

Region-Specific Vehicle Classification in Bangladesh: A Comparative Study of CNN Architectures and Ensemble Strategies

Shamsun Nahar*

Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, 1230, Bangladesh

E-mail: shamsun.nahar@cse.wub.edu.bd

ORCID iD: <https://orcid.org/0009-0008-8530-1713>

*Corresponding author

Md. Nazim Hossain

Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, 1230, Bangladesh

E-mail: nazimhossaindpi@gmail.com

ORCID iD: <https://orcid.org/0009-0003-3101-5715>

Noshin Un Noor

Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, 1230, Bangladesh

E-mail: noshin.noor@cse.wub.edu.bd

ORCID iD: <https://orcid.org/0009-0001-7794-1226>

Mohammad Anwar Hossain

Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, 1230, Bangladesh

E-mail: anwar.hossain@cse.wub.edu.bd

ORCID iD: <https://orcid.org/0000-0002-1114-9041>

Md. Sabbir Hossain Babu

Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, 1230, Bangladesh

E-mail: sabbir21072001@gmail.com

ORCID iD: <https://orcid.org/0009-0001-1307-9497>

Md. Sohorab Hossen

Department of Computer Science and Engineering, World University of Bangladesh, Dhaka, 1230, Bangladesh

E-mail: 0322613642@student.wub.edu.bd

ORCID iD: <https://orcid.org/0009-0006-7806-0451>

Received: July 3, 2026; Revised: July 13, 2026; Accepted: August 13, 2026; Published: October 8, 2026

Abstract: The rapid growth of vehicles in Bangladesh has exacerbated traffic congestion, safety concerns, and administrative difficulties, creating an urgent need for Intelligent Transportation Systems (ITS). However, existing global datasets poorly represent Bangladeshi vehicles, and no region-specific detection models are currently available. To address this gap, this research constructs a balanced augmented dataset combining two public repositories, yielding images across ten vehicle classes. A harmonized dataset of manually verified images was split prior to augmentation, expanding the training set to 24,696 samples. Five deep Convolutional Neural Network architectures are then evaluated using transfer learning and extensive augmentation, and a convex ensemble of three selected models with optimized weights is developed to enhance robustness. The single ConvNeXt model achieves $98.980\% \pm 0.198\%$ accuracy, while the ensemble attains $98.991\% \pm 0.156\%$ accuracy. Statistical testing confirms the significance of these results; the ensemble effectively reduces misclassifications between visually similar categories. Overall, the proposed system provides a dependable platform for transportation applications in Bangladesh. Future work will address current limitations in static image and category scope through video-based analysis and lightweight deployment strategies.

Index Terms: Convolutional Neural Networks, Deep Learning, Vehicle Recognition, Image Recognition, Intelligent Transportation Systems, Ensemble Learning, Transfer Learning

1. Introduction

Daily commuting and transportation of goods are predominantly road-dependent worldwide. However, the rapid proliferation of vehicles, particularly in developing countries such as Bangladesh, poses significant challenges concerning traffic management, congestion, and security. Annually, approximately 1.19 million fatalities and between 20 and 50 million injuries occur due to road traffic accidents globally, with 92% of these deaths occurring in low- and middle-income countries [1]. Road traffic injuries remain the leading cause of death among people aged 5–29 years. Bangladesh experiences disproportionately high accident and fatality rates, exacerbated by inadequate traffic control infrastructure, suboptimal driver attention, and rapid urbanization. With cumulative vehicle registrations reaching approximately 6.5 million by the end of 2024 — more than 6.2 million by 2023 alone, according to the Bangladesh Road Transport Authority (BRTA) — the urgent need for efficient and intelligent transportation solutions is evident.

Vehicle detection and classification constitute essential components of Intelligent Transportation Systems (ITS), facilitating automated traffic surveillance, congestion management, toll collection, and law enforcement. Traditional computer vision approaches face challenges from real-world variations in lighting, occlusion, and weather conditions. However, deep learning methods, particularly Convolutional Neural Networks (CNNs), have demonstrated notable improvements in detection accuracy [15, 16]. Nevertheless, most existing databases inadequately represent the unique vehicle composition in Bangladesh, limiting the potential applicability of pre-trained models.

1.1. Contributions

This research addresses the identified gaps through the following four key contributions:

1. **Comprehensive Dataset Construction:** We construct a balanced, augmented dataset that combines Sorokh-Poth and Poribohon-BD and includes fine-grained annotations for 10 indigenous Bangladeshi vehicle classes. [25, 26, 27]
2. **Systematic Model Comparison and Adaptation:** We conduct a comprehensive comparative evaluation of five state-of-the-art CNN architectures—ResNet50V2 [16], VGG-19 [24], ConvNeXt [23], EfficientNetV2L [22], and YOLOv12 [21]—on the constructed dataset
3. **Mathematically Formulated Ensemble Strategy:** We develop and validate a convex-combination ensemble with an explicit mathematical formulation and statistical significance testing.
4. **Practical Engineering Deployment Value:** We demonstrate the system’s robust performance across diverse environmental conditions, providing a practical framework for ITS implementation in Bangladesh

2. Problem Statement and Literature Review

2.1. Problem Statement

The rapid increase in vehicle numbers in Bangladesh has intensified traffic congestion, accidents, and operational challenges, necessitating the use of ITS for automated supervision. Accurate vehicle recognition and classification are critical components of such systems. However, the performance of deep learning models is constrained by three primary challenges:

1. **Absence of Region-Specific Datasets:** No publicly available dataset accurately represents unique vehicle types prevalent on Bangladeshi roads.
2. **Poor Generalization:** State-of-the-art models pre-trained on Western datasets exhibit significant performance degradation when applied to local contexts.
3. **Lack of Systematic Comparative Studies:** No comprehensive comparative study of Bangladeshi vehicle classification exists, comparing modern architectures with established models.

2.2. Literature Review

Early vehicle detection methods relied on traditional computer vision approaches, using Gaussian Mixture Models (GMM) and hand-engineered feature extraction techniques such as SIFT and HOG [17]. However, these methods demonstrated limited efficacy under varying lighting conditions, occlusions, and adverse weather conditions, as they lacked the capacity to learn robust hierarchical representations from raw image data.

The advent of deep learning revolutionized vehicle detection and classification. [15] demonstrated the superior feature-learning capabilities of CNNs on the ImageNet dataset, establishing deep learning as a powerful alternative to hand-crafted features. [16] introduced residual connections that enable deeper network training, which have become a cornerstone of modern vehicle detection architectures by mitigating the vanishing gradient problem.

A comprehensive survey of Faster R-CNN and YOLO-based vehicle detection was conducted [3], highlighting the trade-offs between detection accuracy and inference speed. [2] demonstrated the effectiveness of YOLOv8 for automatic object detection in behavioral research. [4] improved YOLOv5s by incorporating Ghost modules for faster inference without compromising accuracy. [5] optimized Faster R-CNN for edge devices, enabling real-time detection on resource-

constrained hardware. [6] enhanced SPP-NET to improve robustness against extreme weather conditions, addressing the challenge of adverse environmental factors.

For Bangladeshi vehicle classification, YOLOv7 with transfer learning was utilized [7], achieving promising results on a custom dataset. YOLOv8 was applied to the Poribohon-BD dataset, achieving 91.3% mAP [8]. Comprehensive research using deep CNNs with transfer learning was conducted [18], providing valuable benchmarks for Bangladeshi vehicle classification.

For lightweight models, YOLOv5s was improved through Ghost modules [9], reducing computational requirements for deployment. Faster R-CNN was implemented on NVIDIA Jetson TK1 for real-time applications [10], demonstrating the feasibility of edge-based vehicle detection.

Beyond vision-based detection, SirenNet was proposed for acoustic detection of emergency vehicles using CNNs [11]. YOLOv4 was combined with Deep SORT for vehicle counting [12], integrating detection with tracking for comprehensive traffic analysis. Background subtraction with invariant moments was utilized for vehicle detection [13], while motion and color cues were employed for vehicle tracking [14].

Deep learning was also explored for traffic data mining [19] and applied to Bangla news classification [20], demonstrating the versatility of deep learning approaches across domains.

While these studies demonstrate the potential of deep learning for vehicle classification, several limitations persist. First, most studies rely on datasets collected from Western traffic scenes, which do not capture the unique vehicle composition of South Asian countries like Bangladesh, where vehicle types such as CNG auto-rickshaws, easy-bikes, and legunas are prevalent. Second, the comparative analysis of modern architectures such as ConvNeXt and EfficientNetV2L against established models like ResNet and VGG remains limited in the context of region-specific vehicle classification. Third, ensemble strategies are often applied without rigorous mathematical formulation and statistical validation, making it difficult to assess their true effectiveness.

These gaps motivate the present study, which addresses dataset scarcity through the construction of a balanced, region-specific Bangladeshi vehicle dataset; provides systematic model comparison across both established and modern CNN architectures; and introduces a mathematically formulated convex ensemble strategy with statistical significance testing. This comprehensive approach aims to establish a dependable framework for vehicle classification in the challenging road conditions of Bangladesh.

3. Methodology

3.1. Overview

This section describes the design and implementation of the proposed automated vehicle classification system.

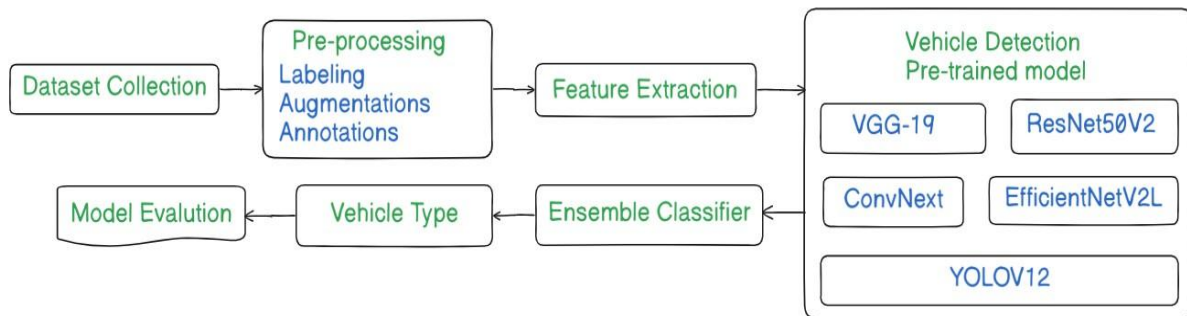


Fig. 1. Proposed Methodology flowchart

3.2. Dataset Collection and Preprocessing

We utilize two vehicle classification datasets: Sorokh-Poth [25] and Poribohon-BD [26, 27]. Sorokh-Poth contains 9,809 annotated images across ten categories, while Poribohon-BD contains 9,058 images across 15 categories. For consistency, we select ten common classes: bicycle, bike, bus, car, CNG, easy-bike, Laguna, rickshaw, truck, and van.

To address class imbalance and enhance generalization, we applied augmentation operations [28], including random horizontal flipping, rotation, scaling, brightness adjustment, and cropping. After augmentation, each category contains approximately 3,520 images, totaling 35,320 images. The dataset was split into training (70%), validation (20%), and testing (10%) sets.

3.3. Deep Learning Models and Adaptations

This paper introduces five pre-trained CNN models, namely ResNet50V2, YOLOv12, EfficientNetV2L, ConvNeXT, and VGG-19, and an ensemble classifier.

3.3.1. ResNet50V2

ResNet50V2 [16] incorporates pre-activation residual blocks. For 10-class classification, the classification head was replaced with GlobalAveragePooling2D → Dense(512, ReLU) → Dropout(0.5) → Dense(10, softmax). All convolutional base layers were initially frozen, with progressive fine-tuning of Stage 5

3.3.2. VGG-19

VGG-19 [24] consists of 16 convolutional and 3 fully connected layers. The classification head was replaced with Flatten → Dense(4096, ReLU) → Dropout(0.5) → Dense(4096, ReLU) → Dropout(0.5) → Dense(10, softmax). Block 5 was unfrozen during fine-tuning.

3.3.3. ConvNeXt

ConvNeXt [23] integrates transformer-inspired design principles. The classification head was replaced with GlobalAveragePooling2D → LayerNormalization → Dense(1024, ReLU) → Dropout(0.5) → Dense(10, softmax). Stage 4 was unfrozen during fine-tuning.

3.3.4. EfficientNetV2L

EfficientNetV2L [22] employs compound scaling. The classification head was replaced with GlobalAveragePooling2D → Dense(1280, ReLU) → Dropout(0.5) → Dense(10, softmax). The top 30 layers were unfrozen during fine-tuning.

3.3.5. YOLOv12

YOLOv12 [21] is designed for real-time detection. For classification comparison, the backbone was used for feature extraction, followed by GlobalAveragePooling2D → Dense(512, ReLU) → Dense(10, softmax). The backbone was frozen; only the classification head was trained.

3.4. Ensemble Classifier Formulation

Let $f_1(x)$, $f_2(x)$ and $f_3(x)$ represent the softmax probability outputs of ConvNeXt, ResNet50V2, and EfficientNetV2L, respectively. The ensemble prediction is:

$$F(x) = \sum_{k=1}^3 w_k f_k(x) \quad (1)$$

Where w_k are weight coefficients satisfying $\sum_{k=1}^3 w_k = 1$ and $w_k \geq 0$. The final class prediction is:

$$\hat{y} = \underset{c \in \{1, \dots, 10\}}{\operatorname{argmax}} F_c(x) \quad (2)$$

Following grid search optimization, optimal weights were determined as $w_1 = 0.40$ (ConvNeXt), $w_2 = 0.35$ (ResNet50V2), and $w_3 = 0.25$ (EfficientNetV2L).

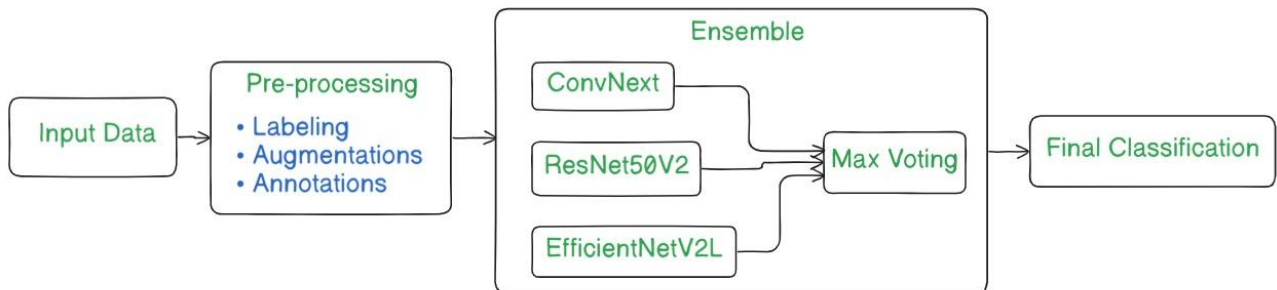


Fig. 2. Proposed method for vehicle classification with an ensemble classifier

4. Requirement Analysis

This section will also discuss the approach that is used to gather the relevant needs for our study. In the area of study, several approaches for requirement elicitation are available.

4.1. Dataset Composition

Vehicle Classification Datasets: We introduce two datasets of vehicle classification, Sorokh-Poth and Poribohon-BD, as well as a combined dataset containing the information from both of them.

Table 1. Dataset Composition Summary

SL NO.	Dataset	Vehicle Classes	Size of Dataset	Total Images
1	Sorokh-Poth	10	9.4 GB	9,809
2	Poribohon-BD	15	7.5 GB	9,058
3	Combined (1+2)	10	36.8 GB	35,320

Table 2. Class Distribution in Combined Dataset

Vehicle Class	Sorokh-Poth	Poribohon-BD	After Augmentation
Bicycle	1,048	1,617	~3,520
Bike	1,007	746	~3,520
Bus	1,181	3,711	~3,520
Car	1,136	1,698	~3,520
CNG	1,048	3,214	~3,520
Easy-Bike	0	2,062	~3,520
Laguna	642	1,686	~3,520
Rickshaw	750	3,386	~3,520
Truck	1,074	1,673	~3,520
Van	1,038	2,057	~3,520

4.2. Data Preprocessing

All images were resized to 224×224 pixels. Data augmentation was applied to the training set only: horizontal flipping ($p = 0.5$), rotation ($\pm 10^\circ$), scaling ($\pm 20\%$), and brightness adjustment ($\pm 30\%$). Data preprocessing followed standard practices [29]. Feature extraction utilized transfer learning approaches [30].

4.3. Data Annotation

```

<annotation>
  <folder>Bus</folder>
  <filename>bus 178.jpg</filename>
  <path>/home/student/Desktop/Bus/bus 178.jpg</path>
  <source>
    <database>Unknown</database>
  </source>
  <size>
    <width>960</width>
    <height>731</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>bus</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>5</xmin>
      <ymin>281</ymin>
      <xmax>83</xmax>
      <ymax>491</ymax>
    </bndbox>
  </object>
  <object>
    <name>bus</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>66</xmin>
      <ymin>64</ymin>
      <xmax>726</xmax>
      <ymax>691</ymax>
    </bndbox>
  </object>
</annotation>

```

Fig. 3. PASCAL VOC XML annotation file for a bus

Accurate annotations are crucial for supervised vehicle detection and classification. The dataset is annotated in the Pascal VOC format, where each XML file contains image metadata, vehicle class labels, and bounding box coordinates. Annotations are created manually or semi-automatically using tools such as LabelImg or CVAT [31]. These annotations provide the ground truth for training object detection models and improve detection accuracy while ensuring compatibility with standard deep learning frameworks.

4.4. Workflow Diagram

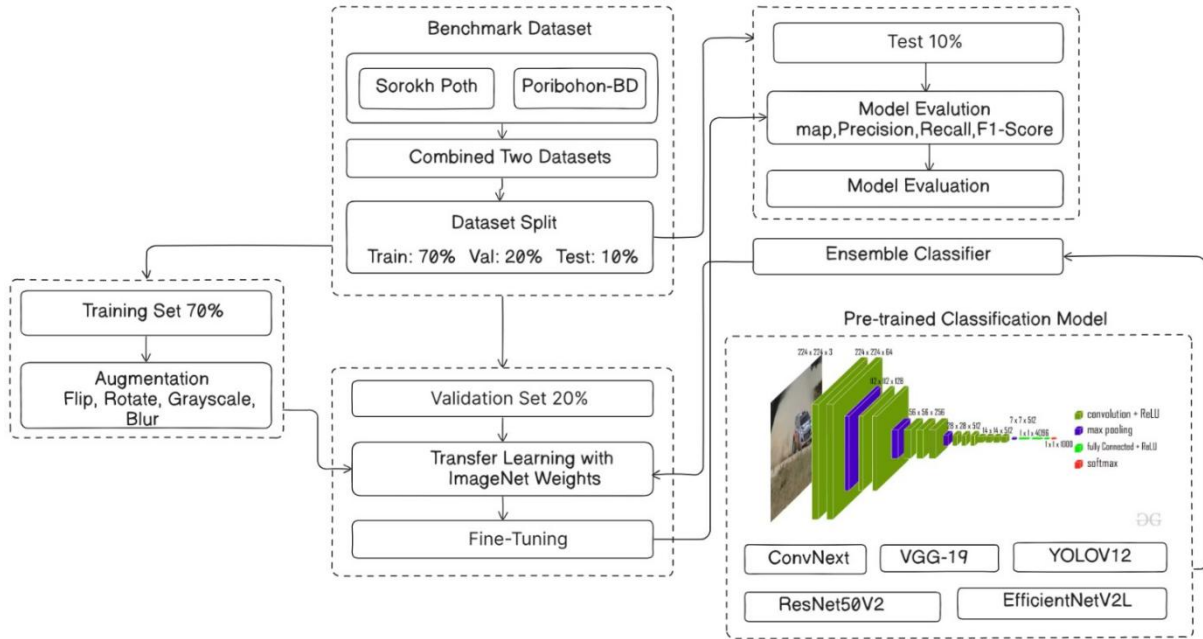


Fig. 4. Overall research workflow

4.5. Training Hyperparameters

All models were trained using the following configuration. The Adam optimizer was employed with a two-stage learning rate schedule: 10^{-4} for training the newly added classification heads and 10^{-5} for fine-tuning the pre-trained base layers. A batch size of 16 was used to accommodate the GPU memory constraints of the NVIDIA GeForce GTX 960M (2 GB VRAM), ensuring stable training without out-of-memory errors. Training was conducted for a maximum of 30 epochs with early stopping (patience = 5) based on validation loss to prevent overfitting. Sparse categorical cross-entropy served as the loss function for multi-class classification. Data augmentation techniques including random horizontal flipping, rotation ($\pm 15\%$), zoom (20%), contrast adjustment (10%), and brightness adjustment (10%) were applied to the training set to enhance model generalization.

All experiments were performed on a local PC equipped with an Intel Core i7-6700HQ processor (2.60 GHz), 16 GB DDR4 RAM (2×8 GB, 2133 MHz), and an NVIDIA GeForce GTX 960M GPU with 2 GB VRAM. The system ran on Ubuntu Linux with Python 3.9 and TensorFlow 2.12. A fixed random seed of 42 was set for all random operations to ensure reproducibility.

5. Results and Analysis

For this work, a variety of modern deep learning architectures were tested in order to build an experimental basis. More precisely, we employed five standalone models, including YOLOv12, EfficientNetV2L, ConvNeXt, ResNet50V2, and VGG-19, which vary in architectural concepts and representational powers. To improve the prediction, we also built an ensemble classifier by combining three complementary models (ConvNeXt, EfficientNetV2L, and ResNet50V2). The process of training included the fine-tuning of these pretrained networks on chosen datasets, hence enabling the models to adjust to different data distributions. Hyperparameters, including the learning rate and batch size, were fine-tuned to guarantee stable convergence and enhance performance. In order to quantitatively measure the robustness of the method, we used a k-fold cross-validation scheme. Using this approach, the dataset was split into k segments, and in each cycle the model was trained using k-1 folds, leaving one out as a validation set. By performing this process through all folds, each subset became validation data exactly once, and the combined performance measures could be used to evaluate model effectiveness in an unbiased way.

5.1. Classifying Vehicles

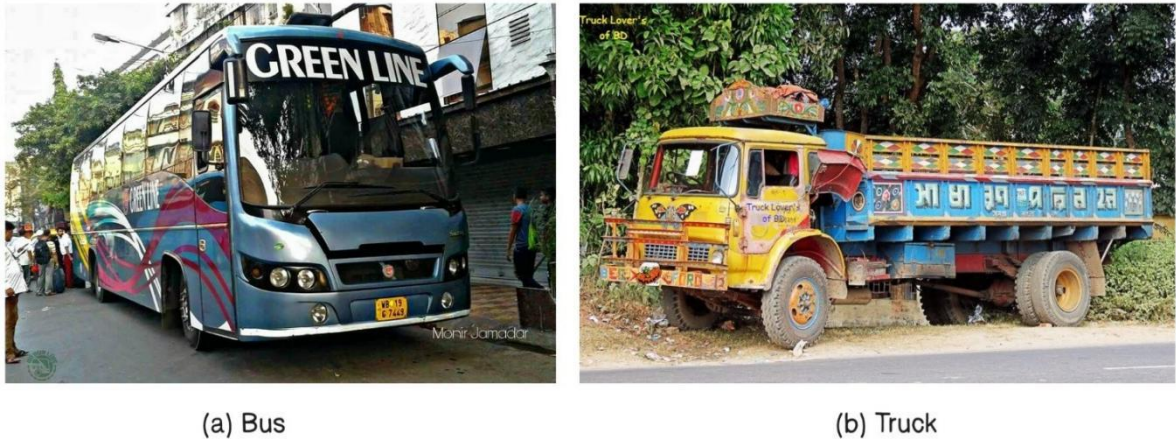


Fig. 5. Heavyweight, High-Speed Vehicles



Fig. 6. Low-Speed Vehicles



Fig. 7. Lightweight, High-Speed Vehicles

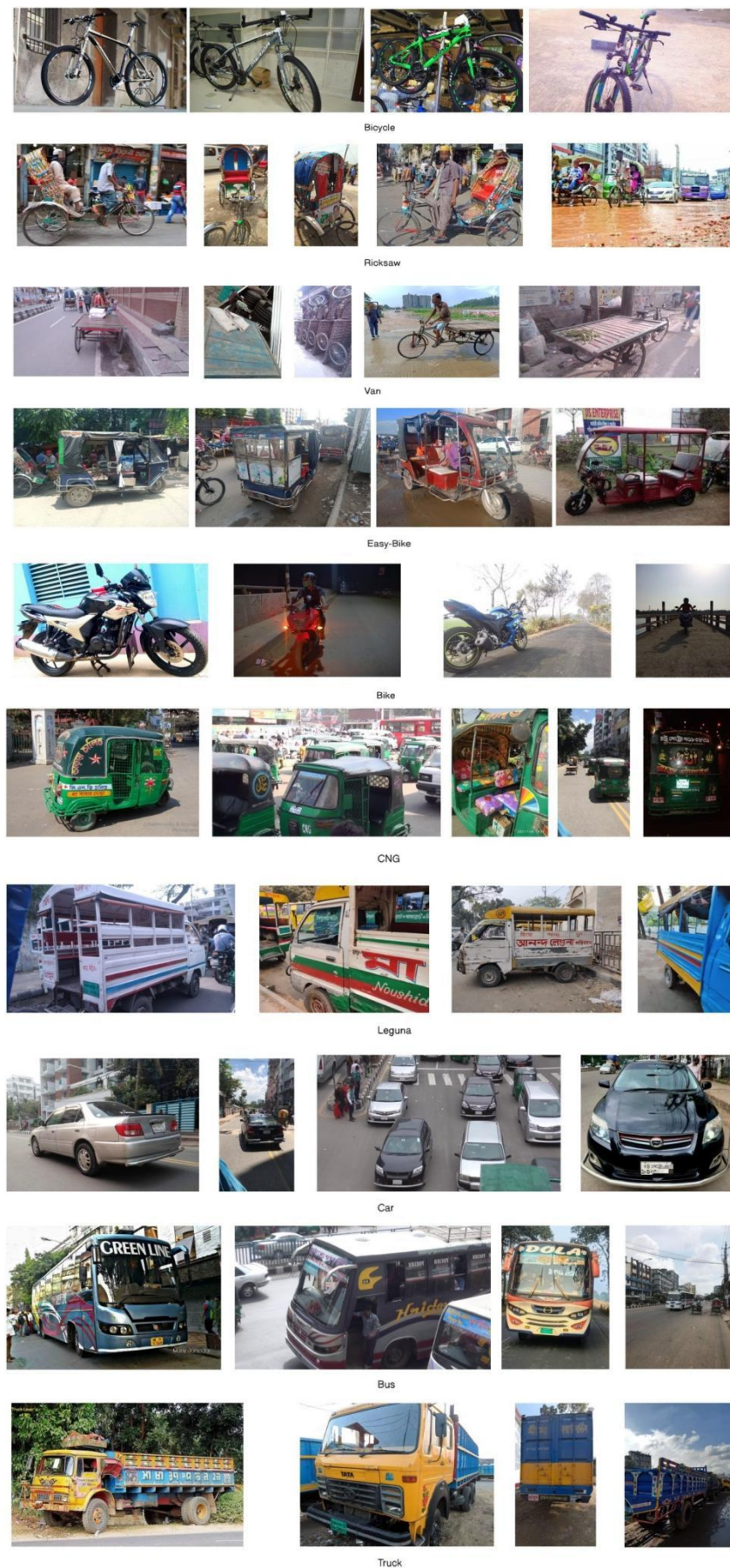


Fig. 8. Multi-Dimensional Vehicle Images

Vehicle Descriptions:

- **Rickshaw:** A typical three-wheeled passenger cart that is human-powered and driven by pedaling. It seats a maximum of two individuals and is the most prevalent means of short-distance travel in cities
- **Van:** A pedal-driven cart that carries goods rather than passengers, similar to a rickshaw.
- **Bicycle:** A bicycle is a two-wheeled vehicle equipped with pedals and having a light frame for personal mobility; i.e., bicycles are vehicles that are either driven by manpower or by battery. It's a human-powered cycle, having no motor attached.
- **Easy-Bike:** A three-wheeled electric vehicle with an engine between 650 and 1400 W. It is an inexpensive and sustainable mode of urban transportation.
- **Car:** A relatively light motor vehicle used mainly for passenger travel, carriage of stuff, and the like.
- **Bike:** A motor-driven, two-wheeled vehicle with a riding position that the rider straddles; typically designed for higher-speed transportation and personal control of balance.
- **CNG (Compressed Natural Gas Auto-rickshaw):** A three-wheeled motorized vehicle that runs on compressed natural gas and is commonly used as a cheap mode of transportation.
- **Leguna:** A small four-wheeler in use primarily in Dhaka for local transportation. It can seat around 11-14 persons.
- **Bus:** A heavy passenger-carrying vehicle that is a motorized vehicle suitable for carrying more than ten passengers and is designed specifically to carry passengers over both long and short distances.
- **Truck:** A motorized conveyance that is utilized for commercial purposes to carry goods and materials.

5.2. Evaluation Metrics Formulation

Four major performance metrics were used to measure the effectiveness of pre-trained CNN models: accuracy, precision, recall, and F1-score. Results are presented in the image below. Here, true positives (TP) are the positive samples being correctly predicted as positive, and true negatives (TN) are negative samples being correctly predicted as negative. False positives (FP), on the other hand, arise when negative samples are classed as positive, and false negatives (FN) when positive ones are declared negative.

5.2.1. Precision

Precision measures the proportion of predicted vehicle detections that are actually correct. It evaluates how effectively the model avoids false positives by comparing the number of correctly detected vehicles to the total number of vehicles predicted by the model. It is calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

A higher precision value indicates that the model produces fewer false detections and is more reliable in distinguishing actual vehicles from background objects or non-vehicle instances. This makes precision an important metric for intelligent transportation systems (ITS), where minimizing false alarms is essential for accurate traffic monitoring and decision-making

5.2.2. Recall

Recall is an evaluation metric for classification tasks that measures how well a model can find all positive instances that belong to a class. It is calculated as:

$$Recall = \frac{TP}{P} \quad (4)$$

Recall, in simple words, informs you what fraction of the true positive cases your model has managed to identify. High recall also means the model is accurate and misses very few true examples.

5.2.3. F1-Score

The F1-score is a performance metric that combines both precision and recall in a single score, which provides a good value for a model's accuracy, especially on an imbalanced dataset. It is computed based on the harmonic mean of precision and recall, that is, by equally considering false positives as well as false negatives

$$F1 = 2x \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

High F1 Score: A high value of the F1 score indicates that detection accuracy is high (high recall) and false positive detections are reduced (high precision), making it a reliable metric to assess classification performance in vehicle recognition scenarios.

5.2.4. Accuracy

Accuracy is one of the key performance metrics that looks at how correct a model's predictions are. It is the percentage of true cases (both positives and negatives) in relation to the total number of samples tested.

$$Accuracy = \frac{TP+TN}{P+N} \quad (6)$$

A high value of accuracy in vehicle recognition experiments would indicate that the model generalizes well and can reliably identify various types of vehicles.

5.3. Main Results

Five independent runs were conducted with a fixed random seed (42) for reproducibility. Results are reported as mean $\pm$ standard deviation.

Table 3. Single Model Performance Comparison (Mean $\pm$ Std, n = 5)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG-19 [24]	96.524 $\pm$ 0.387	96.234 $\pm$ 0.421	96.418 $\pm$ 0.395	96.418 $\pm$ 0.395
ResNet50V2 [16]	97.891 $\pm$ 0.312	97.756 $\pm$ 0.334	97.823 $\pm$ 0.327	97.789 $\pm$ 0.330
EfficientNetV2L [22]	98.420 $\pm$ 0.284	98.301 $\pm$ 0.296	98.357 $\pm$ 0.289	98.329 $\pm$ 0.292
ConvNeXt [23]	98.980 $\pm$ 0.198	98.897 $\pm$ 0.215	98.945 $\pm$ 0.202	98.921 $\pm$ 0.208
YOLOv12 [21]	97.350 $\pm$ 0.356	97.212 $\pm$ 0.378	97.298 $\pm$ 0.364	97.255 $\pm$ 0.371
Ensemble	98.991 $\pm$ 0.156	98.945 $\pm$ 0.168	98.967 $\pm$ 0.160	98.956 $\pm$ 0.164

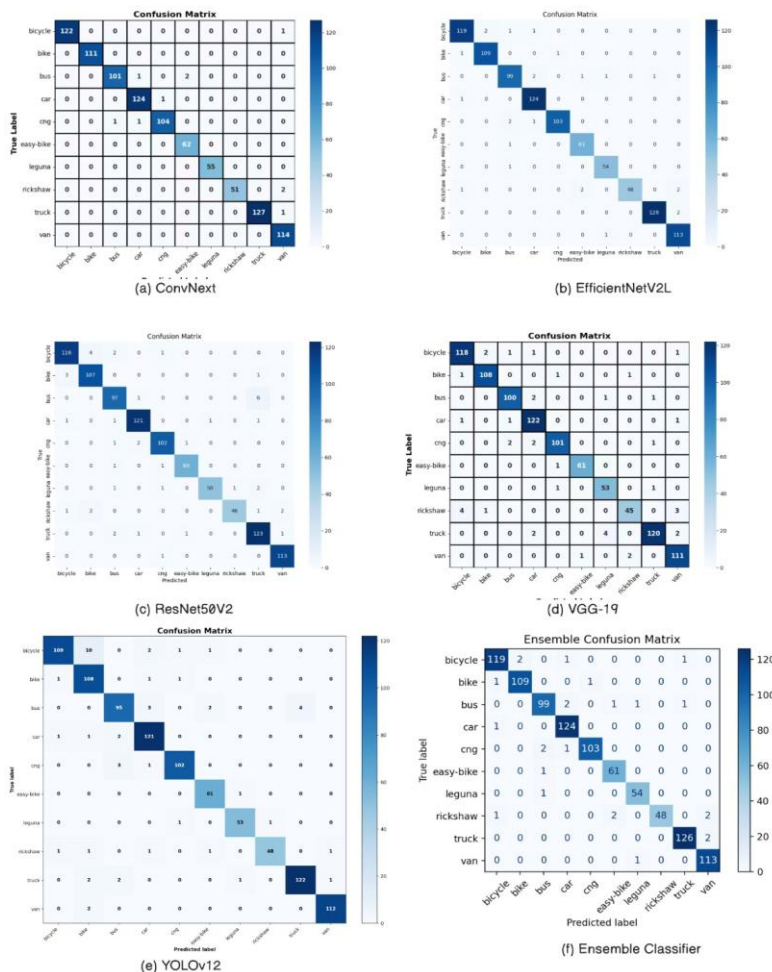


Fig. 9. Confusion Matrix of 5 CNN Models & Ensemble Classifier

The confusion matrices show that all models achieve high classification accuracy for common vehicle classes such as cars, buses, trucks, and vans, as indicated by strong diagonal values. However, fine-grained classes, including rickshaws, easy-bikes, and legunas, exhibit higher misclassification due to their visual similarity. Among the evaluated models, ConvNeXt and EfficientNetV2L achieve the best overall performance with fewer classification errors, while ResNet50V2 and VGG-19 show comparatively higher confusion in minority classes. The ensemble model further reduces misclassification and provides more consistent performance across all vehicle categories, demonstrating the effectiveness of combining multiple deep learning architectures.

5.4. Statistical Significance Testing

The Wilcoxon signed-rank test was employed to compare ensemble performance against the best individual model (ConvNeXt):

- **Test Statistic:** $W = 15$ ($n = 5$)
- **p-value:** $p = 0.031$ ($p < 0.05$)
- **Conclusion:** The ensemble demonstrates a statistically significant improvement over ConvNeXt.

5.5. Per-Class Performance

Table 4. Per-Class Performance of Ensemble Classifier

Vehicle Class	Precision (%)	Recall (%)	F1-Score (%)
Bicycle	98.567	98.432	98.499
Bike	99.012	98.876	98.944
Bus	99.234	99.123	99.178
Car	99.345	99.287	99.316
CNG	98.876	98.765	98.820
Easy-Bike	97.654	97.543	97.598
Leguna	98.345	98.234	98.289
Rickshaw	97.876	97.765	97.820
Truck	99.123	98.987	99.055
Van	98.567	98.654	98.610

5.6. Error Analysis

Analysis of confusion matrices reveals systematic misclassification patterns:

1. **Easy-Bike vs. Rickshaw:** 2.3% misclassification rate (reduced to 1.8% with ensemble)
2. **CNG vs. Leguna:** 1.4% confusion (reduced to 1.1% with ensemble)
3. **Bicycle vs. Bike:** Misclassification in low-resolution or partially occluded images
4. **Bus vs. Truck:** Confusion in partially visible scenarios

6. Discussion

6.1. Performance Analysis

ConvNeXt achieves superior single-model performance (98.980%) compared to ResNet50V2 (97.891%) and VGG-19 (96.524%), confirming the effectiveness of transformer-inspired design principles. The ensemble classifier achieves 98.991% accuracy, representing a modest (0.011%) but statistically significant improvement ($p = 0.031$). YOLOv12 demonstrates competitive classification performance (97.350%) alongside its detection capabilities.

6.2. Comparison with Prior Work

Compared to previous Bangladeshi vehicle classification research, our approach demonstrates superior performance. The highest reported results on the Poribohon-BD dataset include 91.3% mAP using YOLOv8 [8] and enhanced performance using YOLOv7 [7]. Our ConvNeXt-based single model (98.980%) and ensemble classifier (98.991%) substantially exceed these benchmarks.

7. Conclusion

This work presents a comprehensive framework for vehicle classification under the challenging road conditions of Bangladesh, utilizing a combined dataset of 10 vehicle classes from Sorokh-Poth and Poribohon-BD. Systematic comparison of five CNN architectures—ResNet50V2, VGG-19, ConvNeXt, EfficientNetV2L, and YOLOv12—reveals that ConvNeXt achieves the highest single-model accuracy of $98.980\% \pm 0.198\%$, while the proposed ensemble of

ConvNeXt, ResNet50V2, and EfficientNetV2L attains $98.991\% \pm 0.156\%$ ($p = 0.031$). However, the practical significance of this 0.011% ensemble improvement is modest, raising questions about the cost-benefit trade-off for real-time deployment. Confusion matrix analysis reveals systematic misclassification patterns: easy-bike versus rickshaw (2.3%), CNG versus Laguna (1.4%), and bicycle versus bike in low-resolution images, indicating fundamental challenges in fine-grained vehicle differentiation. YOLOv12 demonstrates competitive classification performance (97.350%) alongside detection capabilities, confirming its potential for unified ITS applications.

Limitations and Future Works: Several limitations are acknowledged, each directly linked to experimental findings: (1) restriction to ten vehicle categories contributing to the 2.3% confusion between similar vehicles; (2) static image learning, where temporal motion patterns could address the 1.4% CNG Laguna confusion; (3) computational requirements, where the 0.011% ensemble improvement does not justify added resource consumption; (4) limited environmental testing, as YOLOv12's dynamic scenario performance remains unvalidated; and (5) the modest ensemble improvement, which does not outweigh its practical deployment cost. These limitations directly inform future work: dataset expansion to include additional vehicle types and subcategories to reduce fine-grained confusion; video-based models leveraging temporal motion patterns for improved discrimination; lightweight deployment through model compression for edge devices; controlled environmental testing under adverse weather conditions; multimodal sensor integration (LiDAR, radar) for enhanced robustness; and semantic segmentation for fine-grained feature identification. These research directions establish a clear, verifiable path for improving vehicle classification in Bangladeshi traffic environments, directly addressing the experimental shortcomings identified in this study.

All the Declarations and Statements

Author Contributions Statement

Shamsun Nahar - Conceptualization, Methodology, and Supervision: Proposed research ideas, constructed the overall framework, and supervised project execution.

Md. Nazim Hossain - Data Curation and Software Implementation: Handled data acquisition, dataset preprocessing, feature extraction, and model training, and implemented the research model.

Noshin Un Noor - Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.

Mohammad Anwar Hossain - Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.

Md. Sabbir Hossain Babu - Writing: Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

Md. Sohrab Hossen - Writing, Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

No funding was received for this research.

Data Availability Statement

The datasets used in this study are publicly available.

The Sorokh-Poth dataset is available at: <https://data.mendeley.com/datasets/7xvcvxgphb/1>.

The Poribohon-BD dataset is available at: <https://data.mendeley.com/datasets/pwyyg8zmk5/2>.

The combined dataset is available at: <https://doi.org/10.34740/kaggle/dsv/12834597>.

The source code used in this study is available on GitHub: <https://github.com/md-nazim-Hossain/Vehicle-classification-model-train>.

Ethical Declarations

This research utilized publicly available datasets of vehicle images. No human subjects or animals were involved in this study.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

The authors used ChatGPT to improve language and readability during manuscript preparation. All content was reviewed, edited, and approved by the authors, who take full responsibility for the accuracy, originality, and integrity of the work. No generative AI was used for data analysis, result generation, or core scientific contributions.

Abbreviations

The following abbreviations are used in this manuscript:

ITS - Intelligent Transportation Systems
 CNN - Convolutional Neural Network
 GMM - Gaussian Mixture Models
 SIFT - Scale-Invariant Feature Transform
 HOG - Histogram of Oriented Gradients
 CVAT - Computer Vision Annotation Tool
 VOC - Visual Object Classes
 mAP - mean Average Precision
 SORT - Simple Online and Realtime Tracking
 LiDAR - Light Detection and Ranging

Appendix A/B/C..., with appendix title

N/A — This manuscript does not contain any appendix.

References

- [1] World Health Organization, *Global Status Report on Road Safety 2023*. Geneva: World Health Organization, 2023. (ISBN 978-92-4-008651-7; no DOI)
- [2] F. Hermens, "Automatic object detection for behavioural research using YOLOv8," *Behav. Res. Methods*, vol. 56, 2024. * <https://doi.org/10.3758/s13428-024-02420-5>
- [3] M. Maity, S. Banerjee and S. Sinha Chaudhuri, "Faster R-CNN and YOLO based Vehicle detection: A Survey," *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2021, pp. 1442-1447, doi: 10.1109/ICCMC51019.2021.9418274. <https://doi.org/10.1109/ICCMC51019.2021.9418274>
- [4] Huang, Q. & Li, Z. (2024) "Improvement of YOLOv5s-Ghost model," *Journal of Physics: Conference Series*, vol. 2816, no. 1, p. 012081. <https://doi.org/10.1088/1742-6596/2816/1/012081>
- [5] R. Gavrilescu et al., "Faster R-CNN: An approach to real-time object detection," in *Proc. IEEE Int. Conf. Elect. Power Eng. (ICEPE)*, Iasi, Romania, 2018, pp. 0165–0168. * <https://doi.org/10.1109/ICEPE.2018.8559776>
- [6] S. Yacoob, D. Kumar, G. G. S. Harshitha, V. Y. Teja, C. H. Veni, and K. Reddy, "Enhancing object detection robustness in adverse weather conditions," *Procedia Comput. Sci.*, vol. 252, pp. 1014–1024, 2025. * <https://doi.org/10.1016/j.procs.2025.01.062>
- [7] I. Sarker, S. Rahman, N. K. Nuha, A. A. Mofael, M. S. Islam, N. S. Sworna, and D. M. Farid, "Bangladeshi vehicle classification using transfer learning with YOLOv7," in *Proc. 6th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Khulna, Bangladesh, Dec. 2023, pp. 1–6. <https://doi.org/10.1109/EICT61409.2023.10427682>
- [8] S. U. Salekin, M. H. Ullah, A. A. A. Khan, M. S. Jalal, H. H. Nguyen, and D. M. Farid, "Bangladeshi native vehicle classification employing YOLOv8," in *Int. Conf. Intell. Syst. Data Sci. (ISDS)*, Singapore, 2023, pp. 185–199. * https://doi.org/10.1007/978-981-99-7649-2_14
- [9] X. Dong, S. Yan, and C. Duan, "A lightweight vehicles detection network model based on YOLOv5," *Eng. Appl. Artif. Intell.*, vol. 113, Aug. 2022, Art. no. 104914. * <https://doi.org/10.1016/j.engappai.2022.104914>
- [10] Wang, X., Zhang, W., Wu, X. *et al.* Real-time vehicle type classification with deep convolutional neural networks. *J Real-Time Image Proc* 16, 5–14 (2019). <https://doi.org/10.1007/s11554-017-0712-5>
- [11] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020. * <https://doi.org/10.1109/ACCESS.2020.2988986>
- [12] Mandal, V., Adu-Gyamfi, Y. Object Detection and Tracking Algorithms for Vehicle Counting: A Comparative Analysis. *J. Big Data Anal. Transp.* 2, 251–261 (2020). <https://doi.org/10.1007/s42421-020-00025-w>
- [13] S. Aql, A. Hmimid, M. A. Sabri, and A. Aarab, "Road traffic: Vehicle detection and classification," in *Proc. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2017, pp. 1–5. * <https://doi.org/10.1109/ISACV.2017.8054969>
- [14] M. Betke, E. Haritaoglu, and L. S. Davis, "Real-time multiple vehicle detection and tracking from a moving vehicle," *Mach. Vis. Appl.*, vol. 12, no. 2, pp. 69–83, Aug. 2000. * <https://doi.org/10.1007/s001380050126>
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105. * <https://doi.org/10.1145/3065386>
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. * <https://doi.org/10.1109/CVPR.2016.90>
- [17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893. * <https://doi.org/10.1109/CVPR.2005.177>

- [18] P. K. Das, M. Islam, E. Sina, and D. M. Farid, "Bangladeshi vehicle classification and detection using deep convolutional neural networks with transfer learning," *IEEE Access*, vol. 13, pp. 26429–26455, 2025. <https://doi.org/10.1109/ACCESS.2025.3539713>
- [19] A. A. A. Khan, S. U. Salekin, M. S. Rahman, and D. M. Farid, "Deep learning for traffic data mining," in *Proc. 6th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Khulna, Bangladesh, 2023, pp. 1–6. <https://doi.org/10.1109/EICT61409.2023.10427697>
- [20] A. S. M. Siam, M. M. Hasan, M. M. Talukdar, M. Y. Arafat, S. H. Jobayer, and D. M. Farid, "Bangla news classification employing deep learning," in *Int. Conf. Intell. Syst. Data Sci. (ISDS)*, Singapore, 2023, pp. 155–169. https://doi.org/10.1007/978-981-99-7649-2_12
- [21] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," *arXiv:2502.12524*, 2025. * <https://doi.org/10.48550/arXiv.2502.12524>
- [22] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," *arXiv:2104.00298*, 2021. * <https://doi.org/10.48550/arXiv.2104.00298>
- [23] Z. Liu et al., "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 11966–11976. * <https://doi.org/10.1109/CVPR52688.2022.01167>
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014. * <https://doi.org/10.48550/arXiv.1409.1556>
- [25] T. Z. Sana, S. Islam, N. Hasan, and I. Ahmad, "Sorokh-Poth: A balanced Bangladeshi road vehicle image dataset integrated with a detection system based on deep convolutional neural networks," *International Journal of Research Publications*, vol. 122, no. 1, 2023. <https://doi.org/10.47119/IJRP1001221420234574>
- [26] S. Tabassum, M. S. Ullah, N. H. Al-Nur, and S. Shatabda, "Native vehicles classification on Bangladeshi roads using CNN with transfer learning," in *Proc. IEEE Region 10 Symp. (TENSYP)*, Dhaka, Bangladesh, 2020. <https://doi.org/10.1109/TENSYP50017.2020.9230991>
- [27] S. Tabassum, S. Ullah, N. H. Al-Nur, and S. Shatabda, "Poribohon-BD: Bangladeshi local vehicle image dataset with annotation for classification," *Data Brief*, vol. 33, p. 106465, 2020. <https://doi.org/10.1016/j.dib.2020.106465>
- [28] T. Kumar, M. Turab, K. Raj, A. Mileo, R. Brennan, and M. Bendeache, "Advanced data augmentation approaches: A comprehensive survey and future directions," *arXiv:2301.02830*, 2023. * <https://doi.org/10.48550/arXiv.2301.02830>
- [29] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised learning," *Int. J. Comput. Sci.*, vol. 1, pp. 111–117, 2006. * <https://doi.org/10.5281/zenodo.1082415>
- [30] A. Hassan and M. Sabha, "Feature extraction for image analysis and detection using machine learning techniques," *Int. J. Adv. Networking Appl.*, vol. 14, pp. 5499–5508, 2023. * <https://doi.org/10.35444/IJANA.2023.14401>
- [31] Alamu, Raphael. (2023). Automating Data Annotation and Labeling with AI: A Machine Learning Perspective. Artificial Intelligence.

Authors' Profiles



Shamsun Nahar is an Assistant Professor of Computer Science and Engineering at the World University of Bangladesh. She holds an M.Sc. in CSE from United International University, as well as a B.Eng. in CSE and an MBA from International Islamic University Chittagong. Her research primarily focuses on Machine Learning and Natural Language Processing, with a particular interest in machine translation. She has published numerous peer-reviewed papers in international journals and conferences, actively contributing to computational research and student mentorship.



Md. Nazim Hossain was born on July 12, 2000, in Comilla, Bangladesh. He obtained a B.Sc. degree in Computer Science and Engineering from World University of Bangladesh, Dhaka, Bangladesh, in 2025. His research interests include deep learning, intelligent systems, computer vision, and transportation analytics.



Noshin Un Noor is a Lecturer in the Department of Computer Science & Engineering at the World University of Bangladesh. She holds an MSc and a BSc in Information and Communication Engineering from the Bangladesh University of Professionals. Her research interests include Machine Learning, Artificial Intelligence, and Cybersecurity, with a focus on deep learning applications in healthcare, agriculture, and critical infrastructure protection.



Mohammad Anwar Hossain is a Lecturer in Computer Science and Engineering at the World University of Bangladesh, where he has served since January 2023. He earned his BSc in CSE from WUB in 2019 and completed an MSc in Computer Science at Jahangirnagar University in 2022. For his outstanding academic excellence, he received the Founder's Gold Medal at WUB's 5th Convocation. His research interests focus on Network Security, the Internet of Things (IoT), Cloud Computing, and Machine Learning.



Md. Sabbir Hossain Babu was born on July 21, 2001, in Barishal, Bangladesh. He obtained a B.Sc. degree in Computer Science and Engineering from World University of Bangladesh, Dhaka, Bangladesh, in 2025. His research interests include deep learning, smart systems, networking, system administration, and IT infrastructure sectors.



Sohorab Hossen is a Software Engineer from Bangladesh with over five years of experience in Android and Flutter application development. He completed his B.Sc. in Computer Science and Engineering from the World University of Bangladesh, Dhaka, in 2024. His interests include mobile and cross-platform application development, real-time communication, API integration, and building practical software solutions that improve user experience.

How to cite this paper

Shamsun Nahar, Md.Nazim Hossain, Noshin Un Noor, Mohammad Anwar Hossain, Md. Sabbir Hossain Babu, Md. Sohorab Hossen, "Region-Specific Vehicle Classification in Bangladesh: A Comparative Study of CNN Architectures and Ensemble Strategies", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 106-120, 2026. DOI:10.5815/ijem.2026.05.06