Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Cross-Architecture Performance Evaluation of Transfer Learning Models for Multi-Class Vehicle Damage Severity Classification

Mochammad Fatih Ulumuddin[1)*], Anggay Luri Pramana[2)]
[1,2)]Universitas Nahdlatul Ulama Sidoarjo, Indonesia
[1)]ulumuddinfatih421@gmail.com, [2)]luri409.tif@unusida.ac.id

Submitted :Feb 24, 2026 | **Accepted** : March 4, 2026 | **Published** : April 3, 2026

Abstract: Automated vehicle damage classification supports objectivity and scalability in insurance claim processing and digital inspection systems; however, prior studies often report performance improvements without controlled experimental settings or statistical validation, limiting methodological reliability. This study establishes a statistically controlled cross-architecture evaluation framework to determine whether pretrained convolutional neural networks significantly outperform a custom baseline model in multi-class vehicle damage classification. A dataset of 891 labeled vehicle images categorized into heavy, medium, light, and normal damage was partitioned using stratified sampling (70% training, 15% validation, 15% testing). Four architectures Baseline (CustomCNN), VGG16, ResNet50, and MobileNetV2 were trained under identical preprocessing and optimization settings with two training durations (30 and 50 epochs). Five-fold cross-validation and paired t-test analysis were applied to assess statistical significance. At 30 epochs, MobileNetV2 achieved the highest accuracy (75.76%), while at 50 epochs VGG16 obtained the best performance (78.03%). Extending training duration did not produce statistically significant improvement ($p > 0.05$). Pretrained architectures significantly outperformed the baseline model, whereas ResNet50 did not demonstrate superior performance. The novelty of this study lies in its statistically validated comparative framework. Although limited by moderate dataset size and single-source imagery, the findings provide practical guidance for selecting efficient convolutional neural networks in vehicle damage classification systems.

Keywords: computer vision, convolutional neural networks, cross-architecture evaluation, statistical validation, vehicle damage classification

INTRODUCTION

Vehicle damage severity assessment is a fundamental component of automotive insurance claim processing and repair cost estimation. Conventional inspection procedures rely heavily on manual visual evaluation, making the assessment process subjective and potentially inconsistent across inspectors. Such variability may affect claim approval decisions, repair cost calculation, and overall operational efficiency. As intelligent systems continue to be integrated into transportation and insurance services, automated image-based vehicle damage classification has become increasingly important to enhance objectivity, scalability, and standardization.

Recent developments in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved image classification through hierarchical feature learning. CNN-based architectures have demonstrated strong performance in structural damage detection, infrastructure inspection, and road surface analysis. In vehicle-related applications, pretrained CNN models have been widely adopted for automated damage identification and severity classification. Reported classification accuracies in comparable multi-class damage recognition tasks generally range between approximately 70% and 85%, depending on dataset size, class granularity, and experimental configuration.

Despite these advances, important methodological limitations remain. Many prior studies evaluate individual architectures independently or propose modifications to specific pretrained networks without conducting systematic cross-architecture comparisons under identical preprocessing and optimization conditions. As a result, differences in reported performance may stem from variations in dataset partitioning, augmentation strategies, or

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

training configurations rather than intrinsic architectural superiority. Furthermore, evaluation procedures frequently rely on a single train–test split without cross-validation or inferential statistical testing, limiting the reliability and generalizability of reported improvements. Although deeper architectures are often assumed to yield better accuracy, limited evidence demonstrates whether such gains are statistically significant when applied to moderate-sized vehicle damage datasets. These limitations indicate a clear quantitative and methodological gap in statistically validated cross-architecture evaluation.

To address this gap, this study develops a controlled experimental framework for comparative evaluation of CNN-based architectures in multi-class vehicle damage severity classification. Four models Baseline (CustomCNN), VGG16, ResNet50, and MobileNetV2 are trained under identical preprocessing pipelines, consistent dataset partitioning, and uniform optimization parameters to ensure experimental fairness. Two training durations (30 and 50 epochs) are implemented to analyze convergence behavior and performance stability. In addition, 5-fold cross-validation and paired t-test analysis are employed to determine whether observed performance differences are statistically significant rather than partition-dependent.

This study is guided by three research questions: whether pretrained architectures significantly outperform a custom-designed CNN under controlled experimental conditions; whether extending training duration from 30 to 50 epochs results in statistically significant performance improvement; and how architectural complexity influences the trade-off between classification accuracy and computational efficiency. It is hypothesized that pretrained architectures will outperform the Baseline (CustomCNN) however, deeper and more complex models will not necessarily provide statistically significant gains under moderate dataset constraints.

The contributions of this research are threefold. First, it establishes a statistically grounded and experimentally controlled comparative framework for evaluating CNN architectures in vehicle damage severity classification. Second, it integrates cross-validation and inferential statistical testing to improve methodological rigor beyond single-split evaluation. Third, it provides empirical evidence on efficiency accuracy trade-offs and class-level boundary ambiguity, offering practical insights for deployment in resource-constrained insurance and inspection environments.

By shifting the focus from isolated architecture benchmarking to statistically validated comparative evaluation, this study strengthens methodological reliability in automated vehicle damage assessment research.

LITERATURE REVIEW

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved image classification performance across diverse visual recognition tasks. Comprehensive surveys emphasize that hierarchical feature extraction, transfer learning, and architectural efficiency are central factors influencing model generalization and computational trade-offs. Prior studies also highlight that performance gains are not solely determined by network depth but by the interaction between architecture design, dataset scale, and training configuration.

In damage-related image analysis, CNN-based approaches have demonstrated strong capability in structural and infrastructure inspection. Structural damage identification models have shown robustness under experimental validation, while road damage detection systems have achieved reliable classification performance for safety monitoring applications. These studies confirm that CNN architectures can effectively capture complex surface irregularities and deformation patterns from image data. However, most of these works primarily focus on detection accuracy within specific architectures rather than conducting controlled cross-architecture comparisons under identical experimental conditions.

In vehicle damage assessment, several studies have proposed automated classification frameworks. Multi-class vehicle exterior damage detection models have reported promising performance, particularly when leveraging pretrained networks. Architectural enhancements and optimization strategies have also been introduced to improve residual-based networks. Nonetheless, the majority of these studies concentrate on detection frameworks, enhancement of a single pretrained model, or performance reporting without inferential statistical validation. Direct comparative evaluations across multiple pretrained architectures using consistent preprocessing pipelines, uniform training configurations, and cross-validation procedures remain limited.

Comparative Overview of Previous Studies

In table 1 the comparative synthesis indicates recurring methodological patterns. Many studies focus on improving a single architecture or proposing architectural refinements without systematically comparing multiple pretrained networks under identical experimental settings. Evaluation procedures frequently rely on single data splits, limiting statistical reliability. Moreover, multi-class severity classification receives less emphasis compared to binary damage detection tasks. These tendencies suggest that while CNN effectiveness is well established, rigorous comparative validation remains underexplored.

*name of corresponding author
This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1 summarizes representative prior studies in damage-related image classification and highlights methodological characteristics relevant to this research positioning

Study	Task Focus	Architecture Strategy	Evaluation Strategy	Identified Limitation
Khan (2021)	Multi-class vehicle damage classification	Pretrained CNN	Accuracy-based evaluation	No statistical cross-architecture comparison
Ruitenbeek & Bhulai (2022)	Vehicle damage detection	CNN-based detection framework	Single evaluation setting	Emphasis on detection rather than severity levels
Qaddour & Siddiqa (2023)	Automated damaged vehicle estimation	Enhanced deep learning architecture	Performance improvement reporting	Limited inferential validation
Kyu & Woraratpanya (2026)	Car damage detection and classification	CNN-based models	Experimental evaluation	No controlled multi-architecture comparison
Wan et al. (2024); Zhao & Leong (2025); Ambar (2025)	ResNet optimization	Improved ResNet variants	Model-specific enhancement experiments	Not evaluated under uniform comparative protocol

A critical examination of prior studies indicates that performance improvements are frequently reported without assessing statistical reliability or cross-architecture robustness. Consequently, it remains unclear whether observed accuracy gains stem from architectural superiority or experimental variability. This methodological ambiguity highlights the need for controlled and statistically validated comparative frameworks.

Research Gap and Positioning of the Present Study

Although previous research confirms the effectiveness of CNN-based approaches for damage detection and classification, a methodological gap persists in the statistically validated comparison of multiple pretrained architectures for multi-class vehicle damage severity classification under consistent experimental protocols. The literature reveals limited use of cross-validation combined with inferential statistical testing to assess whether observed performance differences are statistically reliable. Furthermore, detailed analysis of class-level boundary ambiguity in fine-grained severity classification is rarely emphasized.

This study addresses these gaps by performing a structured and controlled comparison of CustomCNN, VGG16, ResNet50, and MobileNetV2 using identical preprocessing steps, consistent training configurations, 5-fold cross-validation, and paired statistical testing. In addition, per-class performance analysis and confusion matrix evaluation are conducted to examine boundary ambiguity between adjacent severity levels. The contribution of this research therefore lies in methodological rigor and statistically grounded comparative evaluation rather than proposing a novel architecture. Through this positioning, the study advances understanding of how pretrained CNN architectures behave under moderate data constraints in multi-class vehicle damage severity classification.

METHOD

Research Design

This study adopts a quantitative experimental design to systematically compare the performance of multiple Convolutional Neural Network (CNN) architectures for multi-class vehicle damage severity classification. The overall experimental workflow is presented in Fig. 1.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Fig. 1 Complete experimental pipeline including data collection, verification, preprocessing, model training under two epoch configurations, performance evaluation, and comparative analysis.

Two training duration scenarios were implemented to analyze convergence behavior and performance stability: 30 epochs and 50 epochs. Both configurations were consistently applied to all evaluated models to ensure a fair and controlled comparison of learning dynamics and classification performance. To minimize experimental bias, identical preprocessing procedures, dataset partitioning strategies, hyperparameter configurations, optimization settings, and evaluation metrics were maintained across all architectures. This controlled experimental setup ensures that performance differences primarily reflect architectural characteristics rather than inconsistencies in training configuration.

The methodological framework aligns with established CNN based image classification practices as discussed in recent deep learning literature (Alzubaidi et al., 2021; Bhatt et al., 2021). For pretrained architectures, transfer learning was implemented to leverage prior feature representations learned from large-scale datasets, a strategy widely adopted in recent classification studies (Ruitenbeek & Bhulai, 2022; Wan et al., 2024). Model optimization was performed using the Adam optimizer due to its adaptive learning rate mechanism and efficient convergence behavior in deep neural networks (Ambar, 2025; Zhao & Leong, 2025).

Categorical cross-entropy was applied as the loss function for multi-class classification, ensuring appropriate probabilistic modeling of class predictions. Performance evaluation was conducted using consistent metrics across all architectures to enable objective comparative analysis of classification accuracy, convergence characteristics, and computational efficiency.

Dataset Description

The dataset was collected from publicly available vehicle damage image repositories and supplemented with manually verified images to ensure label accuracy. All images were reviewed to confirm correct severity categorization prior to model training. The dataset consists of 891 labeled vehicle images categorized into four damage severity levels: heavy, medium, light, and normal condition. The images reflect realistic variations in lighting conditions, viewing angles, and structural damage characteristics.

Fig. 2 Representative samples of vehicle damage severity categories: (a) Normal, (b) Light, (c) Medium, (d) Heavy.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Representative examples of each damage severity category are presented in Fig. 2 to illustrate the visual characteristics used in the classification process. The Normal class represents vehicles without visible structural damage, while the Light damage category includes minor surface scratches or small dents. The Medium damage category reflects moderate deformation affecting body panels without significant structural collapse. The Heavy damage category includes severe structural deformation that substantially alters vehicle geometry.

Stratified sampling was employed to preserve proportional class distribution across subsets. The dataset was divided into 70% training data, 15% validation data, and 15% testing data, resulting in 627 training images, 132 validation images, and 132 testing images.

Table 1 Dataset Distribution

Class	Train	Validation	Test	Total
Heavy	249	53	53	355
Medium	129	27	27	183
Light	166	35	35	236
Normal	83	17	17	117
Total	627	132	132	891

The dataset exhibits moderate class imbalance, particularly in the heavy damage category. Therefore, per-class evaluation metrics are emphasized in the analysis.

Data Preprocessing

All images were resized according to the input dimension requirements of each CNN architecture. Pixel normalization was applied using:

$$X_{norm} = \frac{X}{255} \quad (1)$$

Where X represents the original pixel intensity and X_{norm} is the normalized value in the range [0,1]. Class labels were transformed into one-hot encoded vectors. For C classes, the encoded label vector is defined as:

$$y_i = [0,0,\dots,1,\dots,0] \quad (2)$$

Where the value 1 corresponds to the true class index. Data augmentation techniques, including limited rotation, horizontal flipping, and slight zooming, were applied during training to enhance generalization while preserving semantic damage characteristics.

Model Architectures

Four CNN architectures were evaluated in this study: Baseline CNN (CustomCNN), VGG16, ResNet50, and MobileNetV2. The Baseline CNN (CustomCNN) was designed as a conventional convolutional architecture to provide a reference model for controlled performance comparison. The architecture consists of three convolutional layers with a kernel size of 3×3, each followed by a Rectified Linear Unit (ReLU) activation function and a 2×2 max pooling layer for spatial down-sampling. This design follows widely adopted CNN feature extraction principles that emphasize hierarchical representation learning and computational efficiency (Alzubaidi et al., 2021; Bhatt et al., 2021).

Following feature extraction, the resulting feature maps are flattened into a one-dimensional vector and processed through two fully connected dense layers with 128 and 64 neurons, respectively. The output layer employs a Softmax activation function to classify vehicle damage images into four predefined severity classes. This configuration enables multi-class probabilistic prediction consistent with recent deep learning-based classification systems (Ige & Sibiya, 2024).

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

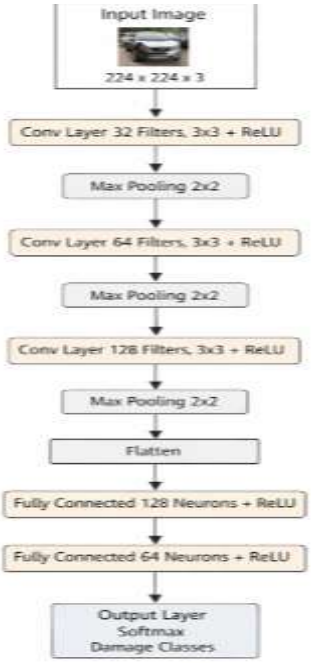

Fig. 3 illustrates the detailed architecture of the proposed Baseline CNN model used in this study.

For comparative evaluation, three pretrained architectures VGG16, ResNet50, and MobileNetV2 were implemented using ImageNet-initialized weights to leverage previously learned large-scale visual representations. Transfer learning strategies have been shown to enhance classification performance on moderate-sized datasets and improve convergence stability (Ruitenbeek & Bhulai, 2022; Wan et al., 2024). The original classification layers of these pretrained networks were replaced with new fully connected layers containing four output neurons corresponding to the vehicle damage severity categories. A Softmax activation function was applied in the final layer to generate normalized class probabilities:

$$P(y = j \mid z) = \frac{e^{z_j}}{\sum_{k=1}^{C} e^{z_k}} \qquad (3)$$

where z_j denotes the logit value for class j, and C is the number of classes.

Table 2 Model Complexity

Model	Total Parameters	Trainable Parameters	Model Size (MB)
Baseline CNN (CustomCNN)	11,169,476	11,169,476	127.86
VGG16	17,926,596	3,211,268	92.96
ResNet50	23,850,500	4,198,404	93.37
MobileNetV2	2,422,468	1,403,908	10.85

The comparison indicates that ResNet50 possesses the highest number of parameters, reflecting greater representational capacity, consistent with recent optimization studies on ResNet based models (Ambar, 2025; Zhao & Leong, 2025). In contrast, MobileNetV2 exhibits significantly lower parameter complexity, making it more suitable for deployment in resource constrained environments while maintaining competitive performance characteristics.

Cross-Validation Strategy

To enhance the robustness of performance estimation and reduce potential bias caused by a single data split, a 5-fold cross-validation strategy was additionally implemented. The dataset was partitioned into five mutually exclusive subsets while preserving class distribution using stratified sampling.

In each fold, four subsets were used for training and one subset for validation/testing. This process was repeated five times so that each subset served once as validation data. Performance metrics were recorded for each fold,

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

and the final results were reported as mean and standard deviation across folds. The use of cross-validation ensures that the reported performance reflects generalized model behavior rather than variance caused by a particular data partition. The 5-fold cross-validation procedure was applied independently to each evaluated architecture (Baseline CNN, VGG16, ResNet50, and MobileNetV2) to ensure fair and unbiased comparative analysis.

Reproducibility and Random Seed Control

To minimize stochastic variability inherent in deep learning training, all experiments were conducted using a fixed random seed (42). Random seed initialization was applied to NumPy, TensorFlow, and Python random modules to ensure deterministic behavior during dataset shuffling and weight initialization. This controlled setup enhances experimental reproducibility and reduces randomness-induced performance fluctuations.

Early Stopping and Regularization

To prevent overfitting and improve generalization stability, early stopping was applied based on validation loss monitoring. Training was terminated when validation loss failed to improve for five consecutive epochs (patience = 5), and the best-performing model weights were restored.

Additionally, dropout regularization with a rate of 0.5 was applied to the fully connected layers in all architectures. Data augmentation techniques, including rotation, horizontal flipping, and zooming, were employed during training to enhance robustness against minor spatial variations.

Hardware and Software Environment

All experiments were conducted using Google Colab equipped with an NVIDIA Tesla T4 GPU (12GB RAM). The implementation utilized TensorFlow 2.x and Python 3.x. This configuration ensured sufficient computational capacity for consistent and reproducible deep learning training across all evaluated architectures.

Statistical Analysis and Confidence Estimation

To address stochastic training variance and strengthen inferential validity, model performance was analyzed using statistical evaluation techniques. For each model, performance metrics obtained from 5-fold cross-validation were aggregated to compute the mean accuracy, standard deviation, and 95% confidence interval. Confidence intervals were calculated using the Student's t-distribution:

$$\mathrm{CI} = \mu \pm \mathrm{t\_}(\alpha/2, \mathrm{n}-1) \times (\sigma/\sqrt{\mathrm{n}}) \qquad (4)$$

where μ denotes the mean performance, σ the standard deviation, and n the number of folds. To determine whether observed performance differences between models were statistically significant, Paired t-tests were performed on fold-wise performance scores between competing architectures to account for correlated samples across identical data splits. Statistical significance was determined at the 95% confidence level, with p-values below 0.05 considered significant.

Performance evaluation was conducted using two complementary strategies. Cross-validation results were used for statistical comparison and confidence estimation, while final model performance was reported on the independent hold-out test dataset. This inferential evaluation framework ensures that reported improvements are not attributed solely to stochastic variance.

Reproducibility Statement

The complete implementation code, hyperparameter configurations, and experimental setup are available upon request to ensure transparency and reproducibility of the reported findings.

Experimental Setup

All models were trained under identical configurations to ensure objective comparison. The Adam optimizer was used with an initial learning rate of 0.001 to update parameters according to:

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \qquad (5)$$

where $\theta_t$represents model parameters at iteration t, $\alpha$is the learning rate, and $\hat{m}_t$and $\hat{v}_t$are bias-corrected moment estimates. The categorical cross-entropy loss function was applied:

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic}) \quad (6)$$

Training was conducted using a batch size of 32 under two epoch scenarios: 30 epochs and 50 epochs. Early stopping with a patience of 5 epochs was applied based on validation loss monitoring. For cross-validation experiments, the 5-fold procedure was applied to the training portion of the dataset, while the independent 15% test set was reserved exclusively for final model evaluation.

Table 3 Summary of Experimental Setup

Component	Specification
Total Images	891
Number of Classes	4
Data Split Ratio	70% Train, 15% Validation, 15% Test
Training Samples	627
Validation Samples	132
Testing Samples	132
Image Processing	Resize and Normalization
Label Encoding	One-hot Encoding
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Batch Size	32
Epoch Variations	30 and 50
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, ROC-AUC

Table 3 presents the standardized experimental configuration applied to all models to ensure fair comparison. The dataset consists of 891 images categorized into four classes and divided using a stratified split of 70% training, 15% validation, and 15% testing to maintain class proportionality. All images were resized and normalized to improve numerical stability during training. One-hot encoding was applied to support multi-class classification with Softmax activation. The models were trained using the Adam optimizer and categorical cross-entropy loss function, which are appropriate for multi-class prediction tasks.

A batch size of 32 was used to balance computational efficiency and convergence stability. Two training scenarios (30 and 50 epochs) were implemented to analyze convergence behavior and generalization performance. Evaluation was conducted using Accuracy, Precision, Recall, F1-score, and ROC-AUC to provide comprehensive performance assessment across all classes. By maintaining identical experimental settings, performance differences are attributed to architectural characteristics rather than training variability.

Performance Evaluation

Model performance was evaluated using two complementary strategies to ensure both generalization assessment and statistical reliability. Final predictive performance was computed on the independent hold-out test dataset, while statistical comparisons were derived from fold-wise cross-validation results.

Accuracy was computed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

For multi-class classification, precision and recall were computed on a per-class basis:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

The F1-score was defined as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

Macro-averaging was applied across all classes to ensure balanced evaluation under moderate class imbalance conditions. Receiver Operating Characteristic (ROC) curves were generated using a one-vs-rest (OvR) approach for multi-class settings, and the Area Under the Curve (AUC) was computed to evaluate the discriminative

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

capability of each architecture. Confusion matrices were analyzed to examine inter-class misclassification patterns and error distribution.

In addition to predictive performance, computational efficiency was assessed based on parameter count and model size to analyze trade-offs between classification accuracy and deployment feasibility. For inferential analysis, statistical comparisons between models were performed using paired t-tests on fold-wise cross-validation scores. Performance variability was quantified using standard deviation and 95% confidence intervals to ensure robustness against stochastic training variance.

RESULT

Results at 30 epochs

Table 4 presents the classification performance of all evaluated models after 30 training epochs. At 30 epochs, MobileNetV2 achieved the highest accuracy (75.76%), followed closely by VGG16 (74.24%). Both pretrained models outperformed Baseline CNN (CustomCNN) and ResNet50. The highest F1-score was also obtained by MobileNetV2, indicating balanced class prediction performance.

Table 4 Performance of CNN Models at 30 Epochs

Model	Accuracy	Precision	Recall	F1-score
MobileNetV2	0.7576	0.7466	0.7576	0.7408
VGG16	0.7424	0.7358	0.7424	0.7380
Baseline CNN (CustomCNN)	0.6212	0.5958	0.6212	0.5965
ResNet50	0.5985	0.4794	0.5985	0.5254

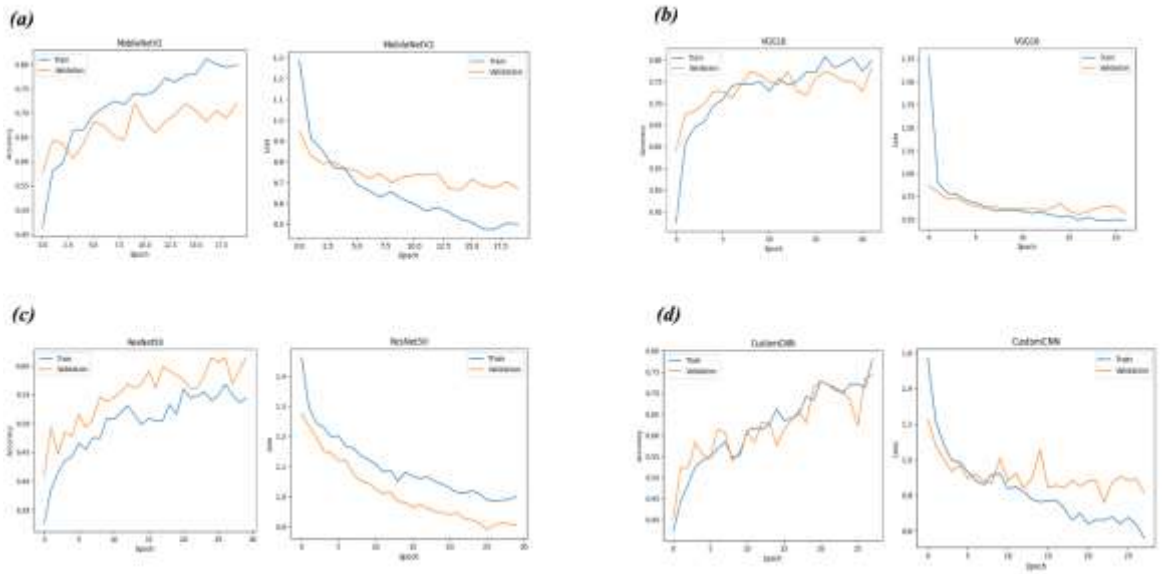

Fig. 4 shows the training and validation accuracy and loss curves at 30 epochs for each model: (a) MobileNetV2, (b) VGG16, (c) ResNet50, and (d) Baseline CNN (CustomCNN).

Results at 50 epochs

Table 5 presents the performance after increasing the training duration to 50 epochs. After 50 epochs, VGG16 achieved the highest accuracy (78.03%), followed by MobileNetV2 (77.27%). Both models showed performance improvement compared to their 30-epoch results.

Table 5 Performance of CNN Models at 50 Epochs

Model	Accuracy	Precision	Recall	F1-score
VGG16	0.7803	0.7573	0.7803	0.7481
MobileNetV2	0.7727	0.7604	0.7727	0.7471
ResNet50	0.6364	0.5063	0.6364	0.5626
Baseline CNN (CustomCNN)	0.5985	0.5795	0.5985	0.5207

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

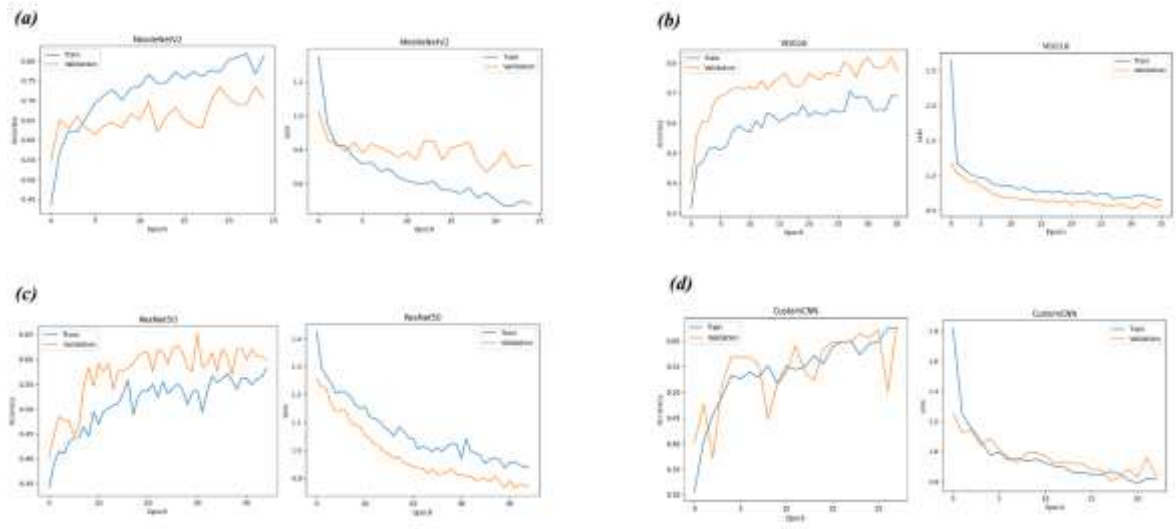

Fig. 5 shows the training and validation accuracy and loss curves at 50 epochs for each model: (a) VGG16, (b) MobileNetV2, (c) ResNet50, and (d) Baseline CNN (CustomCNN).

Comparative Analysis of 30 and 50 Epoch Results

Table 6 summarizes model performance under a single train–test split at 30 and 50 epochs. VGG16 and MobileNetV2 appear to achieve higher accuracy when trained for 50 epochs, while ResNet50 also shows improvement. Baseline CNN (CustomCNN) exhibits a slight decrease in performance, possibly indicating overfitting.

Table 6 Comparison of Model Performance at 30 and 50 Epochs

Model	Accuracy (30)	Accuracy (50)	Δ Accuracy	F1 (30)	F1 (50)
VGG16	0.7424	0.7803	+0.0379	0.7380	0.7481
MobileNetV2	0.7576	0.7727	+0.0151	0.7408	0.7471
ResNet50	0.5985	0.6364	+0.0379	0.5254	0.5626
Baseline CNN (CustomCNN)	0.6212	0.5985	−0.0227	0.5965	0.5207

These single-split results should be interpreted cautiously. The 5-fold cross-validation analysis indicates that the performance improvements for VGG16 and MobileNetV2 are not consistently observed across different data partitions. Paired t-test confirms that the difference between 30 and 50 epochs is not statistically significant ($p > 0.05$). Given the small dataset (891 images), differences of 1–4% may fall within expected statistical variability.

Cross-Validation Results

5-fold cross-validation was conducted for all models at 30 and 50 epochs to reduce potential bias and ensure robustness. Table 7 presents the mean accuracy, standard deviation, and mean F1-score.

Table 7 5-Fold Cross-Validation Results (Mean ± Std)

Model	Epoch	Mean Accuracy	Std Dev	Mean F1
VGG16	30	0.7428	±0.0279	0.7481
VGG16	50	0.7405	±0.0428	0.7437
MobileNetV2	30	0.7157	±0.0338	0.7199
MobileNetV2	50	0.7088	±0.0505	0.7152
ResNet50	30	0.5101	±0.0483	0.4270
ResNet50	50	0.5643	±0.0243	0.4978
Baseline CNN (CustomCNN)	30	0.6660	±0.0367	0.6764
Baseline CNN (CustomCNN)	50	0.6682	±0.0156	0.6769

Cross-validation results show that VGG16 achieves the highest average accuracy at 30 epochs, slightly decreasing at 50 epochs, with increased variability. MobileNetV2 shows a similar trend. ResNet50 improves at 50 epochs with reduced variability, and CustomCNN shows minimal improvement with lower variability. Overall, differences between 30 and 50 epochs are within fold-level variability (±0.02 to ±0.05).

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15939
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Performance Evaluation of the Highest Performing Model

Since VGG16 achieved the highest classification accuracy, further evaluation was conducted at 50 epochs to examine class-level prediction performance.

Table 8 Per-Class Metrics of VGG16 at 50 Epochs

Class	Precision	Recall	F1-score	Support
Heavy	0.80	0.91	0.85	53
Normal	1.00	1.00	1.00	17
Light	0.66	0.77	0.71	35
Medium	0.64	0.33	0.44	27
Accuracy	—	—	0.77	132
Macro Avg	0.78	0.75	0.75	132
Weighted Avg	0.76	0.77	0.75	132

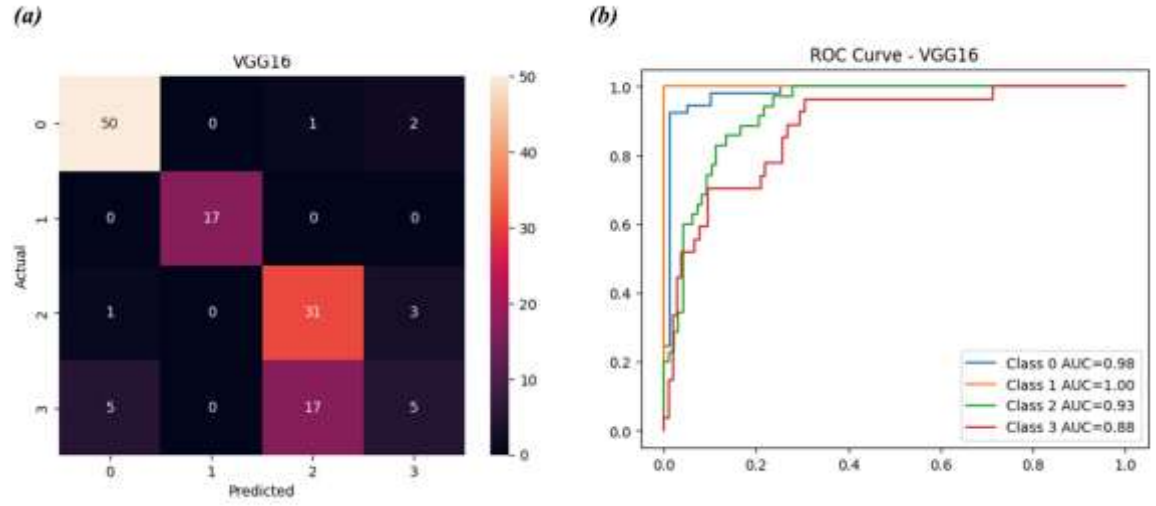

Fig. 6 Performance evaluation of VGG16 at 50 epochs: (a) confusion matrix and (b) ROC curve.

The confusion matrix (Fig. 6a) and ROC curve (Fig. 6b) further illustrate VGG16's discrimination capability. The model performs best on Normal and Heavy classes, while Medium shows lower recall, indicating misclassifications. Overall metrics confirm balanced performance across classes.

Statistical Significance Test

A paired t-test was conducted on fold-level accuracies from 5-fold cross-validation to determine if extending training from 30 to 50 epochs significantly affects performance. The test yielded a t-statistic of −1.1924 and p-value = 0.3188. Since $p > 0.05$, the performance difference is not statistically significant, indicating that observed variations are likely due to dataset variability rather than true improvement.

DISCUSSIONS

Comparative Interpretation with Previous Studies

The results confirm established findings in deep learning literature that pretrained convolutional neural networks provide strong performance under limited-data conditions. As discussed in Journal of Big Data and Electronics, transfer learning enables hierarchical feature reuse from large-scale datasets, reducing the dependency on extensive task-specific data. The superior performance of VGG16 and MobileNetV2 compared to the custom CNN architecture in this study supports this theoretical premise.

In vehicle damage classification research, studies such as IEEE Access (Khan, 2021) and Machine Learning with Applications (Ruitenbeek & Bhulai, 2022) report multi-class classification accuracies typically ranging between approximately 70% and 85%, depending on dataset size and class granularity. The highest accuracy obtained in this study (78.03% using VGG16) lies within this reported range, indicating competitive but not inflated performance under moderate data constraints. Similarly, James (2025) demonstrated that pretrained CNN architectures tend to yield consistent yet incremental improvements over baseline models. The marginal difference

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

observed between 30 and 50 epochs in the present study aligns with this pattern, suggesting diminishing returns beyond a certain training depth.

Research on optimized residual networks (Wan et al., 2024; Zhao & Leong, 2025; Ambar, 2025) highlights that deeper architectures such as ResNet50 require architectural refinement or advanced optimization strategies to outperform simpler models consistently. In this study, ResNet50 did not surpass VGG16 despite its higher parameter count. This outcome reinforces findings in structural and road damage detection studies (Bouhsissin et al., 2025; Duran et al., 2024), where feature separability and data representation were found to exert stronger influence on performance than architectural depth alone.

Theoretical Implications on Model Capacity and Generalization

The results reflect the interaction between model capacity, sample size, and generalization behavior. Increasing model complexity without proportional data expansion does not necessarily yield improved predictive performance. The absence of statistically significant differences between 30 and 50 epochs indicates performance saturation, where additional training primarily increases computational cost rather than discriminative capability.

From a learning-theoretic perspective, this suggests that the effective capacity of deeper models was not fully utilized due to limited sample complexity. Transfer learning mitigates this constraint by leveraging pretrained feature hierarchies; however, once decision boundaries are sufficiently stabilized, additional epochs do not substantially alter generalization behavior. This explains why VGG16 achieved stable performance without requiring extended training cycles.

MobileNetV2's near-equivalent performance further illustrates the efficiency–accuracy trade-off. Depthwise separable convolutions maintain representational strength while reducing parameter redundancy. The results therefore support the argument that parameter efficiency can achieve comparable discrimination when pretrained features are transferable and classes are moderately separable.

Class-Level Error Characteristics and Boundary Ambiguity

The per-class analysis reveals that prediction errors are not uniformly distributed. The moderate damage class exhibits the lowest recall, indicating systematic boundary ambiguity rather than random misclassification. Confusion patterns predominantly occur between adjacent severity levels (light–moderate and moderate–heavy), while extreme classes show relatively higher precision.

This asymmetric error distribution suggests that the principal challenge lies in subtle visual transitions, such as minor variations in dent depth, surface deformation continuity, and texture irregularities. These characteristics create overlapping feature representations in latent space, reducing inter-class separability. Similar observations have been reported in fine-grained damage classification studies (Jiang, 2024; Qaddour & Siddiqa, 2023), where classification difficulty stems from visual similarity rather than model underfitting.

Therefore, increasing network depth alone is unlikely to resolve misclassification at severity boundaries. Approaches emphasizing spatial attention mechanisms, higher-resolution feature extraction, or contrastive feature learning may provide more effective improvements than simply extending training duration or increasing parameter count.

Statistical Reliability and Threats to Validity

The use of 5-fold cross-validation combined with paired t-test analysis strengthens inferential reliability by reducing dependency on a single data split. The absence of statistically significant performance differences between 30 and 50 epochs indicates that apparent improvements in single-run experiments may reflect partition-specific variance rather than genuine learning gains.

However, several limitations must be acknowledged. First, the dataset size remains moderate, which constrains sample complexity and may limit the generalizability of deeper architectures. Second, external validation on independent datasets was not conducted, restricting cross-domain generalization claims. Third, advanced optimization techniques for residual networks were not implemented, which may have limited the achievable performance of ResNet50. Finally, visual severity labeling inherently involves subjective interpretation, potentially contributing to inter-class overlap and boundary ambiguity.

These limitations suggest that future research should incorporate larger and more diverse datasets, cross-dataset validation protocols, and architecture-specific optimization strategies. Nonetheless, within the experimental scope, the findings consistently demonstrate that transfer learning and parameter-efficient architectures provide stable and statistically reliable performance for multi-class vehicle damage classification under moderate data conditions.

CONCLUSION

This study investigated the comparative performance of pretrained CNN architectures for multi-class vehicle damage classification under moderate data constraints. The results demonstrate that transfer learning–based models provide stable and competitive performance, with VGG16 achieving the highest accuracy (78.03%) and

*name of corresponding author
This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

F1-score at 50 epochs. However, statistical testing using 5-fold cross-validation and paired t-test analysis indicates that the performance difference between 30 and 50 epochs is not statistically significant, suggesting performance saturation beyond a certain training duration.

MobileNetV2 achieved near-equivalent performance while maintaining substantially lower parameter complexity, confirming the existence of an efficiency–accuracy trade-off. ResNet50, despite having the largest parameter capacity, did not outperform the other architectures, indicating that increased model depth does not necessarily translate into superior generalization when dataset size is limited.

Confusion matrix analysis revealed that misclassifications predominantly occur between adjacent severity categories, particularly between light and moderate damage. This finding answers the second research question by demonstrating that classification difficulty arises primarily from inter-class visual similarity and boundary ambiguity rather than insufficient architectural complexity.

The primary contribution of this study lies in providing a statistically validated comparative evaluation of pretrained CNN architectures for multi-class vehicle damage severity classification under constrained data conditions. Unlike single-split evaluations commonly reported in similar studies, this work integrates cross-validation and inferential statistical testing to assess the reliability of observed performance differences. The findings contribute empirical evidence that parameter-efficient architectures can achieve competitive results without requiring extended training or excessive model depth.

Nevertheless, several limitations must be acknowledged. The dataset size remains moderate, which constrains sample complexity and may limit the full utilization of deeper architectures. This dataset limitation constitutes a potential threat to external validity, as generalization to larger or more diverse real-world damage scenarios cannot be fully guaranteed. Additionally, the absence of external dataset validation restricts cross-domain generalization claims.

Future research should address these limitations by expanding dataset scale and diversity, incorporating external validation protocols, applying systematic hyperparameter optimization, and exploring feature-enhancement strategies such as attention mechanisms or contrastive learning. Such improvements may enhance fine-grained damage discrimination while maintaining computational efficiency for deployment in resource-constrained environments.

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al Dujaili, A., Duan, Y., Al Shamma, O., Santamaría, J., Fadhel, M. A., Al Amidie, M., & Farhan, L. (2021). *Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions*. Journal of Big Data, 8, 53. https://doi.org/10.1186/s40537-021-00444-8

Ambar, M. K. (2025). *Optimizing ResNet-50 for multiclass classification: A multi-stage learning approach*. IEEE Access, 13, 142517–142534.

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., & Pandya, S. (2021). *CNN variants for computer vision: History, architecture, application, challenges and future scope*. Electronics, 10(20), 2470. https://doi.org/10.3390/electronics10202470

Bouhsissin, S., Assemlali, H., & Sael, N. (2025). *Enhancing road safety: A convolutional neural network based approach for road damage detection. Machine Learning with Applications*, 20, 100668. https://doi.org/10.1016/j.mlwa.2025.100668

Duran, B., Emory, D., Azam, Y. E., & Linzell, D. G. (2024). *A novel CNN architecture for robust structural damage identification via strain measurements and validation via full-scale experiments*. Engineering Structures, 304, 117486.

Ige, A. O., & Sibiya, M. (2024). *State-of-the-art in 1D convolutional neural networks: A survey*. IEEE Access, 12, 15023–15045.

Janiesch, C., & Heinrich, K. (2021). *Machine learning and deep learning*. Electronic Markets, 31(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Jiang, Y. (2024). *Road damage detection and classification using deep neural networks*. Discover Applied Sciences, 6, 324. https://doi.org/10.1007/s42452-024-06129-0

James, A. (2025). *Comparative analysis of pre-trained CNN architectures for damage detection*. Multimedia Tools and Applications. Advance online publication.

Khan, M. H. (2021). *Automated detection of multi-class vehicle exterior damages using deep learning*. IEEE Access, 9, 159234–159245.

Kyu, P., & Woraratpanya, K. (2026). *Car damage detection and classification*. In Proceedings of the ACM International Conference on Multimedia (pp. 1120–1128). https://doi.org/10.1145/3406601.3406651

Qaddour, J., & Siddiqa, S. A. (2023). *Automatic damaged vehicle estimator using enhanced deep learning algorithm*. Intelligent Systems with Applications, 18, 200231.

Ruitenbeek, R. E. van, & Bhulai, S. (2022). *Convolutional neural networks for vehicle damage detection*. Machine Learning with Applications, 9, 100330.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Wan, J., Li, B., Wang, K., Teng, X., Wang, T., & Mao, B. (2024). An *improved ResNet50 for environment image classification*. Expert Systems with Applications, 237, 121556.

Zhao, C. H., & Leong, W. Y. (2025). *Optimisation solutions and simple innovative solution research on ResNet50 model*. Applied Sciences, 15(1), 112.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.