: zulfiqarahmad@hu.edu.pk; 020419501012@enu.kz

collision-risk prediction. The proposed feature-fusion model performs best in the perception stage, with an accuracy of 0.97. In the IoV-based prediction stage, the proposed stacked ensemble architecture outperforms all baseline models, achieving an accuracy of 0.9847. Robustness analysis shows that the perception module tolerates blur and rotation but remains sensitive to Gaussian noise. Cross-stage integration incorporates image-derived features into the IoV feature space, providing complementary visual context for collision-risk prediction and demonstrating its potential for intelligent transportation and cyber-physical IoV applications.

1. INTRODUCTION

Intelligent transportation systems are evolving rapidly, and a major paradigm driving this evolution is the Internet of Vehicles (IoV), which aims to provide connected, autonomous, and data-driven vehicle transportation^[1,2]. IoV extends the Internet of Things (IoT) by integrating vehicles, roads, sensors, and communication networks into a seamless environment. Vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-everything (V2X) are three fundamental communication environments that support real-time information exchange with various entities^[3–5]. This integrated system allows vehicles to communicate with each other and provides valuable information such as speed, location, traffic data, and driver behavior. Therefore, IoV plays a significant role in optimizing traffic flow, decreasing congestion, and enhancing road safety. However, an IoV environment is complex and dynamic, with high mobility, heterogeneous data sources, and unpredictable driving scenarios. These challenges make collision risk prediction a key concern. Traditional methods that merely extract structured sensor data cannot fully represent the context of a driving situation, especially visual information such as surrounding objects, road type, and environment. This constraint underscores the need for more comprehensive and context-aware solutions that can effectively combine multiple sources of information to gain higher accuracy and reliability in their predictions^[6].

Cyber-physical systems (CPS) provide a solid basis for designing intelligent and adaptive transportation systems. CPS combines computation with physical processes through continuous sensing, communication, and control. In an IoV-based CPS, vehicles act as intelligent agents that communicate with physical and digital infrastructure. Sensors gather real-time data from the surrounding environment, computational models process the data to provide actionable information, and the control mechanism takes action based on dynamic environments. The integration of the physical and cyber worlds allows for real-time monitoring, predictive analytics, and automated decision-making in safety-critical CPS applications. By incorporating these elements into a CPS, the system becomes more sophisticated and can operate in real time, continuously adapting to its surroundings. The CPS-based IoV systems can be more complete and predictive by incorporating data from various sources. However, designing systems that can manage heterogeneous data, scale, and remain reliable amid uncertainty is challenging. These requirements drive the need for an intelligent framework that bridges the gap between perception and decision-making and combines data-driven and context-aware technologies^[7–9].

As intelligent transportation systems become more complex, advanced artificial intelligence techniques, specifically deep learning and machine learning, are increasingly used for perception and decision-making tasks. Deep learning, particularly convolutional neural networks (CNNs), has achieved outstanding performance in vehicle image classification by automatically extracting hierarchical, discriminative features from raw images. Unlike handcrafted computer vision algorithms, CNN-based algorithms learn spatial patterns such as edges, shapes, textures, and object structures in the data. This ability allows vehicles to be classified correctly under different lighting conditions, orientations, occlusions, and background complexities. Therefore, deep learning has become an integral part of perception systems for many modern vehicle applications. In addition to perception, machine learning is also significant for predicting collision

risks and analyzing structured data from IoV environments^[10,11]. Logistic regression (LR), support vector machines (SVMs), multi-layer perceptrons (MLPs), and ensemble models are among the popular algorithms used to model relationships among factors such as vehicle speeds, distance, driving behavior, and environmental conditions^[12–14]. These models can identify complex patterns and relationships in large amounts of data, potentially predicting possible collisions. In contrast to rule-based systems, machine learning (ML)-based methods provide better adaptability, scalability, and predictive capabilities in dynamic traffic situations. Deep learning and machine learning can be used together to overcome the shortcomings of using one modality. Deep learning is well suited to capture context and visual information from images, and machine learning is well suited to model linear sensor data and decision boundaries in images. They can be used together to combine visual and numerical modalities for more comprehensive, contextual predictions. This type of synergy is particularly important for collision risk prediction. The advantages of ensemble methods lie in combining multiple models with varying learning characteristics, thereby reducing bias^[15–17].

In IoV environments, accurate collision risk prediction remains a critical challenge because traffic systems are dynamic and heterogeneous. Let the IoV dataset be represented as $D_V = \{(X_i, y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$ denotes structured features such as vehicle velocity, inter-vehicle distance, driver status, braking capability, and environmental conditions, and $y_i \in \{0, 1\}$ represents the collision label. Traditional approaches aim to learn a function $f: \mathbb{R}^d \rightarrow \{0, 1\}$ to predict collision risk based solely on these features. However, such models often fail to capture contextual information from the surrounding environment. To address this limitation, consider an additional image dataset $D_I = \{I_i\}_{i=1}^n$, where each image I_i contains visual information related to the driving environment. A deep learning model extracts image-based features, $f_i^{img} \in \mathbb{R}^k$ such that:

$$f_i^{img} = g(I_i),$$

where $g(\cdot)$ represents the feature extraction function. The augmented feature space can then be defined as:

$$X'_i = [X_i, f_i^{img}],$$

which combines structured IoV data with visual information.

The objective of this research is to learn an optimal prediction function:

$$F: \mathbb{R}^{d+k} \rightarrow \{0, 1\},$$

that minimizes the prediction error:

$$\min_F E[L(y_i, F(X'_i))],$$

where $L(\cdot)$ is a suitable loss function.

The direct learning of such a function, however, is difficult because of (i) the diversity in data modalities; (ii) the complex non-linear relationships between features; and (iii) real-world variations in driving conditions. Thus, a strong framework is required to integrate visual and IoV data and to leverage multiple learning models to improve prediction accuracy and generalization.

In this work, we tackle these issues by constructing a two-stage context-aware meta-ensemble approach, which is composed of deep learning techniques for extracting image features and a stacked ensemble learning technique for collision risk prediction. The aim is to improve prediction capability, robustness, and

context awareness in CPSs based on IoV. The key contributions of this research are summarized as follows:

- A unified two-stage architecture is proposed to connect the image-based perception and IoV-based collision risk prediction, allowing the seamless integration of visual understanding and decision-making in cyber-physical vehicle systems.
- To incorporate image-derived information into the IoV dataset, a novel cross-stage integration strategy is introduced, in which image-derived features (class probabilities, predicted labels, and confidence scores) are injected into the IoV dataset, and the model uses both visual and sensor-based information in a unified predictive framework.
- Multiple pre-trained CNN models (VGG16, VGG19, DenseNet121, DenseNet169, InceptionV3, and MobileNetV2) are used to obtain rich and diverse visual representations through feature-level fusion, thereby enhancing the system's perception capability.
- An ensemble learning model is constructed by combining heterogeneous base learners [LR, SVM, MLP, and XGBoost (XGB)] with augmented features and applying a random forest meta-learner to boost the learning precision and generalization.
- To ensure the trustworthiness of models for safety-critical applications, the study includes explainable AI techniques such as Grad-CAM, SHapley Additive exPlanations (SHAP), and local interpretable model-agnostic explanations (LIME) analysis, offering both visual and feature-level interpretability.
- The robustness of the perception stage is evaluated under controlled noise, blur, and rotation perturbations, revealing its tolerance to blur and rotation as well as its sensitivity to Gaussian noise.

The rest of the paper is organized as follows: Section 2 provides the related work. Section 3 describes the proposed framework. Section 4 highlights the performance evaluation. Section 5 presents experiments, results, and discussion, and Section 6 concludes the article with several future directions.

2. RELATED WORK

With the recent development of IoV and intelligent transportation systems, deep learning and machine learning have been effectively applied in accident identification, collision prediction, and safety improvement.

In^[18], a federated deep learning model for IoV-based collision prediction is proposed, which combines federated learning and reinforcement mechanisms to optimize both storage and collision prediction. The model performs well with high accuracy, which proves the effectiveness of temporal learning in IoV systems. But the method is primarily geared towards sequential data and lacks visual perception and heterogeneous feature integration. In^[19], an IoV-based accident detection system with an edge-assisted architecture is presented, where on-vehicle inference and validation from a roadside unit (RSU) are done jointly to ensure real-time detection with low latency. A hybrid deep learning ensemble is used for the model, and high accuracy (98.65%) is obtained. Although this represents a strong performance, the framework is mainly based on visual information and does not integrate structured features of IoV. The study in^[20] suggests that accident risk prediction can be performed by a deep learning-based CNN-long short-term memory (LSTM)-graph neural network (GNN) model, which is capable of capturing spatial, temporal, and relational dependencies in traffic data. They model spatiotemporal relationships, improving prediction accuracy to a great extent. The model, however, is based solely on trajectory and sensor data and lacks any visual contextual information.

In^[2], an IoV fog-assisted accident detection framework is proposed where machine learning techniques are applied to learn from sensor data from vehicles in edge and fog layers. The study shows high latency and efficient detection in a distributed environment. But the method is based on data processing using sensors and does not discuss the integration of perception-based features. In^[21], a real-time accident detection system

based on deep learning is proposed in which CNN-based models are deployed on the vehicles and combined with the IoV communication systems. The study shows the need for distributed intelligence and doesn't include an organized mechanism for data fusion across multiple modalities. An uncertainty-aware collision prediction model based on edge computing and a recurrent neural network is presented in [22]. The framework can forecast vehicle movements and evaluate intersection collision risk with high accuracy. While the approach correctly models uncertainty, it is still mainly based on the prediction of a trajectory and neglects image-based contextual features. Recent developments on ensemble learning for IoV systems described in [6] show that stacking is a powerful approach that combines several systems to improve prediction performance. However, these techniques are typically used in cybersecurity and are not used for collision prediction and cross-domain feature integration.

The new developments in IoV and intelligent transportation systems (ITS) increasingly focus on real-time sensing, reliable V2V communication, and smart network management. The study presented in [23] introduced a multi-access edge computing (MEC) system for real-time vehicle tracking in large-scale roadside sensor networks. By combining camera and radar data, the authors used machine learning techniques to predict trajectories and provide scalable tracking with low processing latency. The study in [24] presents an optical communication-based V2V approach for vehicle platooning, in which an LED matrix serves as the transmitter and a monocular camera as the receiver. The study therefore complements the conventional V2V communication approach and creates network-independent communication. The authors in [25] explored combining of LoRa and distributed machine learning to optimize network connectivity for green and intelligent transportation by dynamically tuning the transmission parameters and optimizing the allocation of gateways and end-devices to minimize interference, congestion, and energy consumption. The studies indicate that IoV is rapidly evolving into a platform for sensing, resilient V2V communication, and intelligent network optimization, but they mainly focus on tracking, communication, and network connectivity rather than cross-modal visual-context integration with structured IoV data for collision risk prediction.

The current literature shows that deep learning, machine learning, edge computing, and ensemble learning techniques have been successfully used for accident detection and collision risk prediction in IoV scenarios. However, most existing studies focus on structured IoV data or visual information in isolation, limiting context awareness and adaptability in complex traffic scenarios. Moreover, the integration of heterogeneous data sources, robust ensemble learning mechanisms, and decision explanations has not been thoroughly examined. This study introduces a context-aware cyber-physical framework combining image-based perception with IoV sensor information via a two-stage meta-ensemble architecture to overcome these shortcomings. The proposed algorithm is expected to enhance the accuracy, reliability, and explainability of collision risk prediction in ITSs by leveraging deep feature extraction, cross-modal feature fusion, stacked ensemble learning, and explainable AI methods.

3. PROPOSED FRAMEWORK

This study presents a two-stage context-aware collision risk prediction system aimed at addressing collision issues in IoV environments, leveraging visual perception and data-driven decision-making. The system architecture shown in Figure 1 is based on the concept of a CPS pipeline, which involves the cooperation between the perception and prediction elements for intelligent and reliable collision risk assessment. The first stage uses a deep learning-based perception module to process vehicle images and provide meaningful visual information. Various pre-trained CNN architectures are used to extract different representations of features and are combined with a feature-level fusion strategy. The visualization in this stage is implemented

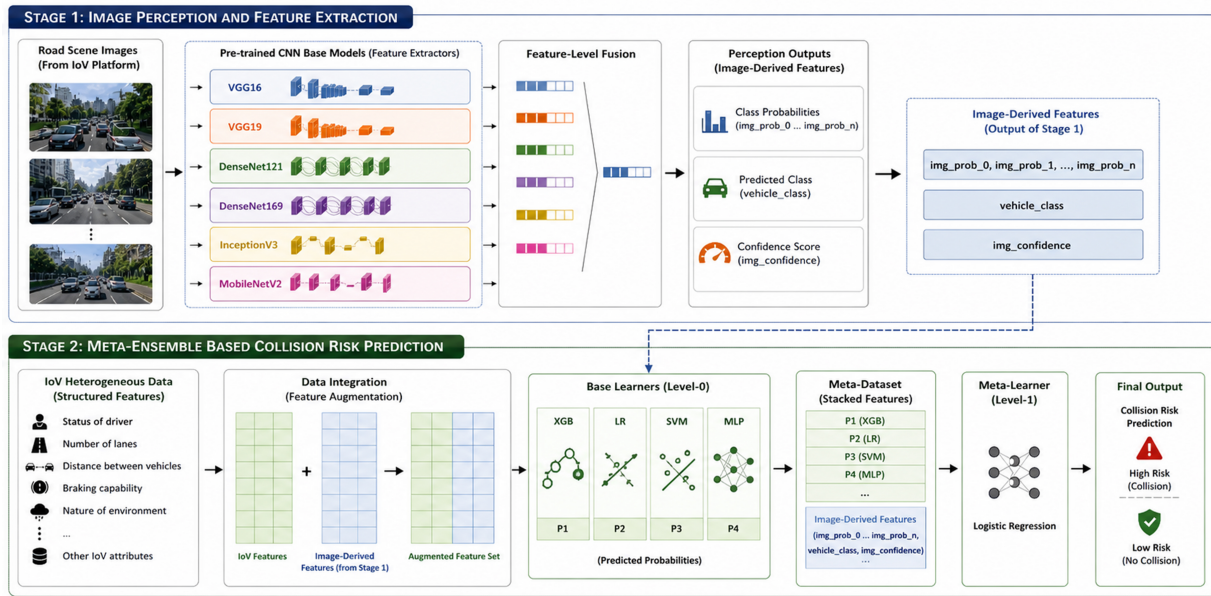


Figure 1. Two-stage context-aware framework for collision risk prediction in IoV. IoV: Internet of Vehicles; XGB: XGBoost; LR: logistic regression; SVM: support vector machine; MLP: multi-layer perceptron.

in the form of high-level visual descriptors, predicted labels, and confidence scores. These outputs provide context about the driving environment and serve as an additional source of knowledge beyond conventional IoV sensor inputs.

In the second stage, the stacked ensemble learning approach predicts collision risk. The features in the first stage, such as vehicle velocity, distance between vehicles, vehicle brakes, driver type, environmental conditions, and road conditions, are augmented with image-derived features in the IoV dataset. This integration combines structured sensor information and visual information for context-aware learning. Base models are trained independently, and then their probabilities are fused in a weighted fashion with the expanded feature vector in order to generate a unified representation. This combined input is then passed to a meta-classifier, which produces the final collision prediction. The proposed framework's novelty lies in feature fusion, ensemble learning, and a cross-stage integration mechanism, where the perception stage output is systematically fed into the prediction stage. The design enables the model to learn intricate relationships between visual context and vehicular dynamics, and it leads to more accurate and robust predictions. The proposed framework also leverages visual intelligence, which provides a more realistic and complete understanding of traffic scenarios. The system is scalable and flexible for deployment in CPS-enabled smart transportation systems. The modular design enables optimization of each stage without compromising their seamless integration. The perception stage can be improved by using more sophisticated vision models, and the prediction stage can be extended by adding more IoV features or learning algorithms without impacting the architecture.

3.1. Algorithm for the proposed framework

The proposed algorithm [Algorithm 1] is designed to be a structured two-stage pipeline, combining image-based perception and IoV-driven collision prediction using a context-aware meta-ensemble strategy. In the first phase, the images of the vehicles are processed by a number of pre-trained CNNs, namely VGG16, VGG19, DenseNet121, DenseNet169, InceptionV3, and MobileNetV2. These models are independent, and each produces a class probability distribution of the input images. A feature-level fusion mechanism is then used to fuse these outputs to capture different visual representations. Three different image-derived features are extracted from the fused outputs: predicted class label, maximum confidence, and full probability vector. These features represent a condensed and informative visual context representation. The image-derived features from the first stage are appended to the IoV dataset, which includes structured features like driver status, number of lanes, vehicle velocity, inter-vehicle distance, braking capability, and environmental conditions. This creates an enriched feature space that combines sensor and visual information. Later, the different base learning models are independently trained on the expanded dataset, namely, LR, SVM, MLP, and XGB. All base models produce a probability output, which represents the proposed probability of collision. These outputs are then used as meta-features and fused with the original augmented features to create a higher-level representation. Next, a meta-learning layer (random forest classifier) is added, which takes the concatenated features and learns the non-linear, complex relationship between the base-model prediction and the features. This hierarchical learning approach lets the system overcome the limitations of using a single model for parameterization and detect relationships across model classes. The same sequence is followed for inference, which entails extracting image features and injecting them into the IoV data, generating probability scores from the base models, and generating predictions from the meta-classifier. The output is a binary classification indicating collision or no-collision risk. The computational complexity of the proposed two-stage framework is mainly influenced by the deep learning models used in the first stage and the ensemble learning process in the second stage. In Stage 1, a number of pre-trained CNN models are applied to obtain the visual features of input images. These models are used in inference mode with fixed weights and therefore are efficient and can be parallelized across models, leading to a reduction in overall processing time. The complexity is related to training and prediction using multiple base models (LR, SVM, MLP, and XGB) and the meta-learning step performed using a random forest classifier in Stage 2. The base models are structured and relatively light, with only a moderate computational overhead, whereas the random forest meta-learner performs a series of tree-based operations, which results in a moderate computational overhead. This overhead is manageable, however, because of the controlled parameters of the models and the efficient implementation. The multi-stage structure does not impose prohibitive computational costs. Adopting pre-trained models, parallel processing, and effective ensemble methods ensures the system is scalable and can be deployed in IoV and CPS applications in real time or near real time.

Algorithm 1: Two-Stage Context-Aware Meta-Ensemble Framework for Collision Risk Prediction

1. **Begin**
2. **Input:**
3. Image dataset $\mathcal{D}_I = \{I_1, I_2, \dots, I_n\}$
4. IoV dataset $\mathcal{D}_V = \{X_1, X_2, \dots, X_n\}$
5. Labels $Y = \{y_1, y_2, \dots, y_n\}$, where $y \in \{0,1\}$
6. **Output:**
7. Predicted collision risk $\hat{Y}$
- 8.
9. **Stage 1: Image-Based Feature Extraction**
10. **Initialize CNN Models:**
11. $\mathcal{M}_{CNN} = \{M_1, M_2, \dots, M_6\}$
12. where
- $M_i \in \{\text{VGG16, VGG19, DenseNet121, DenseNet169, InceptionV3, MobileNetV2}\}$
13. **For each image** $I_i \in \mathcal{D}_I$:
14. Extract prediction probabilities from each model:
15. $p_i^{(k)} = M_k(I_i), k = 1, 2, \dots, 6$
16. Perform feature fusion:
17. $P_i = \frac{1}{6} \sum_{k=1}^6 p_i^{(k)}$
18. Extract image features:
19. $f_i^{img} = \{\arg \max(P_i), \max(P_i), P_i\}$
20. **Construct Image Feature Set:**
21. $\mathcal{F}_{img} = \{f_1^{img}, f_2^{img}, \dots, f_n^{img}\}$
- 22.
23. **Stage 2: IoV-Based Collision Prediction**
24. **Augment IoV Features:**
25. $X'_i = [X_i, f_i^{img}]$
26. **Initialize Base Models:**
27. $\mathcal{M}_{base} = \{\text{LR, SVM, MLP, XGB}\}$
28. **Train Base Models and Generate Probabilities:**
29. $p_i^{LR}, p_i^{SVM}, p_i^{MLP}, p_i^{XGB}$
30. **Construct Meta-Feature Vector:**
31. $Z_i = [p_i^{LR}, p_i^{SVM}, p_i^{MLP}, p_i^{XGB}, X'_i]$
32. **Meta-Learning Layer**
33. **Train Meta-Classifier (Random Forest):**
34. $\hat{y}_i = M_{meta}(Z_i)$
35. **Prediction Phase**
36. **For a new sample** (I_{new}, X_{new}) :
37. Extract image features f_{new}^{img}
38. Form augmented input:
39. $X'_{new} = [X_{new}, f_{new}^{img}]$
40. Generate base model outputs:
41. $p_{new}^{LR}, p_{new}^{SVM}, p_{new}^{MLP}, p_{new}^{XGB}$
42. Construct meta-input:
43. $Z_{new} = [p_{new}^{LR}, p_{new}^{SVM}, p_{new}^{MLP}, p_{new}^{XGB}, X'_{new}]$
44. Final prediction:
45. $\hat{y}_{new} = M_{meta}(Z_{new})$
46. **Return:**
47. $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$
48. **end**

4. PERFORMANCE EVALUATION

A Python-based simulation environment was used to implement and evaluate the proposed framework. Data preprocessing and manipulation libraries such as NumPy and Pandas were used, and deep learning models were implemented with TensorFlow/Keras. The machine learning algorithms and evaluation metrics were used with the help of Scikit-learn. For visualization and performance analysis, we used Matplotlib and Seaborn.

4.1. Performance evaluation parameters

The effectiveness of the proposed framework was tested through a set of standard classification metrics in order to provide a comprehensive assessment of the effectiveness of prediction. The number of correctly classified instances measures the model's overall correctness and is therefore termed 'accuracy'. Precision reflects how well the model distinguishes and classifies events as collisions, indicating the accuracy of positive predictions. Recall refers to the model's ability to identify real collision events, a property critical for safety-critical applications. The F1 score balances precision and recall and is calculated as their harmonic mean. Moreover, the receiver operating characteristic (ROC) curve and the area under the curve (AUC) are used to evaluate the model's discriminatory power for varying threshold values. These metrics are all used to give a comprehensive analysis of the proposed collision prediction system's performance, reliability, and generalization.

4.2. Datasets

Two benchmark datasets were used for the development and testing of the proposed two-stage framework, both of which are publicly available with the titles "Vehicle Images Dataset"^[26] and "Internet of Vehicles Dataset"^[27]. The image-based dataset (Vehicle Images Dataset) supports visual perception and context extraction, while the sensor-based dataset (IoV) supports collision-risk prediction. The first is the Vehicle Images Dataset, suitable for classifying vehicles based on visual data. The Vehicle Images Dataset contains 15,645 images distributed across seven vehicle categories, including City Car, Multi-Purpose Vehicle, Sedan, Van, Sport Utility Vehicle, Truck, and Big Truck, providing diverse visual patterns for perception-based feature extraction. The images vary in background, lighting, and viewpoint, providing diverse data for training robust deep-learning models. The data is usually stored in a folder hierarchy, with each class containing a set of labeled images, which supports supervised learning such as classification. This dataset is commonly used to train and test CNNs for tasks such as object recognition and scene understanding in computer vision. In Stage 1, the dataset was used to train and evaluate various pre-trained deep learning models for vehicle classification, enabling extraction of high-level information from visual input. The second dataset is the Internet of Vehicles Dataset, which includes structured data representing the environmental and vehicular conditions in IoV environments. The Internet of Vehicles Dataset provides structured vehicular and environmental attributes, including vehicle speed, inter-vehicle distance, braking status, driver-related characteristics, and environmental/road conditions, which are used as inputs for collision risk prediction in Stage 2. Researchers have used these datasets in intelligent transportation research to model real driving environments and build predictive models for safety analysis and risk assessment. This dataset was used in Stage 2 to train machine learning and meta-ensemble models for collision risk prediction in the proposed architecture. The organized data structure enables the implementation of different classification algorithms and evaluation of their performance across traffic scenarios.

4.3. Dataset preprocessing

A structured preprocessing pipeline was applied to the image dataset to ensure data consistency and improve model generalization for efficient deep learning training. The data directory structure was first examined to ensure that data files were organized by class and that all files contained the same number of data points. All categories were grouped into separate folders by vehicle class, making them easy to label for supervised learning.

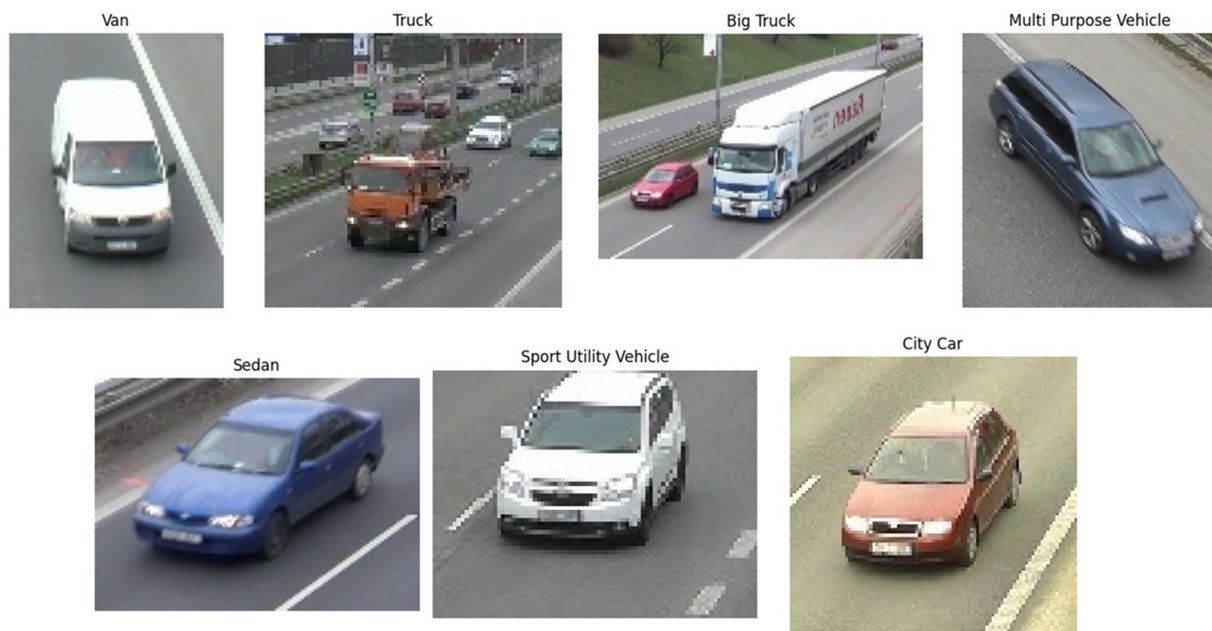


Figure 2. Representative sample from each category (the original vehicle samples are selected from the Vehicle Images Dataset^[26]).

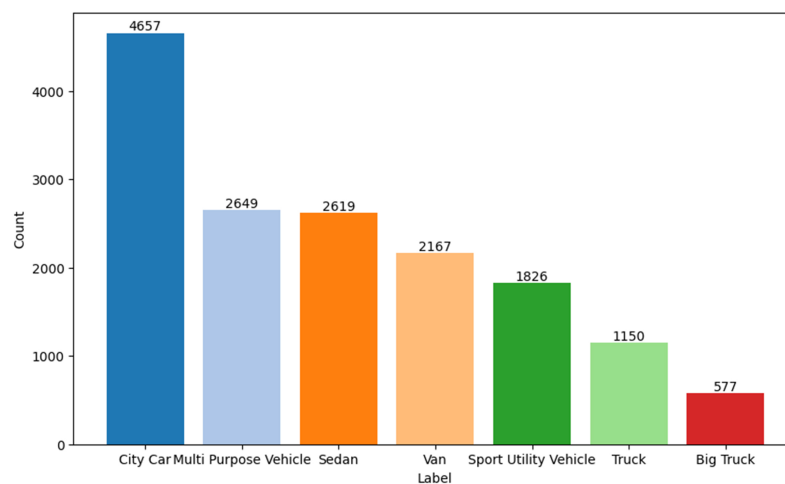


Figure 3. Distribution of samples across different classes.

A representative sample from each category is plotted in [Figure 2](#). This visualization shows a variety of vehicle types, perspectives, and background conditions. This visualization is necessary to build strong classification models. A bar chart shown in [Figure 3](#) and a pie chart shown in [Figure 4](#) are used to analyze the distribution of samples across the different classes. These visualizations provide insight into class balance and help identify potential data imbalance issues that can affect model performance.

After exploring the dataset, a structured data frame was created with the image file paths and associated labels. The data was then split into training, validation, and test sets using stratified splitting to ensure each set has the same class distribution. Specifically, 80% of the data was used for training, and 20% was split into validation and test sets (10% each). This eliminate evaluation bias and avoids data leakage.

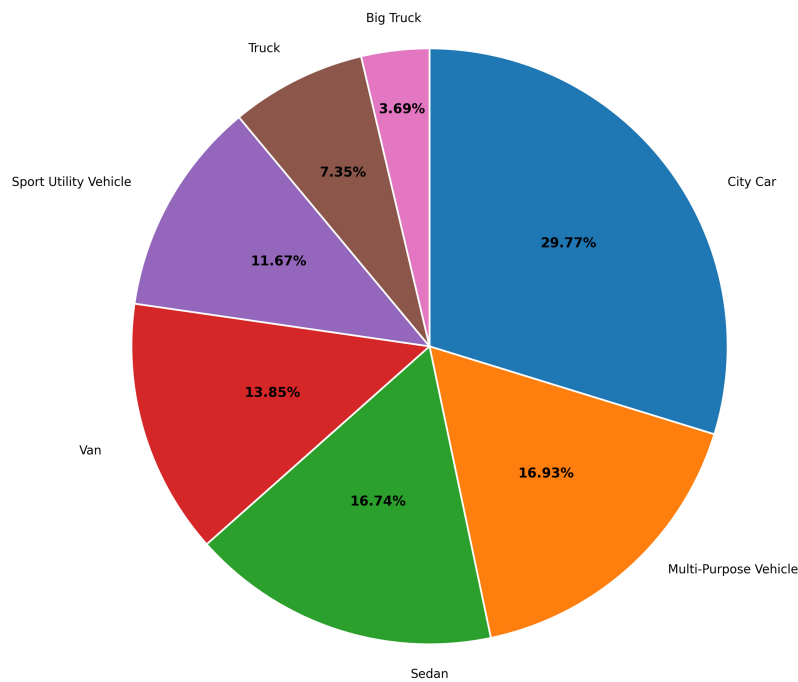


Figure 4. Percentage distribution of samples across different classes using a pie chart.

All images are resized to a fixed size of 224×224 pixels with 3 color channels (RGB) for use with deep learning models. Pixel values are rescaled to the range $[0, 1]$ to stabilize training and speed convergence. Besides, the training set is also augmented using data augmentation techniques to improve generalization and prevent overfitting. Examples include random rotations, width and height shifts, shear transformations, zoom operations, and horizontal flipping. The additional information enables simulation of a wide range of real-world changes, including variation in orientation, scale, and perspective. Efficient data loading and training: image generators dynamically provide data to the models in batches. The training generator performs both rescaling and augmentation, whereas the validation and testing generators only normalize to ensure equal evaluations. All datasets use a batch size of 32 to balance between training stability and computational efficiency.

5. EXPERIMENTS, RESULTS, AND DISCUSSIONS

This section presents the experiments and results used to validate the effectiveness of the proposed framework together with the corresponding discussion. The experimental analysis is structured to provide a comprehensive evaluation across multiple dimensions.

5.1. Stage 1: image-based context extraction

To extract rich visual representations from the vehicle images, six pre-trained CNN architectures were used, namely VGG16, VGG19, DenseNet121, DenseNet169, InceptionV3, and MobileNetV2. The models were chosen because their complementary architectural features enable different and discriminative feature extraction. We fine-tuned each model using transfer learning for the target vehicle classification task. Based on these models, we propose a new ensemble architecture, the hybrid feature fusion network (HFFN). Unlike conventional ensemble learning methods that aggregate predictions (e.g., averaging and voting), the proposed approach uses feature-level fusion by designing an ensemble of deep representations from multiple CNN backbones. Specifically, we loaded each pre-trained model and froze its parameters to retain the knowledge it has learned and reduce computational complexity and overfitting. We extracted feature vectors from the middle layers (third-to-last layer) of each network to capture high-level semantic information such

as structural patterns and visual attributes. All 6 models' extracted features are then concatenated to get a high-dimensional, combined feature representation. The fused feature space represents complementary information from various architectures, leading to a richer representation than any single architecture. To stabilize feature distributions and improve training efficiency, batch normalization was used, and fully connected layers learn more complex interactions among the fused features. In order to further reduce overfitting and improve generalization, dropout regularization was added to the model. A hybrid loss function was introduced to improve learning effectiveness by combining categorical cross-entropy loss and focal loss. This formulation allows the model to retain overall classification performance while focusing on class imbalance and hard-to-classify samples. The hybrid loss combines the two with weighted optimization, balancing overall accuracy with learning difficult samples. The final classification layer uses the softmax activation function to output the classification as one of the 7 vehicle names in the dataset. The Adam optimizer and a controlled learning rate optimize the entire architecture for stable convergence. The proposed HFFN model was designed as a multi-architecture feature extraction, feature-level fusion, and hybrid loss optimization-based vehicle classification model, which is the perception stage of the overall two-stage vehicle classification framework.

A detailed comparative study compares six individual pre-trained CNN models and the proposed HFFN across seven vehicle classes, as shown in [Table 1](#), in terms of precision, recall, F1 score, and accuracy. The results show significant performance differences between architectures, highlighting the strengths and limitations of each architecture in capturing the range of vehicle characteristics. DenseNet121 performs best in overall accuracy (0.93) and in macro and weighted averages. Its dense connectivity enables efficient feature propagation and reuse, achieving a good balance in classifying most vehicle categories. However, DenseNet169 performs worse (accuracy 0.85), showing that greater depth does not necessarily improve generalization in this task.

In a similar manner, VGG16 and VGG19 show moderate performance with accuracies of 0.90 and 0.89, respectively. Although these models obtain high precision in some classes (big truck or truck), they have low recall performance in more difficult classes (multipurpose vehicles), which reduces the F1 scores. The architectural efficiency and the multi-scale feature extraction capabilities of InceptionV3 and MobileNetV2 allow them to show competitive performance with 0.91 accuracy. Both models are not consistent across some classes, such as in terms of class-wise balance and recall. For example, the classification of multi-purpose vehicle categories and city cars indicates that individual models may not capture subtle differences between these classes. The proposed HFFN model outperforms all baseline models, yielding an accuracy of 0.97. This improvement stems from combining complementary feature representations from multiple CNN architectures. The proposed model uses feature-level fusion to learn both global and fine-grained visual patterns, leading to better discrimination across all vehicle types. One of the interesting points is the consistent improvement of the recall and F1 score for all classes in the proposed model. The Multi-Purpose Vehicle category, which was hard for the individual models, is balanced with good precision and recall (0.94), thus showing better reliability in classification. Likewise, the fused feature representation is also strong in categories like Big Truck, Truck, and Van, with near-perfect performance.

The training and validation accuracy and loss curves of all CNN models and the proposed HFFN are presented in [Figures 5-11](#). These plots offer valuable insights into the convergence behavior, learning stability, and generalization capabilities of various models. Both baseline models (VGG16 and VGG19) show stable learning curves, with both training and validation accuracy increasing smoothly and loss decreasing smoothly with a corresponding curve. The gap between the training and validation curves is relatively small, suggesting good generalization and limited overfitting. However, their convergence is slower than that of more advanced architectures because their feature-extraction mechanisms are comparatively simpler. The

Table 1. Comparative performance analysis of base models and proposed HFFN

Models	Metrics	Classes							Macro average	Weighted average
		Big truck	City car	Multi-purpose vehicle	Sedan	Sport utility vehicle	Truck	Van		
VGG16	Precision	1.00	0.79	0.98	0.93	0.90	1.00	0.99	0.94	0.91
	Recall	0.98	0.98	0.53	0.95	0.96	0.99	0.98	0.91	0.90
	F1-score	0.99	0.88	0.69	0.94	0.93	1.00	0.98	0.91	0.89
	Accuracy	0.90								
VGG19	Precision	0.95	0.94	0.71	0.92	0.90	1.00	0.96	0.91	0.90
	Recall	0.98	0.78	0.92	0.90	0.91	0.97	0.98	0.92	0.89
	F1-score	0.97	0.85	0.80	0.91	0.90	0.98	0.97	0.91	0.89
	Accuracy	0.89								
DenseNet121	Precision	0.97	0.91	0.89	0.92	0.92	1.00	0.97	0.94	0.93
	Recall	0.98	0.91	0.86	0.95	0.95	0.97	0.96	0.94	0.93
	F1-score	0.97	0.91	0.88	0.93	0.93	0.98	0.97	0.94	0.93
	Accuracy	0.93								
DenseNet169	Precision	0.98	0.81	0.75	0.84	0.84	0.98	0.96	0.88	0.85
	Recall	0.98	0.83	0.69	0.90	0.81	0.99	0.95	0.88	0.85
	F1-score	0.98	0.82	0.72	0.87	0.82	0.99	0.96	0.88	0.85
	Accuracy	0.85								
InceptionV3	Precision	0.98	0.94	0.77	0.92	0.89	1.00	0.98	0.93	0.91
	Recall	1.00	0.83	0.90	0.92	0.92	0.99	0.99	0.94	0.91
	F1-score	0.99	0.88	0.83	0.92	0.91	1.00	0.99	0.93	0.91
	Accuracy	0.91								
MobileNetV2	Precision	0.94	0.86	0.84	0.92	0.94	1.00	1.00	0.93	0.91
	Recall	1.00	0.92	0.86	0.87	0.90	0.95	0.95	0.92	0.91
	F1-score	0.97	0.89	0.85	0.90	0.92	0.97	0.97	0.92	0.91
	Accuracy	0.91								
HFFN	Precision	1.00	0.97	0.94	0.97	0.97	1.00	1.00	0.98	0.97
	Recall	1.00	0.96	0.94	0.99	0.96	1.00	0.99	0.98	0.97
	F1-score	1.00	0.97	0.94	0.98	0.96	1.00	0.99	0.98	0.97
	Accuracy	0.97								

HFFN: Hybrid feature fusion network.

training and validation curves on DenseNet121 are very similar, indicating the network's ability to converge quickly and maintain stable performance. This behavior may be explained by its high connectivity, which provides better feature reusability and gradient flow. DenseNet169 [Figure 8] shows slightly larger variations in validation loss, suggesting that deeper networks can be less stable and more prone to overfitting. The training and validation curves are both smooth and show little divergence, indicating that InceptionV3 effectively extracts multi-scale features. Likewise, MobileNetV2's lightweight architecture enables fast convergence and efficient learning. However, small variations in validation loss suggest sensitivity of the data to variability and augmentation. The proposed HFFN model is better in terms of training behavior than all the individual models. It provides better convergence, higher accuracy, and much lower loss values with a minimal difference between the training and validation curves. This means that the generalization capability is high and the learning of the discriminative features is effective. The accuracy curve remains stable, and the

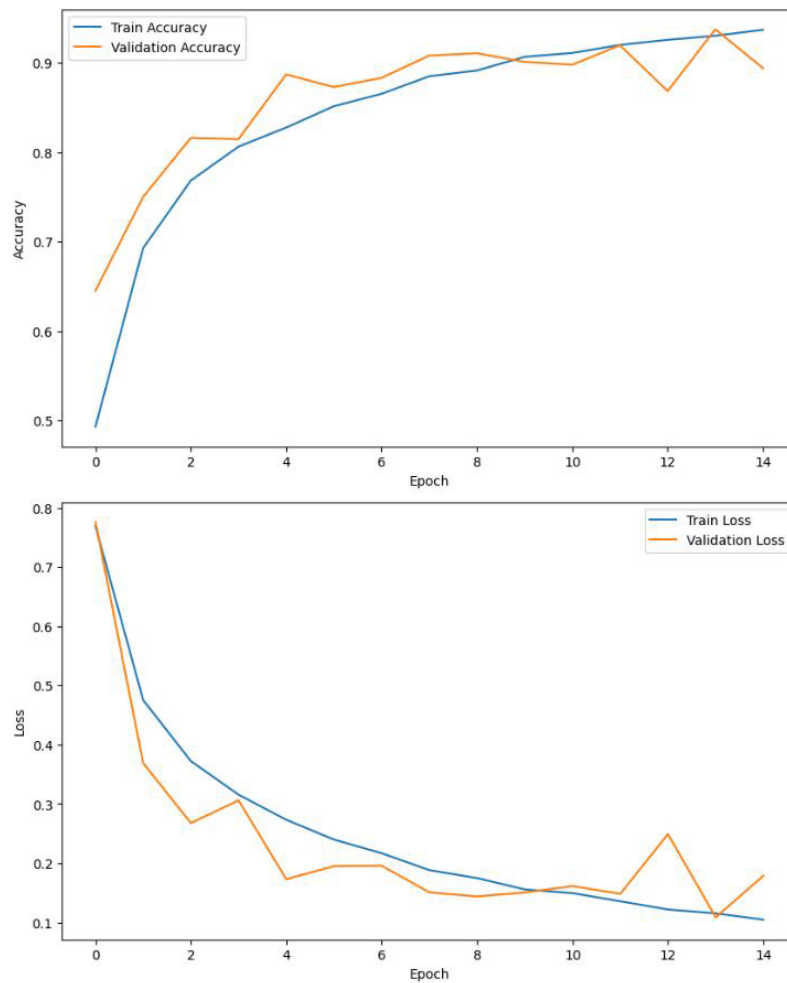


Figure 5. Model accuracy vs. model loss (VGG16).

loss curve also stabilizes, indicating that the model effectively prevents overfitting by incorporating feature-level fusion and regularization methods. The HFFN model is able to converge to a higher level of accuracy in a more consistent way than individual architectures, suggesting the advantage of fitting together complementary feature representations.

The confusion matrices of all individual CNN models and the proposed HFFN are shown in [Figures 12-18](#) to provide detailed knowledge of the prediction performance in the various classes and the misclassification pattern. The baseline models show different levels of misclassification between closely related vehicle classes. For VGG16 and VGG19, misclassification occurs between city car and multi-purpose vehicle, and between sedan and sport utility vehicle. Some of these errors stem from the fact that these architectures cannot sufficiently represent overlapping visual attributes such as size, shape, or structural similarity. Furthermore, the relatively high misclassification rate for multi-purpose vehicles suggests challenges in distinguishing vehicles with various medium-sized characteristics. DenseNet121 shows higher class-wise discrimination, with less confusion across most classes. A low misclassification rate between city cars and multi-purpose vehicles still exists, even with improved feature propagation, indicating similarities between these two classes that remain difficult to classify. Although DenseNet169 performs well in several categories, the deeper network also exhibits greater confusion, especially for the City Car, Sedan, and Multi-Purpose Vehicle categories, suggesting that increasing network depth will not yield additional gains in discriminative power. InceptionV3 and MobileNetV2 perform relatively well and show better class separation, although some

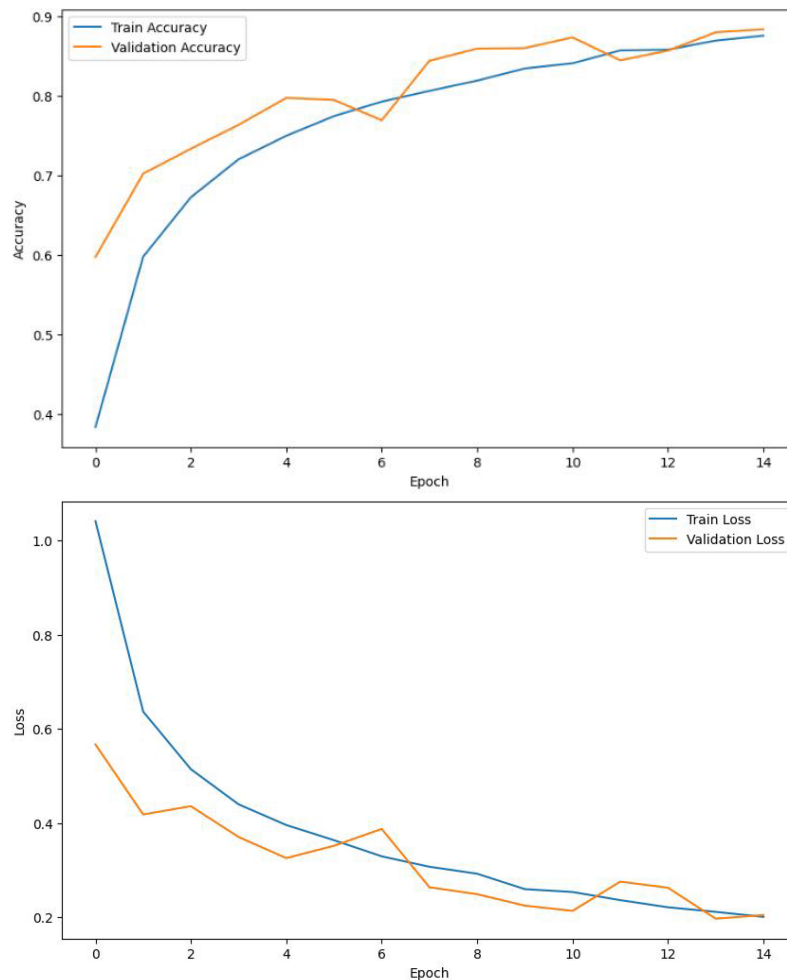


Figure 6. Model accuracy vs. model loss (VGG19).

confusion remains among visually similar vehicle categories. The results emphasize the need for using multiple architectures to obtain a fuller representation of complex visual patterns in every class. The proposed HFFN model shows a substantial decrease in misclassification in all of the categories. The confusion matrix is predominantly diagonal with few cross-class errors, demonstrating high accuracy in predictions. In particular, there is a significant improvement in the classification accuracy for the previously difficult categories, multi-purpose vehicle and city car, proving that nuanced visual differences are successfully captured by the feature-level fusion. Likewise, excellent classification results are obtained for the categories of Big Truck, Truck, and Van, showing good performance both on larger vehicles and in the more structurally different classes.

Figures 19 and 20 show the ROC curve and precision–recall (PR) curve, respectively, for the proposed HFFN, detailing the evaluation of the classification performance for all vehicle categories. The ROCs show that the HFFN model has good discriminatory performance, with all class-wise ROC curves converging toward the top-left corner of the graph. Near-perfect separability ($AUC \approx 1.00$) is achieved by the near-perfect performances by the class with the largest object, Big Truck, and by the classes Truck and Van, for which the AUC values are consistently high. Likewise, sedans ($AUC = 0.98$) and sport utility vehicles ($AUC = 0.97$) yield strong results, showing that the model can effectively capture both structural and contextual features. For City Car ($AUC = 0.95$) and Multi-Purpose Vehicle ($AUC = 0.93$), the AUC values are slightly lower, indicating there is a moderate overlapping feature space, as these vehicles are visually similar. Overall, the

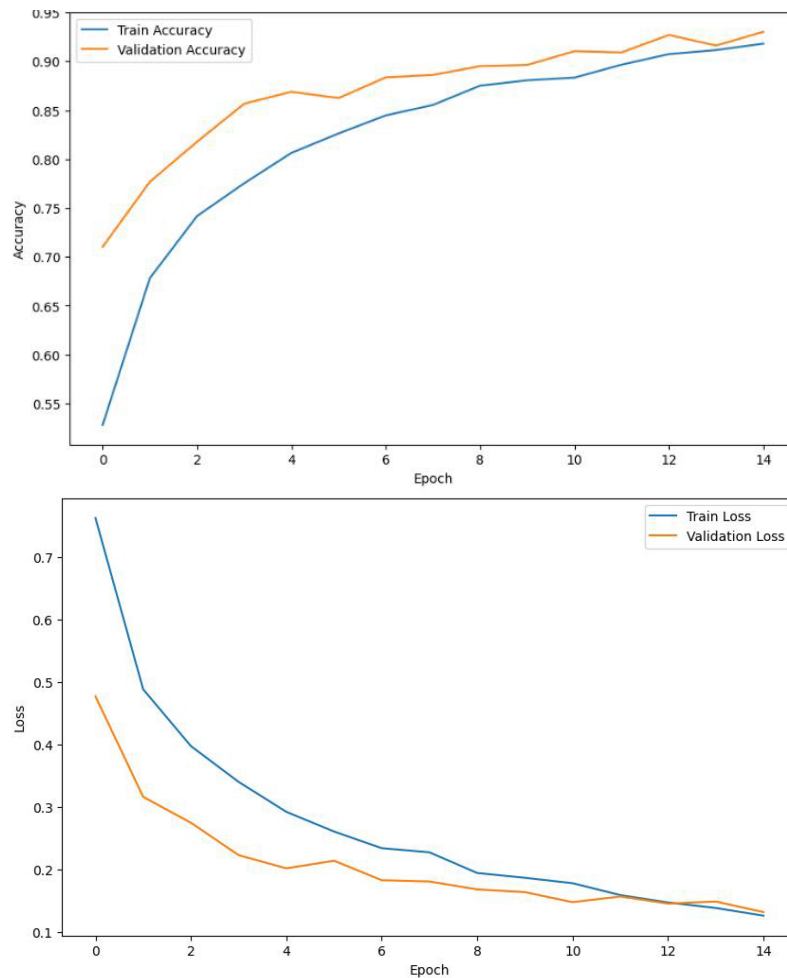


Figure 7. Model accuracy vs. model loss (DenseNet121).

ROC performance indicates that the HFFN model is robust and highly sensitive. The PR curves further validate the proposed model's effectiveness, especially with imbalanced classes. Truck and Van achieve high precision across most of the recall range [e.g., average precision (AP) = ~0.99], with near-perfect performance in most cases. At a higher recall level (AP = 0.98), Big Truck is able to perform quite well. The performance of the multi-purpose vehicle is comparatively low (AP = 0.76), and its precision decreases gradually with the increase of recall; this class has a relatively high similarity with other vehicle types, and this is very common in classification problems. City Car and Sport Utility Vehicle show moderate performance (AP ≈ 0.89), with a typical PR trade-off.

Explainable artificial intelligence (XAI) methodologies, such as gradient-weighted class activation mapping (Grad-CAM) and LIME, improve the interpretability and transparency of the proposed HFFN. These techniques provide visual and feature-level insights into the model's decision processes, improving understanding of how predictions are made. Grad-CAM identifies the most influential regions in an image that a model uses to make its prediction.

In this work, the DenseNet121 architecture is used, leveraging the feature maps of the last convolutional layer and gradient information to create class-specific heatmaps. These heatmaps represent areas that are most important for determining the predicted class. For the Sedan class, which has a high confidence score of 87.40%, as seen in [Figure 21](#), the model's attention is directed mainly towards the structure and central

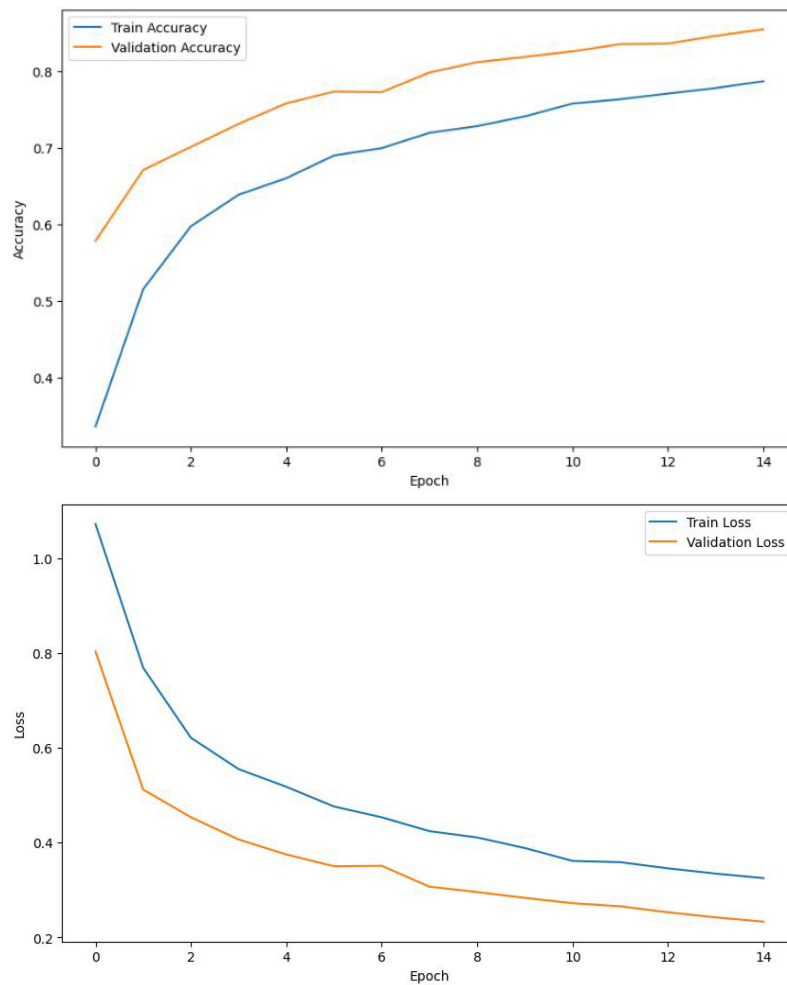


Figure 8. Model accuracy vs. model loss (DenseNet169).

body of the vehicle. The highlighted region closely matches the actual region of interest, indicating that the model learns meaningful spatial features rather than irrelevant background information.

Similarly, [Figure 22](#) shows another example of the model achieving an excellent confidence level of 99.68% when predicting a Big Truck. The Grad-CAM visualization shows a strong focus on the truck’s front structure and cabin, highlighting the model’s ability to identify distinct characteristics like shape and size. The bounding region created from the heatmap further confirms that the model attends to the region of interest associated with the vehicle object, which supports the reliability of the classification.

In addition to Grad-CAM, LIME provides model-agnostic explanations that highlight important superpixels affecting predictions. LIME perturbs input regions and observes changes in prediction to identify important areas, while Grad-CAM is based on deep feature activations. The LIME explanation shows the most important parts of the car that are classified as “good” [[Figure 23](#)].

The selected superpixels capture structural boundaries and salient visual features of the vehicle, further supporting the model’s ability to predict based on relevant features. Grad-CAM together with LIME allows for complementary interpretability. Unlike Grad-CAM, which explains attention at the convolutional-layer level, LIME can explain at the pixel or region level and is more interpretable to humans. The high inter-

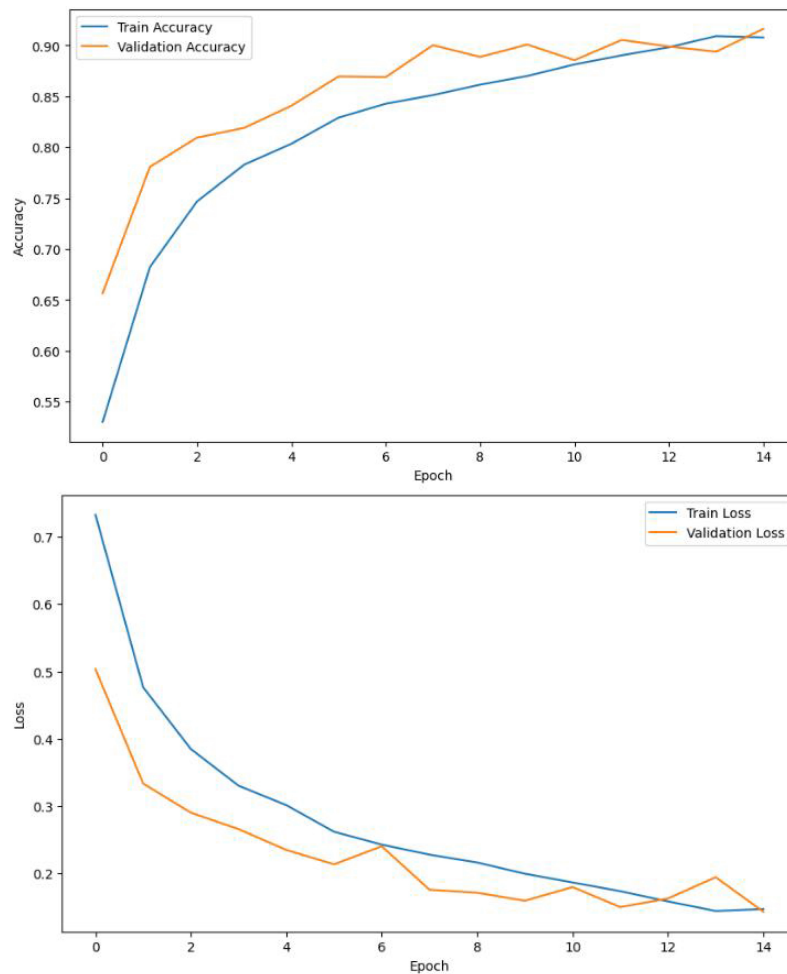


Figure 9. Model accuracy vs. model loss (InceptionV3).

method consistency shows that the proposed model of HFFN is making decisions based on the semantically meaningful features of text and not noise or background artifacts.

A robustness analysis subjects the proposed HFFN to controlled perturbations to the input images, such as Gaussian noise, image blurring, and image rotation, to assess the network's robustness in real-world situations. The transformations model real-life problems like noise in the sensors, motion blur, and changes of viewpoint usually faced in ITS. In the original image, the prototype classification of City Car is correctly classified as a city car with a high confidence level of 98.42%, as shown in [Figure 24](#), indicating strong baseline performance. With the addition of Gaussian noise, the model incorrectly classifies the image as a sport utility vehicle, with a lower confidence of 73.35%. This degradation reveals the model's sensitivity to high-frequency noise, which can cause errors in fine-grained visual features and misclassification, especially between visually similar classes. If the image is blurred, the model still predicts the class with 90.14% confidence, though slightly lower. This suggests the overall vehicle structure remains recognizable even when fine details are missing, explaining the model's robustness to low-frequency distortions. Likewise, the model preserves the correct classification with 96.38% confidence under the rotation transformation, indicating good tolerance to this geometric perturbation in the evaluated example. The results indicate that the perception model demonstrates good tolerance to blur and rotation in the evaluated example, whereas Gaussian noise causes a notable degradation and incorrect classification. These findings suggest that the model has robustness to some perturbations but remains sensitive to high-frequency noise.

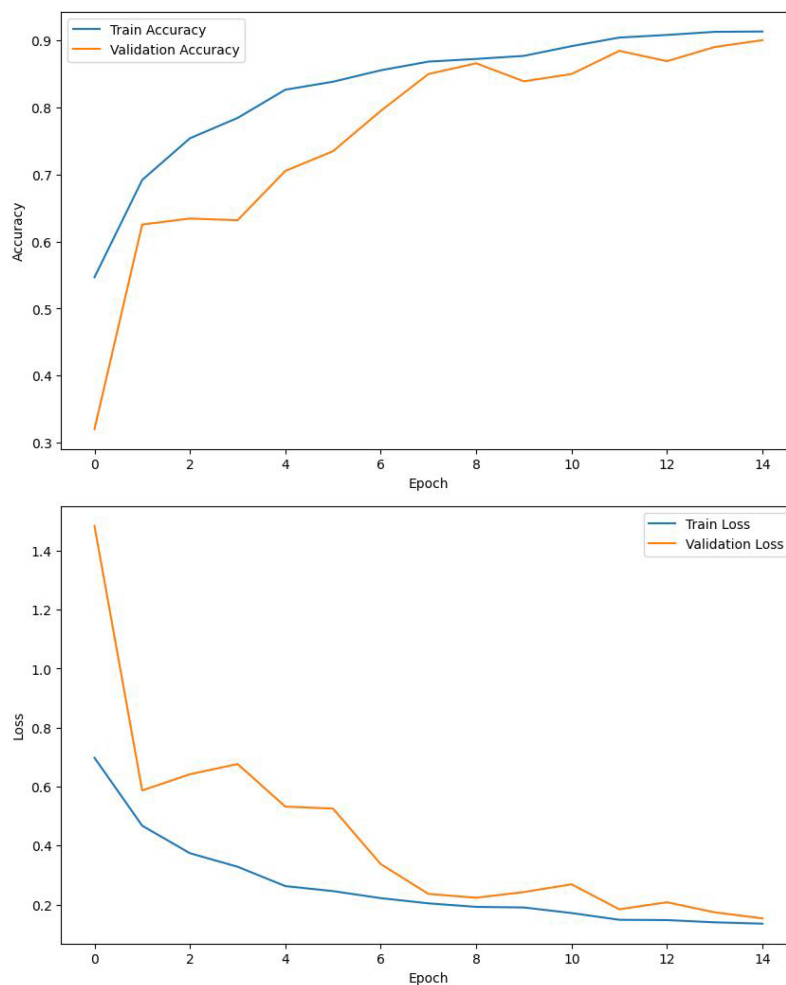


Figure 10. Model accuracy vs. model loss (MobileNetV2).

5.2. Stage 2: IoV-based collision risk prediction

For collision risk prediction in an IoV environment, a stacked ensemble learning scheme uses four machine learning models stacked in a hierarchical prediction architecture. A novel architecture, hybrid stacked ensemble architecture for collision prediction (HSEA-CP), was designed to improve prediction accuracy by combining the complementary strengths of multiple base learners. The first step in pre-processing the IoV data was to normalize the column names, then replace missing values and convert categorical data into numerical values using one-hot encoding. The target variable (collision occurrence) was coded in binary format. The data were then split into training and test sets using a stratified split to maintain the class distribution and allow for unbiased evaluation. The first layer of the proposed architecture uses four heterogeneous base models: LR, SVM, MLP, and XGB. The models complement each other in their learning capabilities, from linear modeling (LR) and margin-based classification (SVM) to nonlinear representation learning (MLP) and gradient boosting (XGB). All models are trained independently based on each particular IoV feature, including vehicle velocity, inter-vehicle distance, braking capability, driver status, environmental conditions, and road characteristics. Feature alignment based on the pre-defined feature column sets, and scaling, when necessary, were performed to ensure consistency across the models. Each base model outputs collision likelihoods. Each probability score is a meta-parameter that represents the confidence of each base learner in its prediction. Besides these meta-features, the original feature set of IoV is concatenated to create an enriched feature representation. The stacking framework uses an independent training-testing protocol where the data is first split into independent training and testing sets. For the Stage 2 comparative

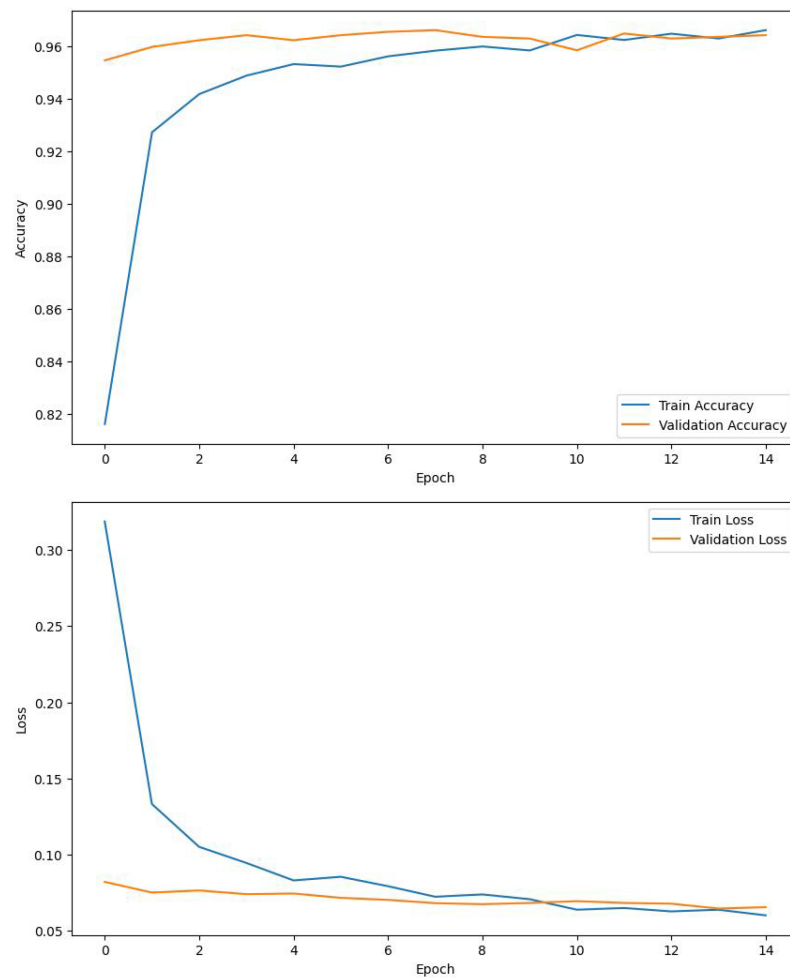


Figure 11. Model accuracy vs. model loss (HFFN). HFFN: Hybrid feature fusion network.

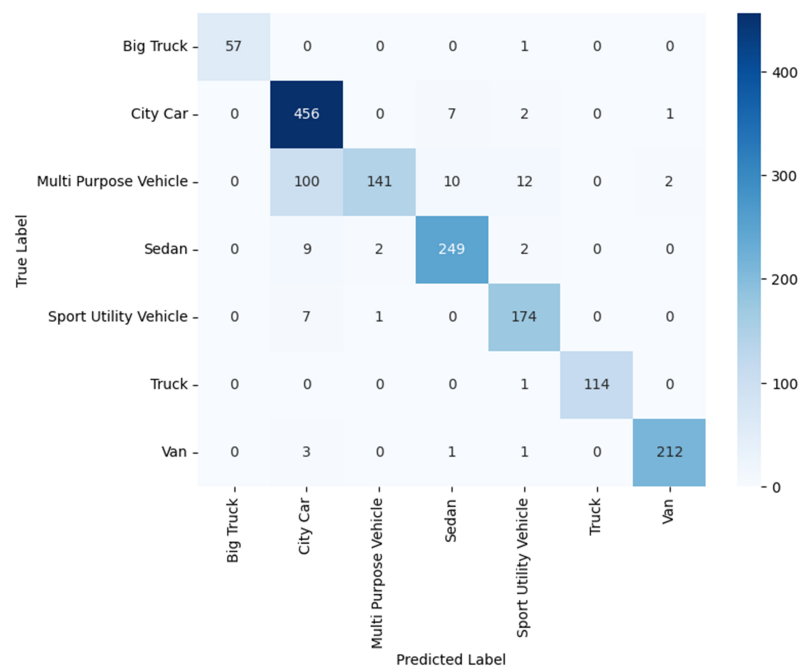


Figure 12. Confusion matrix (VGG16).

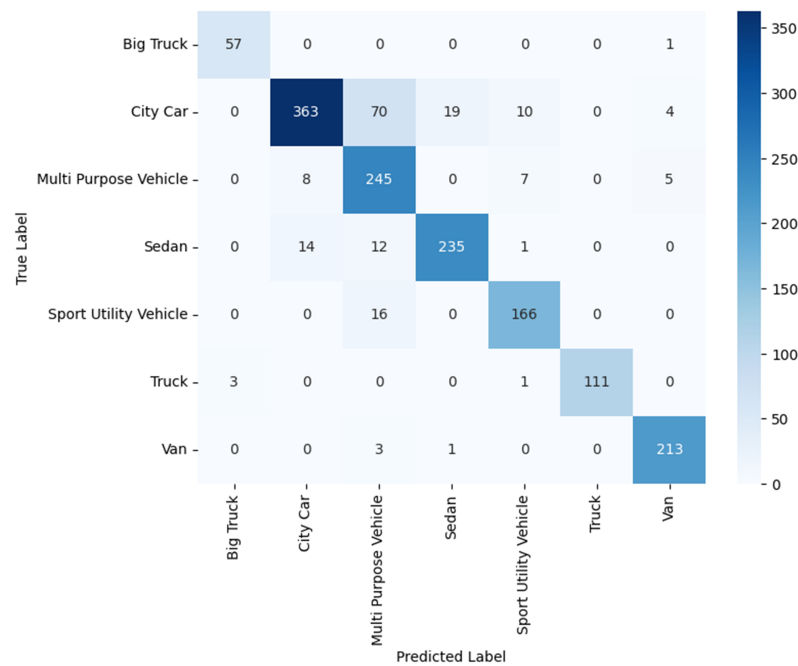


Figure 13. Confusion matrix (VGG19).

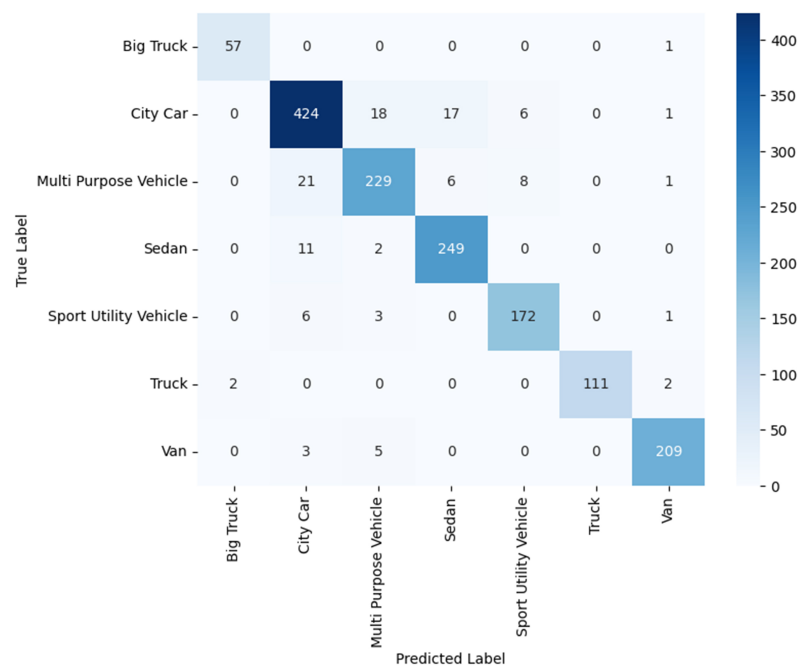


Figure 14. Confusion matrix (DenseNet121).

evaluation reported in Table 2, the base learners were trained and evaluated within the Stage 2 collision-prediction configuration. Their prediction probabilities are then fused with the enhanced feature set to form the meta-feature space, which is then fed into the random forest meta-classifier. In the model evaluation phase, the test subset is never used in any part of the training process, from feature augmentation to base-model learning to meta-classifier training. This evaluation protocol guarantees that the reported performance is based on unseen data and provides an accurate measure of the predictive ability of the framework. The hybrid architecture allows for a second-level model to leverage raw sensor information

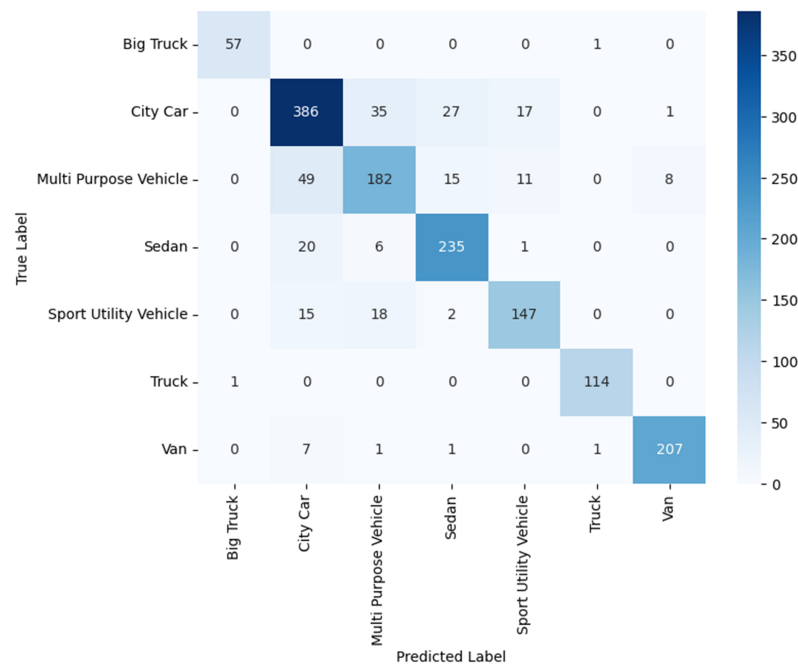


Figure 15. Confusion matrix (DenseNet169).

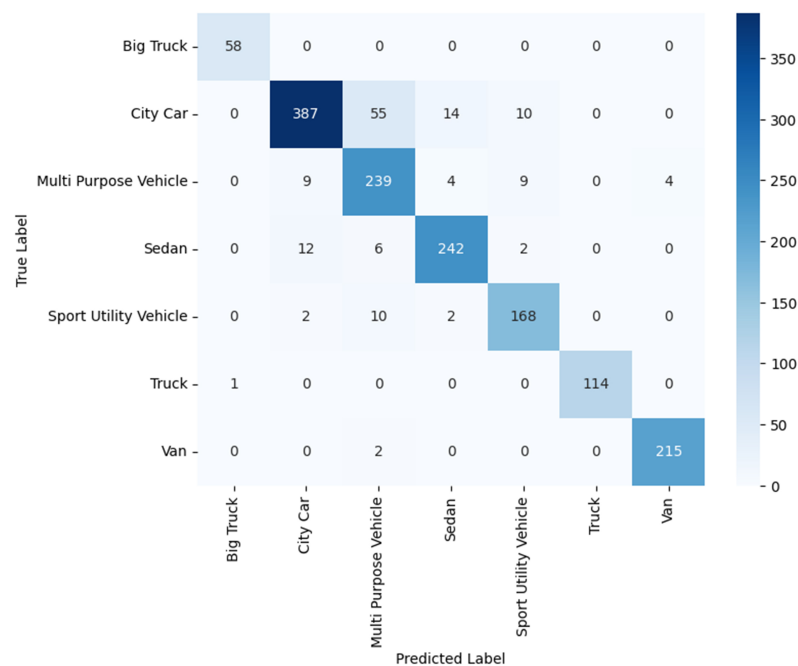


Figure 16. Confusion matrix (InceptionV3).

and learned predictive signals from base models. Finally, a meta-classifier is trained using a random forest algorithm with the combined feature vector. The meta-model balances the classes and sets the model depth to ensure generalization and combat class imbalance. During the testing phase, the trained base learners generate prediction probabilities for the unseen test samples, which are concatenated with the corresponding augmented IoV features and provided as input to the trained random forest meta-classifier for final collision risk prediction. This multi-level hierarchical architecture captures complex



Figure 17. Confusion matrix (MobileNetV2).

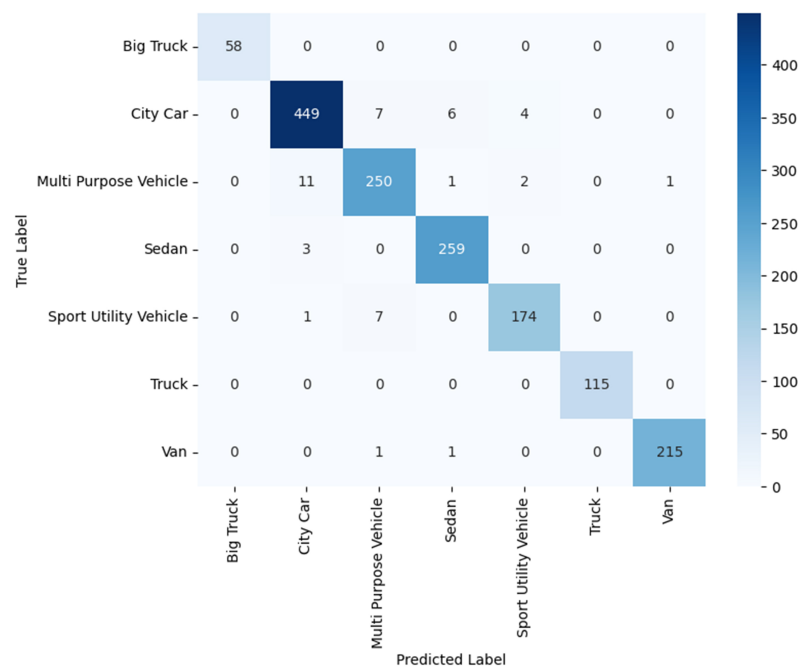


Figure 18. Confusion matrix (HFFN). HFFN: Hybrid feature fusion network.

relationships in the data while overcoming the limitations of any single model. The proposed HSEA-CP is hence a strong and scalable approach for collision risk prediction for IoV environments.

A detailed performance comparison of individual base models such as XGB, LR, SVM, and MLP with the proposed HSEA-CP is shown in [Table 2](#). The evaluation process is performed in terms of precision, recall, F1-score, and accuracy. Results show that individual base models perform inconsistently across evaluation

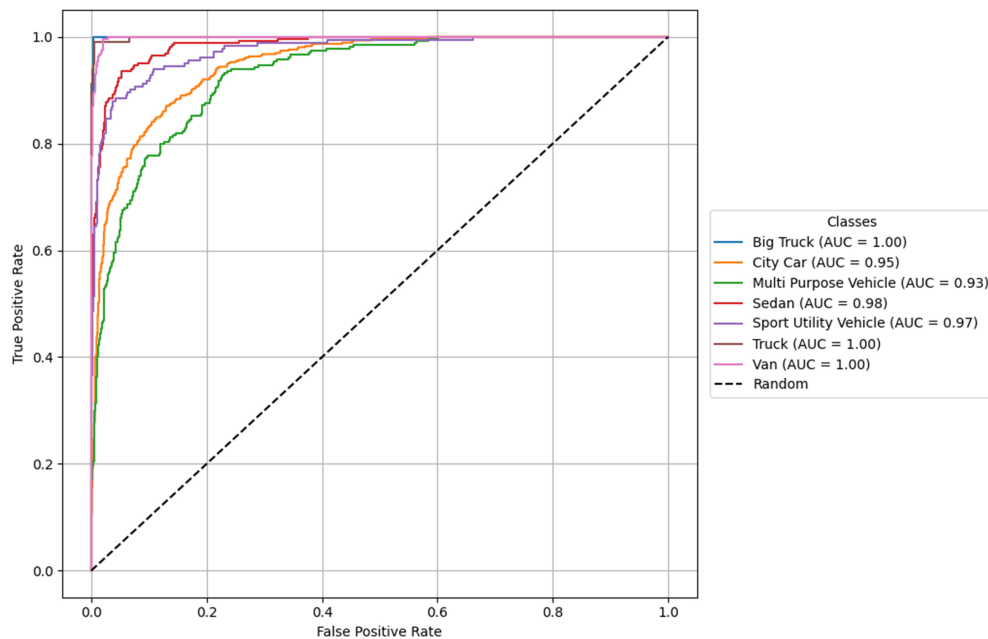


Figure 19. ROC curve for proposed HFFN. ROC: Receiver operating characteristic; HFFN: hybrid feature fusion network; AUC: area under the curve.

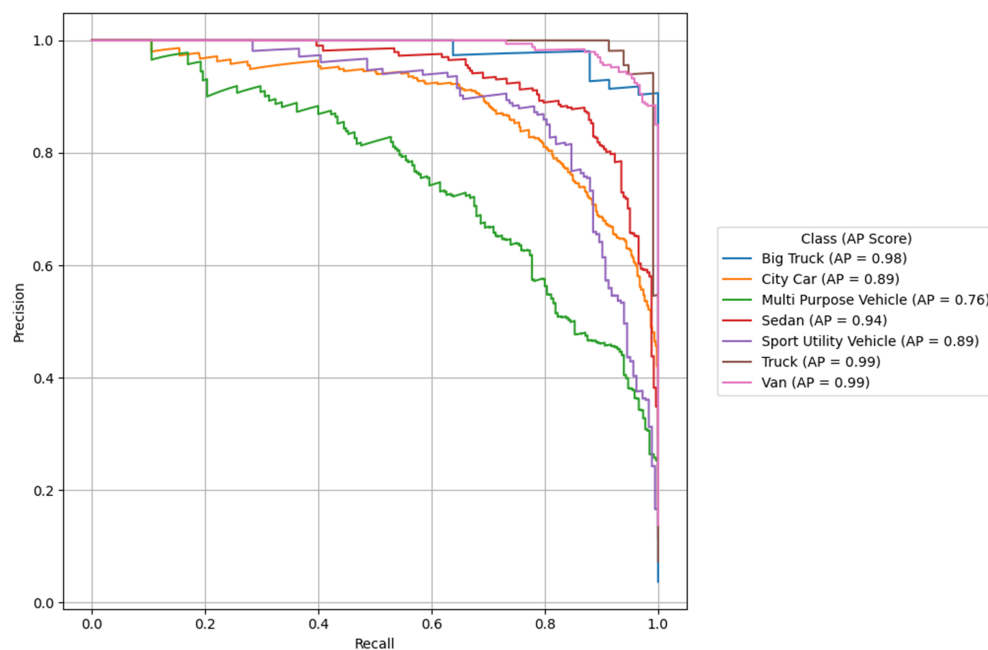


Figure 20. PR curve for proposed HFFN. PR: Precision–recall; HFFN: hybrid feature fusion network; AP: average precision.

measures, though generally well. The LR, SVM, and XGB models have relatively high precision values for both classes, with the Collision class having perfect precision (1.0000) for the LR and SVM models. Though these models have low recall, especially in the No Collision class, which has the lowest recall (0.5765) among all models compared, XGB has this class value in a comparatively low range. This imbalance indicates that although base models are confident in their predictions, they do not necessarily include all instances, resulting in lower sensitivity. The MLP model shows relatively balanced precision and recall; however, it is not always the best performer compared with the others. Although its F1-scores are still moderate,



Figure 21. Grad-CAM visualization for sedan classification highlighting model attention regions (The original vehicle samples are selected from the Vehicle Images Dataset^[26]). Grad-CAM: Gradient-weighted class activation mapping.



Figure 22. Grad-CAM visualization for big truck classification demonstrating high-confidence feature localization (The original vehicle samples are selected from the Vehicle Images Dataset^[26]). Grad-CAM: Gradient-weighted class activation mapping.

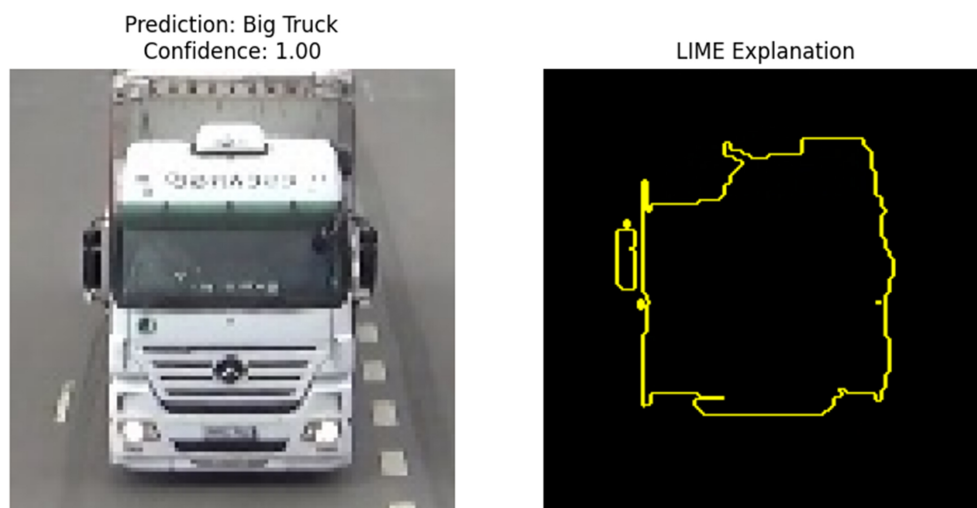


Figure 23. LIME-based explanation highlighting key superpixel regions influencing model prediction for big truck classification (The original vehicle samples are selected from the Vehicle Images Dataset^[26]). LIME: Local interpretable model-agnostic explanations.

suggesting that its ability to balance false positive and false negative rates is mediocre. These observations show that individual model performance is not consistent across all evaluation metrics and classes. The proposed HSEA-CP achieves an accuracy of 0.9847, with macro-averaged precision, recall, and F1-score of

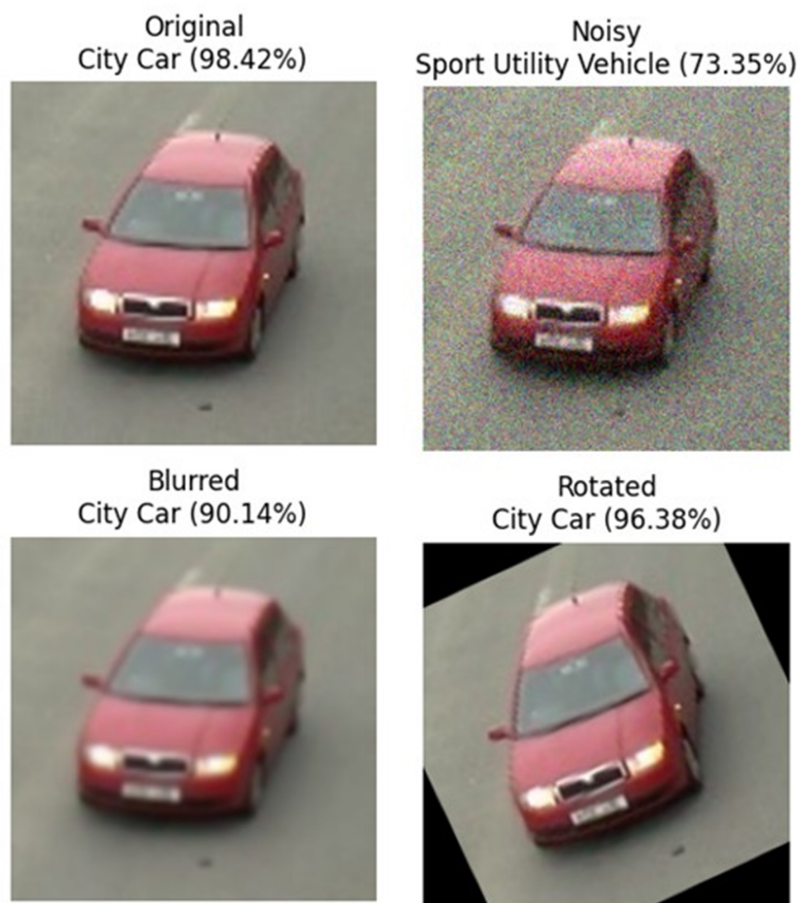


Figure 24. Robustness analysis of the proposed HFFN architecture under noise, blur, and rotation transformations (The original vehicle samples are selected from the Vehicle Images Dataset^[26]). HFFN: Hybrid feature fusion network.

0.9784, 0.9884, and 0.9831, respectively. Its weighted precision, recall, and F1-score are 0.9853, 0.9847, and 0.9847, respectively. Because HSEA-CP uses a hybrid stacking mechanism based on the probabilistic outputs of several base learners and the original IoV features, it achieves improved performance. This integration allows the model to learn the underlying relationships between the features as well as each model's confidence level, leading to improved decision-making.

The confusion matrices of individual base models (XGB, LR, SVM, and MLP) and the proposed HSEA-CP are shown in Figures 25–29. The following matrices show a detailed, class-wise assessment of prediction performance and the number of each type of prediction made. The XGB model performs well at identifying collision instances, with a high number of true positives. However, it has a high false positive rate, misclassifying “no collision” as “collision”. This indicates a bias toward predicting collisions, which can lead to unnecessary alerts in practice. In LR, no-collision instances are correctly classified, but false negatives are relatively high; that is, actual collision cases are classified as no collision. This suggests the model might be conservative in predicting collisions but may not capture critical risk scenarios, limiting its usefulness in safety-critical situations. Likewise, the SVM correctly classifies no-collision instances but misclassifies many collision instances. This imbalance indicates limited ability to handle more complex nonlinear relationships in IoV data and, therefore, low sensitivity towards collisions.

The MLP model has balanced performance with better detection of collision cases than LR and SVM models. However, it has a moderate number of false positives and false negatives, suggesting that predictive performance is affected by data variation and feature complexity. By contrast, the proposed HSEA-CP model

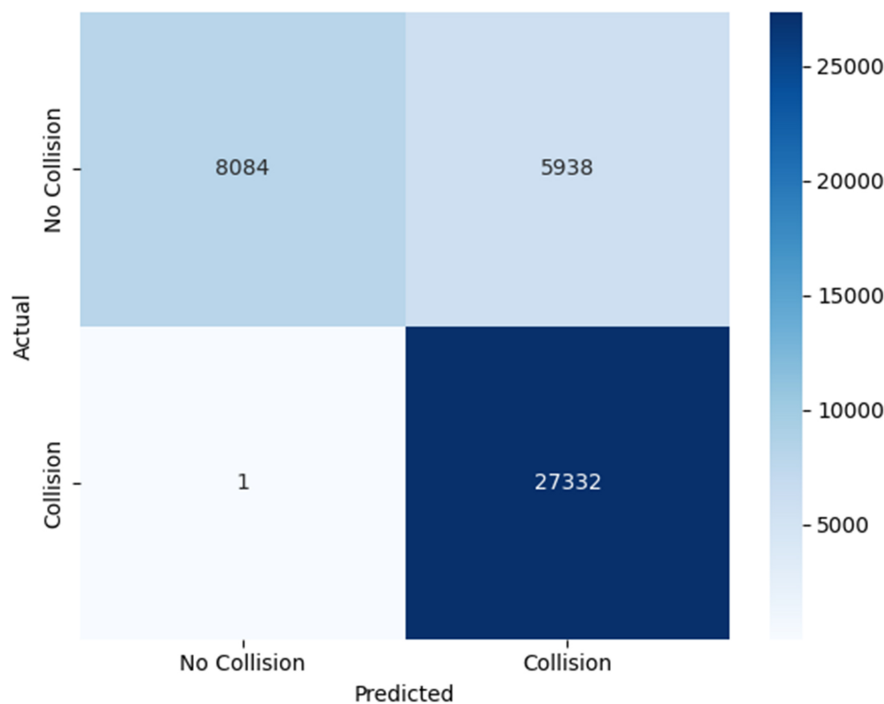


Figure 25. Confusion matrix (XGB). XGB: XGBoost.

Table 2. Comparative performance analysis of base models and proposed HSEA-CP

Models	Metrics	Classes		Macro average	Weighted average
		No collision	Collision		
XGB	Precision	0.9999	0.8215	0.9107	0.8820
	Recall	0.5765	1.0000	0.7882	0.8564
	F1-score	0.7314	0.9020	0.8167	0.8441
	Accuracy	0.8564			
LR	Precision	0.8484	1.0000	0.9242	0.9486
	Recall	1.0000	0.9083	0.9542	0.9394
	F1-score	0.9180	0.9520	0.9350	0.9404
	Accuracy	0.9394			
SVM	Precision	0.8482	1.0000	0.9241	0.9485
	Recall	1.0000	0.9082	0.9541	0.9393
	F1-score	0.9179	0.9519	0.9349	0.9404
	Accuracy	0.9393			
MLP	Precision	0.8782	0.9391	0.9087	0.9185
	Recall	0.8816	0.9373	0.9095	0.9184
	F1-score	0.8799	0.9382	0.9091	0.9185
	Accuracy	0.9184			
HSEA-CP	Precision	0.9567	1.0000	0.9784	0.9853
	Recall	1.0000	0.9768	0.9884	0.9847
	F1-score	0.9779	0.9883	0.9831	0.9847
	Accuracy	0.9847			

HSEA-CP: Hybrid stacked ensemble architecture for collision prediction; XGB: XGBoost; LR: logistic regression; SVM: support vector machine; MLP: multi-layer perceptron.

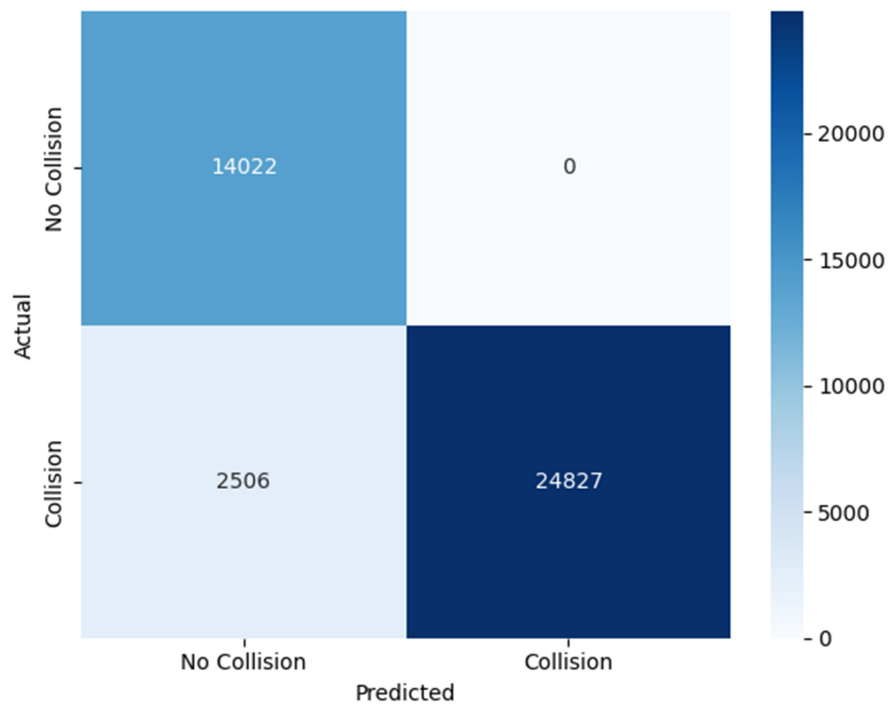


Figure 26. Confusion matrix (LR). LR: Logistic regression.

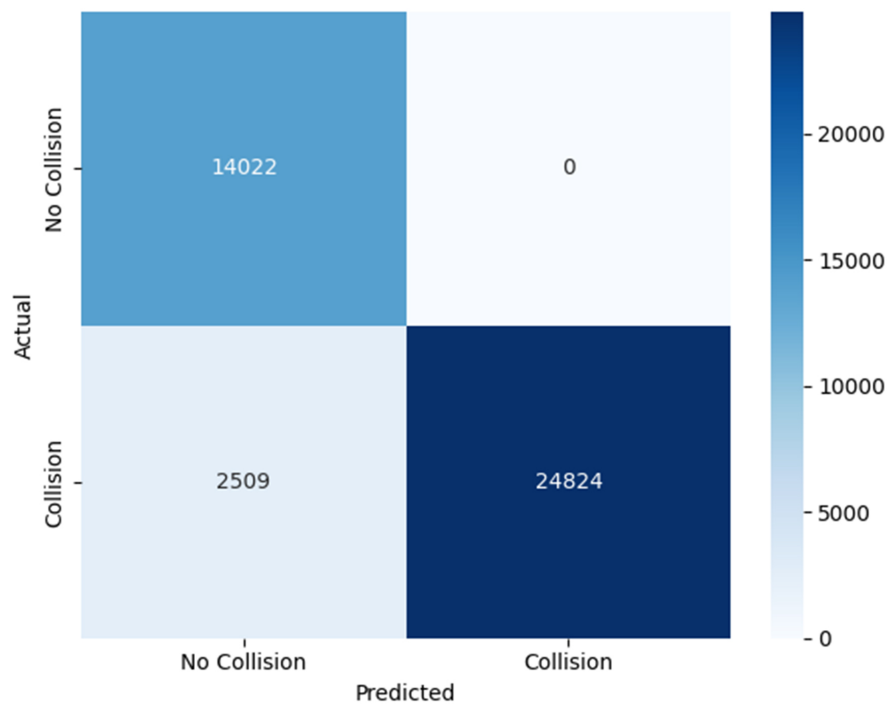


Figure 27. Confusion matrix (SVM). SVM: Support vector machine.

shows a much better balance between the two classes. It has achieved a high rate of true positives and true negatives and significantly lowered the false positive and false negative rates. This balanced classification shows that the model captures the underlying patterns in the data and reduces the bias seen in individual models. The reduction in false negatives is especially important in collision prediction, where missing a

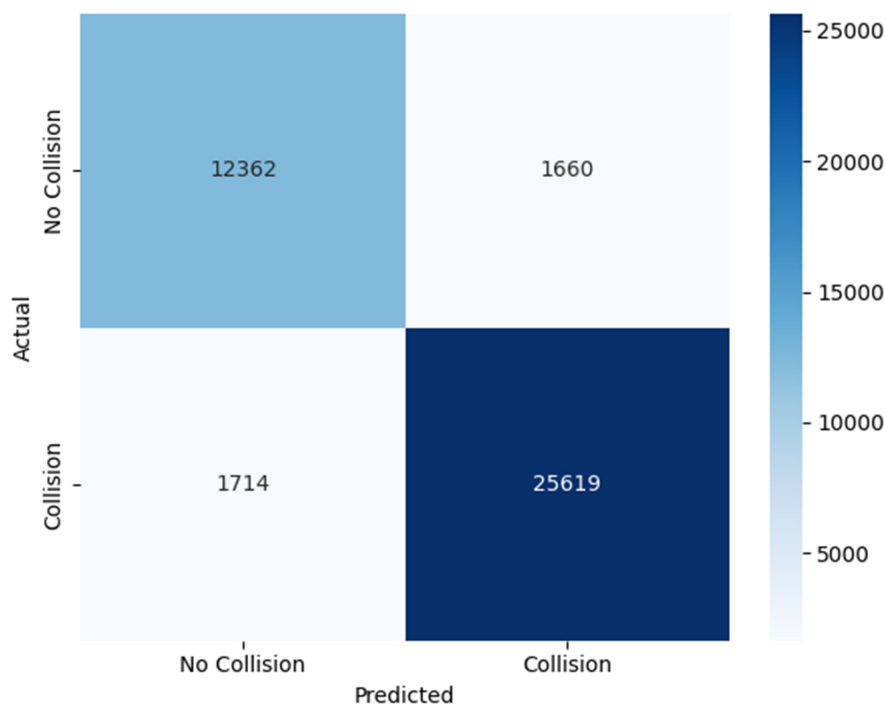


Figure 28. Confusion matrix (MLP). MLP: Multi-layer perceptron.

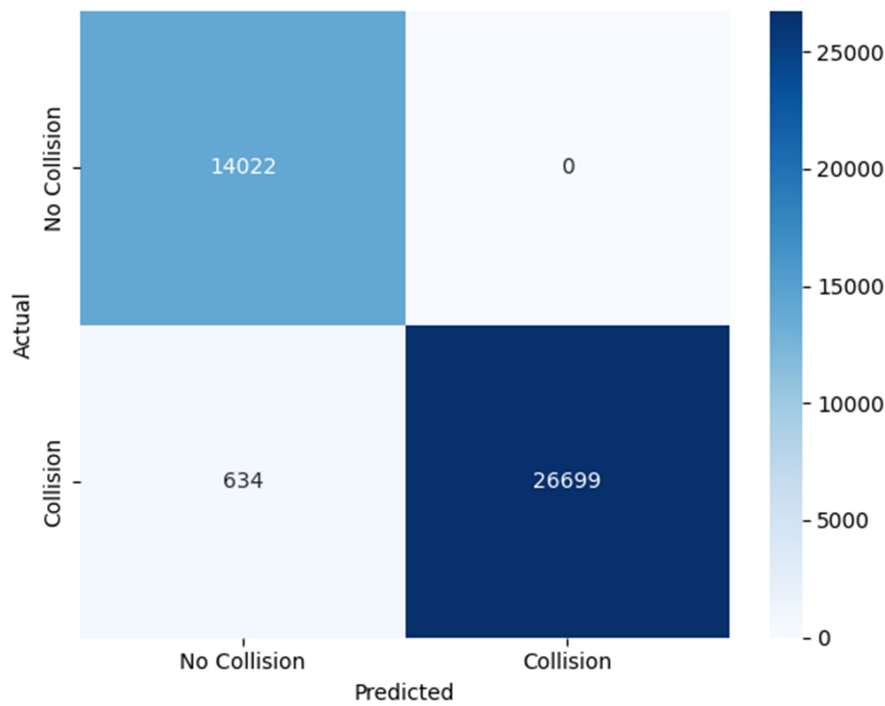


Figure 29. Confusion matrix (HSEA-CP). HSEA-CP: Hybrid stacked ensemble architecture for collision prediction.

collision event can have serious consequences. Meanwhile, the limited number of false positives prevents false alarms from being over-emitted. HSEA-CP performs better because its hybrid stacking procedure combines complementary predictions from different base models with original feature information.

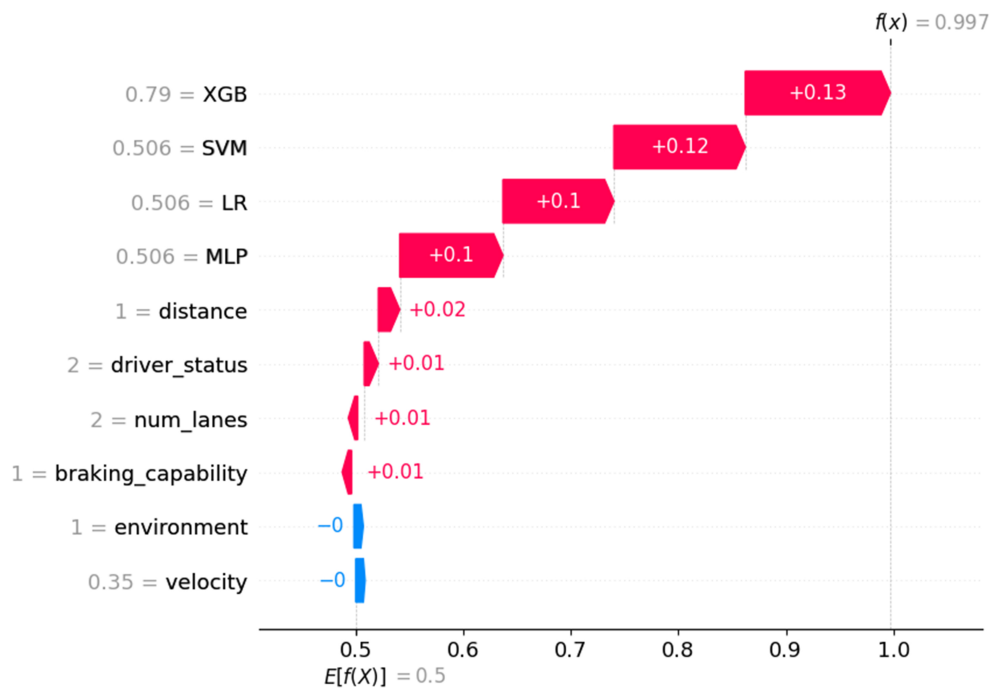


Figure 30. SHAP-based local explanation of HSEA-CP. SHAP: SHapley Additive exPlanations; HSEA-CP: hybrid stacked ensemble architecture for collision prediction; XGB: XGBoost; SVM: support vector machine; LR: logistic regression; MLP: multi-layer perceptron.

SHAP measure each feature's contribution to the model's prediction and enhance the interpretability of the proposed HSEA-CP. SHAP is a theoretically informed method for understanding complex models by quantifying each feature's contribution to the prediction. Figure 30 shows a local explanation for a sample test instance in which the model predicts the collision event with high accuracy ($P = 0.997$). The prediction starts at the baseline value of $E[f(X)] = 0.5$, and as the features increase, the output moves toward the final prediction. The analysis shows that the most influential contributing factors are the probabilistic outputs of the base models XGB, SVM, LR, and the MLP model. Among these, XGB contributes the most (+0.13), followed by SVM (+0.12), while LR and MLP each contribute +0.10. The stacking mechanism clearly exploits the predictive power of the base models, making them the main drivers of the final classification. Besides these meta-features, the original IoV features, including distance (+0.02), driver status (+0.01), number of lanes (+0.01), and braking capability (+0.01), also contribute positively to the prediction. These features provide context about driving conditions and interactions with vehicles. Other characteristics, such as velocity or environment, show very small or slightly negative contributions, suggesting they have little to no influence in this prediction case. The waterfall visualization shows how individual contributions can collectively move the prediction from the baseline towards a high-confidence collision outcome. It is a layered contribution pattern that shows the effectiveness of integrating meta-features with the raw IoV features in the same architecture.

5.3. Cross-stage feature alignment

The proposed two-stage design links visual perception (Stage 1) and IoV-based risk prediction (Stage 2), with image-derived features added to the collision prediction pipeline. The image classification model is converted into structured outputs and added to the IoV dataset, enabling a common learning framework to integrate visual and sensor data. The deep learning model produces class probabilities and confidence scores for each input image, which are added to the IoV dataset as additional features such as the predicted class label, confidence value, and complete probability distribution. This process adds visual context to the

Table 3. Feature importance ranking for collision prediction

Rank	Feature	Importance
1	Status of driver	0.4485
2	Number of lanes	0.2227
3	Distance between vehicles	0.1508
4	Braking capability	0.0850
5	Nature of environment	0.0764
6	img_prob_3	0.0028
7	img_confidence	0.0028

original dataset, making it more representative of real driving situations and improving the input-space representation capability. The extended dataset is then used in the proposed HSEA-CP, where features derived from the IoV and image are combined. These enriched features, along with probabilistic outputs from several base models, are combined in the stacking mechanism, allowing the meta-classifier to learn complex interactions between the visual context and the dynamics of the vehicles.

To specifically evaluate the effect of cross-stage feature alignment, an additional experiment was conducted by integrating the image-derived features with the structured IoV features. In this cross-stage integrated configuration, HSEA-CP achieved an accuracy of 0.9808. For the Collision class, the model achieved a precision of 0.9985, recall of 0.9721, and F1-score of 0.9851, while the corresponding values for the No Collision class were 0.9500, 0.9972, and 0.9730, respectively. The model also achieved macro-averaged precision, recall, and F1-score of 0.9742, 0.9847, and 0.9791, respectively, with weighted averages of 0.9817, 0.9808, and 0.9809. These results represent the cross-stage feature alignment experiment and differ from the primary Stage 2 comparative results reported in [Table 2](#). Integrating visual information provides complementary context to the IoV features and supports reliable prediction, particularly for collision events predicted with high confidence. Furthermore, feature importance analysis shows that features derived from the image are also useful for making the final prediction, which indicates that visual features are not simply redundant to other attributes of an IoV, velocity, distance, and driver status, but instead provide complementary information. The contribution analysis also shows that image-based and IoV features both play a significant role in the decision-making process, highlighting the effectiveness of multimodal fusion. System-wise, this is a cross-stage integration, connecting the perception with the decision-making layer in cyber-physical vehicle systems. The proposed framework can include visual intelligence in prediction models based on IoV, which allows for more realistic and context-aware collision risk assessment. In dynamic, changing environments, visual factors (including scene context and vehicle characteristics) are especially important for perceiving risk.

The top features that affect the prediction of collision risk in the proposed HSEA-CP are shown in [Table 3](#). The results show that IoV-related features play a major role in the prediction process. Among these, driver status is the most important, accounting for almost half of the overall importance (0.4485). This emphasizes the importance of human behavior in collision risk. The number of lanes (0.2227) and vehicle distance (0.1508) also significantly influence accidents, highlighting the importance of traffic density and the space between vehicles in accident conditions. The impact of vehicle dynamics and environmental conditions is further underscored by additional IoV characteristics like braking capability (0.0850) and nature of environment (0.0764). These findings align with real transportation systems: driver-related and environmental factors affect safety outcomes. Interestingly, other image-derived features like `img_prob_3` and `img_confidence` also appear among the top features, although their importance values are lower. Their presence in the top ranks also suggests that visual information plays a role in the decision-making process and helps make integrating image-based insights into the IoV prediction framework successful.

Table 4. Analysis of visual feature contributions

Rank	Feature	Importance
1	img_prob_3	0.0028
2	img_confidence	0.0028
3	img_prob_4	0.0024
4	img_prob_5	0.0013
5	img_prob_1	0.0012
6	img_prob_6	0.0012
7	img_prob_2	0.0010
8	img_prob_0	0.0008
9	vehicle_class	0.0007

Table 4 shows a detailed analysis of the contributions of image-derived features in the proposed HSEA-CP architecture. These results indicate that all the image features together have a smaller contribution than IoV features; the most significant image features are `img_prob_3` and `img_confidence`. These features reflect the model's confidence and the image classification probability distribution, suggesting that certain image patterns are important for predicting collision risk. The remaining image probability features (`img_prob_4`, `img_prob_5`, `img_prob_1`) contribute moderately but consistently, indicating that several visual cues work together during prediction. The relatively small size of the vehicle class indicates that, relative to the probabilistic representation of the class, which includes uncertainty and more contextual information, the raw class labels are less informative. The combined effects of the individual image features are meaningful, but the contribution is small. These features provide additional contextual information that supplements the IoV data, especially when IoV alone is insufficient to capture the environment's complexity. This shows that the proposed approach successfully integrates multimodal features: using image-derived features improves the model's contextual awareness without dominating the effect of IoV-derived features.

The proposed framework does not rely on naturally synchronized image and IoV observations collected from the same driving events. Instead, multimodal integration is realized through a cross-stage semantic feature transfer mechanism. The HFFN maps each vehicle image to a structured semantic descriptor consisting of the inferred vehicle class, a confidence score, and a full probability distribution over all vehicle classes. The method then incorporates these image-derived representations into the IoV dataset as additional explanatory variables through a structured feature augmentation process before training the collision prediction model. Thus, the proposed HSEA-CP leverages structured vehicular features and high-level visual context in the same learning space. This design is understood as a cross-modal semantic feature integration strategy rather than a direct end-to-end multimodal learning framework based on naturally paired multimodal datasets. The experimental results show that including transferred visual semantics provides complementary contextual information to the original IoV feature space and improves collision risk prediction.

The results show that the proposed two-stage visual perception and IoV-based analytics approach improves collision risk prediction. The first stage, which is based on deep learning, utilizes feature-level fusion of various CNN architectures to obtain high classification performance, and visual patterns are extracted from vehicle images with high capability. The second stage is a stacked ensemble architecture in which several machine learning models learn collision risk from IoV features to predict it accurately. The results show that each stage performs well overall: the perception stage provides accurate visual representations, while the prediction stage achieves high accuracy and balanced performance across evaluation metrics. More importantly, cross-stage integration provides a meaningful link between visual data and IoV data, allowing

multimodal feature integration in the unified prediction framework. The proposed approach introduces image-derived features into the IoV dataset, adding contextual information and improving prediction accuracy and reliability. While IoV features are the primary contributors, visual data also provides complementary information that helps inform model decisions, especially in complex scenarios. This research shows that the fusion of perception and decision-making layers can result in a more complete, accurate, and actionable collision prediction system, further supporting the system's potential applications in intelligent transportation and cyber-physical vehicle systems.

The computational cost of the proposed framework relies on the six CNN models in the perception phase. The LR, SVM, MLP, XGB, and random forest models in the prediction phase have relatively small computational requirements. CNN models are trained with pre-trained weights for inference and can run independently with parallel processing to reduce perception latency. Modular design also enables the use of lightweight CNN architectures or model-compression techniques where computational resources are limited. Thus, the framework can predict collision risk quickly in near real time by deploying appropriate edge and/or vehicular computing platforms that support GPUs. However, the real-time performance achieved depends on the capabilities of the hardware, image processing time, communication delay, sensor acquisition rate, and computational resources on the vehicle / edge node. The current study evaluates the framework on benchmark datasets; however, it does not include physical deployment or dedicated hardware latency measurements. Future work will explore such real-world validation, including an inference-time benchmark and implementation on edge devices, model compression, and optimization for resource-constrained vehicular platforms.

6. CONCLUSION

This study proposed a two-stage, context-aware meta-ensemble approach to collision risk prediction in the IoV context, combining image-based perception with IoV-driven predictive analytics in a CPS paradigm. The proposed method effectively bridges this perception-to-decision gap by combining deep learning-based visual feature extraction with a stacked ensemble learning strategy. Experimental results show that the feature fusion-based perception module achieves the best classification performance, and the hybrid stacked ensemble architecture greatly improves collision-prediction accuracy compared with the individual baseline models. Mixing heterogeneous base learners (LR, SVM, MLP, and XGB) with a random forest meta-learner allows the model to learn complex nonlinear relationships and improve its generalization ability. Furthermore, cross-stage integration of image-derived features with the IoV dataset adds contextual information, increasing robustness and predictive confidence. The study also shows that IoV features are the most dominant for prediction, while image-based features act as auxiliary features that enhance the decision-making process. The robustness analysis indicates that the perception module maintains correct predictions under blur and rotation in the evaluated example, while Gaussian noise causes noticeable degradation and misclassification, highlighting the model's sensitivity to high-frequency perturbations. Furthermore, the embedded explainable AI methods increase the interpretability and trust of the proposed system, crucial for safety-related applications. The proposed framework could be a scalable and effective solution to collision risk prediction in intelligent transportation systems. It plays a key role in advancing road safety by supporting enhanced, context- and situation-aware decision-making in connected vehicle environments. Future research could consider real-time deployment, integration with edge-computing infrastructure, and incorporating other data sources, including temporal sequences and traffic flow information, to further improve system performance.

The proposed framework can be extended to real-time deployment by introducing edge and fog computing to support low-latency decision-making for IoV applications. Moreover, integrating temporal information such as video sequences and trajectory data can further improve prediction by capturing the dynamic aspects

of traffic. The next step is to combine other contextual data, such as data about weather conditions and traffic flow analysis, to get a better picture of the situation. The framework will be optimized for lightweight implementation and scalability to support its use in large-scale cyber-physical intelligent transportation systems.

DECLARATIONS

Authors' contributions

Contributed to conceptualization, software, validation, supervision, and writing the original draft: Zhukabayeva, T.; Ahmad, Z.

Performed formal analysis, visualization, project administration, and review and editing of the article: Karabayev, N.; Baumuratova, D.; Benkhelifa, E.

All authors reviewed the article.

Availability of data and materials

Two datasets are used for the development and testing of the proposed two-stage framework, both of which are publicly available in Kaggle platform with the titles “Vehicle Images Dataset”^[26] and “Internet of Vehicles Dataset”^[27].

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tool ChatGPT (GPT-5.5 Instant, released 2026-05-05) was used solely to assist with the generation and visualization of the photographs included in Figure 1. The tool did not influence the study design, data collection, analysis, interpretation, or scientific content. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23489127).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Biswas, A.; Wang, H. C. Autonomous vehicles enabled by the integration of IoT, edge intelligence, 5G, and blockchain. *Sensors* **2023**, *23*, 1963. DOI PubMed PMC
2. Kumar, N.; Sood, S. K.; Saini, M. IoV-fog-assisted framework for accident detection and classification. *ACM. Trans. Embed. Comput. Syst.* **2025**, *24*, 1-19. DOI
3. Mohialden, Y. M.; Hussien, N. M.; Hussain, S. A. K. Mitigating cybersecurity threats in autonomous vehicles: a real-world case study on protecting V2X communication against remote hacking attempts. *Cybersys. J.* **2024**, *1*, 41-50. DOI
4. Rishiwal, V.; Agarwal, U.; Alotaibi, A.; Tanwar, S.; Yadav, P.; Yadav, M. Exploring secure V2X communication networks for human-centric security and privacy in smart cities. *IEEE. Access.* **2024**, *12*, 138763-88. DOI
5. Ouaisa, M.; Ouaisa, M.; Houmer, M.; El Hamdani, S.; Boulouard, Z. A secure vehicle to everything (V2X) communication model for intelligent transportation system. In: Ouaisa, M.; Boulouard, Z.; Ouaisa, M.; Guermah, B., eds. *Computational intelligence in recent communication networks*. Cham: Springer International Publishing; 2022. pp. 83-102. DOI

6. Alharthi, M.; Medjek, F.; Djenouri, D. Ensemble learning approaches for multi-class intrusion detection systems for the Internet of Vehicles (IoV): a comprehensive survey. *Future. Internet.* **2025**, *17*, 317. DOI
7. Seo, J. K.; Lee, J.; Kim, B.; Shim, W.; Seo, J. T. AI-based anomaly detection in industrial control and cyber-physical systems: a data-type-oriented systematic review. *Electronics* **2026**, *15*, 20. DOI
8. Alsulami, A. A.; Al-Hajja, Q. A.; Alturki, B.; Alqahtani, A.; Alsini, R. Security strategy for autonomous vehicle cyber-physical systems using transfer learning. *J. Cloud. Comp.* **2023**, *12*, 564. DOI
9. Xing, W.; Shen, J. Security control of cyber-physical systems under cyber attacks: a survey. *Sensors* **2024**, *24*, 3815. DOI PubMed PMC
10. Mienye, I. D.; Swart, T. G. A comprehensive review of deep learning: architectures, recent advances, and applications. *Information* **2024**, *15*, 755. DOI
11. Salim, M. M.; Shanmuganathan, V.; Loia, V.; Park, J. H. Deep learning enabled secure IoT handover authentication for blockchain networks. *Hum. Cent. Comput. Inf. Sci.* **2021**, *11*, 21. DOI
12. Mozaffaree Pour, N.; Oja, T. Prediction power of logistic regression (LR) and multi-layer perceptron (MLP) models in exploring driving forces of urban expansion to be sustainable in Estonia. *Sustainability* **2022**, *14*, 160. DOI
13. Shah, K.; Patel, H.; Sanghvi, D.; Shah, M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment. Hum. Res.* **2020**, *5*, 32. DOI
14. Algarni, A.; Ahmad, Z.; Alaa Ala'anzy, M. An edge computing-based and threat behavior-aware smart prioritization framework for cybersecurity intrusion detection and prevention of IEDs in smart grids with integration of modified LGBM and one class-SVM models. *IEEE. Access.* **2024**, *12*, 104948–63. DOI
15. Safi, A.; Ahmad, Z.; Jehangiri, A. I.; et al. A fault tolerant surveillance system for fire detection and prevention using LoRaWAN in smart buildings. *Sensors* **2022**, *22*, 8411. DOI PubMed PMC
16. Talaat, F. M. Effective prediction and resource allocation method (EPRAM) in fog computing environment for smart healthcare system. *Multimed. Tools. Appl.* **2022**, *81*, 8235–58. DOI
17. Muthuramalingam, S.; Bharathi, A.; Rakesh Kumar, S.; Gayathri, N.; Sathiyaraj, R.; Balamurugan, B. IoT based intelligent transportation system (IoT-ITS) for global perspective: a case study. In: Balas, V. E.; Solanki, V. K.; Kumar, R.; Khari, M.; eds. *Internet of Things and big data analytics for smart generation*. Cham: Springer International Publishing; 2019. pp. 279–300. DOI
18. Balajee, A.; Vinoth, R.; Suresh, A.; Khan, M.; Mahesh, T.; Sayal, A. Federated hyper LSTM model for storage optimization and collision prediction in an intelligent IoVT. *Egypt. Inform. J.* **2026**, *33*, 100884. DOI
19. Seghir, Z.; Guezouli, L.; Barka, K.; Boubiche, D. E.; Toral-Cruz, H.; Martínez-Peláez, R. A real-time consensus-free accident detection framework for Internet of Vehicles using vision transformer and EfficientNet. *AI* **2026**, *7*, 4. DOI
20. Li, H.; Chen, L. Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories. *PLoS. One.* **2025**, *20*, e0320656. DOI PubMed PMC
21. Khan, F. M.; Zeb, A.; Rahman, T.; et al. Federated deep learning for collision avoidance in IoV with digital twin integration. *Expert. Syst.* **2026**, *43*, e70168. DOI
22. Selvaraj, D. C.; Vitale, C.; Panayiotou, T.; Kolios, P.; Chiasserini, C. F.; Ellinas, G. Edge-assisted ML-aided uncertainty-aware vehicle collision avoidance at urban intersections. *IEEE. Trans. Intell. Veh.* **2026**, *11*, 573–89. DOI
23. Liu, Y.; Gao, B.; Lin, P.; Tian, G.; Li, K. A framework for real-time vehicle tracking in large-scale roadside sensor networks. *Green. Energy. Intell. Transp.* **2025**, *4*, 100362. DOI
24. Zhang, J.; Zhan, R.; Wang, Y.; Qu, X. Optical communication based V2V for vehicle platooning. *Green. Energy. Intell. Transp.* **2025**, *4*, 100278. DOI
25. Khan, M. A. A.; Ma, H.; Farhad, A.; Mujeeb, A.; Mirani, I. K.; Hamza, M. When LoRa meets distributed machine learning to optimize the network connectivity for green and intelligent transportation system. *Green. Energy. Intell. Transp.* **2024**, *3*, 100204. DOI
26. Kaggle. Vehicle Images Dataset. <https://www.kaggle.com/datasets/lyensoetanto/vehicle-images-dataset>. (accessed 2026-09-15).
27. Kaggle. Internet of Vehicles Dataset. <https://www.kaggle.com/datasets/harunachiromagombe/internet-of-vehicles-dataset>. (accessed 2026-09-15).

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.