SN Computer Science (2025) 6:632
https://doi.org/10.1007/s42979-025-04181-0

ORIGINAL RESEARCH

Bridging Reality and Synthetics: Optimizing Image Classification with Hybrid AI-Generated and Real-World Datasets

Abdallah Tariq Hasan Alabed[1] · Jawad Rasheed[1,2,3] · Mirsat Yesiltepe[4] · Shtwai Alsubai[5] · Tunc Asuroglu[6,7]

Received: 3 February 2025 / Accepted: 29 June 2025

Abstract

The rapidly growing revolution of generative Artificial Intelligence software has moved into the counseling and disseminating synthetic images, thereby establishing a new paradigm for machine learning models. This study investigates the impact of combining real-world and AI-generated synthetic images on the performance of image classification models. Using three traffic-related datasets—potholes, speed bumps, and traffic lights—we applied data augmentation and tested seven configurations with varying real-to-synthetic image ratios. The DenseNet201 model, fine-tuned with the Adam optimizer, was used for all experiments. Results show that a 1:3 real-to-synthetic ratio enhances classification accuracy and generalization, with the highest validation accuracy reaching 97.36%. Our findings demonstrate that synthetic data, when properly integrated, serves as a cost-effective and scalable complement to real data, especially in scenarios with limited labeled samples.

Keywords Image classification · Synthetic images · Machine learning · DenseNet201 · Adam

Introduction

Machine learning models have seen significant advancements in recent years. Generative artificial intelligence has entirely changed the field of computer vision by being able to use the generation of synthetic datasets at scale and high quality, particularly in the field of image classification [1, 2]. As the demand for diverse and large data sets continues to grow, synthetic data is cost-efficient, scalable, and controllable; it is unclear whether it is effective in combination with real-world data [3]. This study aims to investigate how real and synthetic images could be best integrated for image classification tasks and explore the best methods of combining real and synthetic images for image classification tasks while trying to address issues such as data scarcity, domain gaps, and overfitting, while the main goal of the research is to optimize image classification performance.

Many studies assess the effectiveness of synthetic datasets in machine-learning tasks: for example [4], focuses on quality and diversity in the dataset and shows how convolutional neural networks (CNNs) can distinguish with an accuracy of 88% between real and synthetically generated images. Similarly [5], showcases both scalable and cost-effective synthetics and demonstrates the improvements they offer in performance on applications such as object detection and segmentation. However, synthetic datasets have benefits and challenges regarding performance gaps compared to original data, as quoted in "Quantifying Performance Gaps." In the study, real images were found to outperform synthetic ones due to domain shifts.

✉ Jawad Rasheed
jawad.rasheed@izu.edu.tr

1 Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul 34303, Turkey

2 Department of Software Engineering, Istanbul Nisantasi University, Istanbul 34398, Turkey

3 Applied Science Research Center, Applied Science Private University, Amman, Jordan

4 Department of Intelligent Transportation System, Istanbul University, Istanbul, Turkey

5 Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, P.O. Box 151, Al-Kharj 11942, Saudi Arabia

6 Faculty of Medicine and Health Technology, Tampere University, Tampere 33720, Finland

7 VTT Technical Research Centre of Finland, Tampere 33101, Finland

Published online: 11 July 2025

There is significantly more work arguing for adding real data with synthetic data to fill these gaps in the literature review section. For instance, Aml Yasser in [6] argued that AI-generated datasets, without real-world data, would not fully replace them; however, it would augment if rightly combined with real data in enhancing robustness and generalization of the models. This trend of combining real with synthetic has already proven successful in previous works, as demonstrated by the hybrid dataset's success for improved classification accuracy [7]. The ongoing study extends the findings by combining multiple datasets pertinent to speed bumps, potholes, and traffic lights into a balanced configuration of real and AI-generated images. The study highlights how various configurations of real alongside synthetic data would impact model performance, with an insight into the optimal ratio to achieve high accuracy plus robust generalization.

The use of synthetic data in a structured way and image augmentation techniques to remedy the limitations previously discussed in the current literature and to further the ongoing research on improvement in the machine learning model application for diverse image data [8]. Through this study, we intend to show the feasibility and advantages of improving classification performance through real and synthetic datasets to advance the state of the art using AI-generated images.

Related Work and Literature Review

In [9], Resmini et al. introduced the Text-Conditioned Knowledge Recycling (TCKR) pipeline, which utilizes generative diffusion models to produce synthetic datasets for image classification. Their approach combines dynamic image captioning, efficient fine-tuning of diffusion models, and generative knowledge distillation. Remarkably, models trained entirely on this synthetic data achieved accuracy comparable to, or better than, those trained on real images, while also demonstrating improved resistance to Membership Inference Attacks. In contrast, our study adopts a hybrid data strategy using a 1:3 real-to-synthetic ratio and employs the DenseNet201 architecture optimized with Adam, focusing on both performance and practicality. While they emphasize data privacy and advanced generative techniques, our work offers a more accessible and resource-efficient solution, highlighting a balanced trade-off between accuracy, complexity, and applicability.

In [10] Nguyen et al. present a theoretical framework for few-shot image classification that quantifies distribution discrepancies between real and synthetic data and introduces an algorithm leveraging prototype learning to optimize data partitioning and training. Their method outperforms state-of-the-art techniques across several datasets and demonstrates that strategically integrated synthetic data can significantly enhance model generalization. While both studies address data scarcity, our work focuses on an empirical, application-specific approach using a 1:3 real-to-synthetic ratio for traffic-related datasets, whereas Nguyen et al. provide a more generalizable, theory-driven solution. Their findings complement our research by offering deeper theoretical insights, though the practical application of their framework may involve added complexity in real-world scenarios.

The study [4] contends within an analysis using a convolutional neural network (CNN)-based model to argue the emerging issue of AI abuse, including deepfakes for cybercrime and fake news. They used two datasets: 1,000 images from Google Images (500 AI and 500 real images) and a smaller self-made dataset with 16 images. The model, comprised of four convolutional layers with max-pooling and dropout, achieves an accuracy of 88% on the Google dataset, with precision and recall scores of 83% and 95% for real images and 94% and 81% for AI-generated images. The accuracy percentage for the smaller self-prepared dataset read 81%, while the scores on F1 were lower. The significance of volume and diversity of data to avert overfitting and attain higher performance is quite prominent in the results from the larger dataset as opposed to the Google dataset. This paper discusses how CNNs may be used for differentiating complicated patterns in images produced from real life against those coming from AI while stressing the need for very reliable detection mechanisms for overcoming challenges posed by deepfake technologies. The findings align with this work's goals, particularly in demonstrating how data quality and size influence model accuracy, reinforcing the approach of combining real and AI-generated datasets while providing benchmarks like precision, recall, and F1-score for performance evaluation.

The study [5] explores the transformative potential of AI-generated images as data sources in visual intelligence, emphasizing their scalability, controllability, and cost-efficiency in creating large, diverse datasets. Synthetic data is increasingly used in image classification, segmentation, object detection, and medical imaging across domains like robotics and autonomous driving. Key resources include fully synthetic datasets (e.g., DiffusionDB, HPD v2) and semi-synthetic ones (e.g., ForgeryNet, DeepArt), evaluated using metrics like FID, R-Precision, mIoU, and mAP. Technologies such as GANs and diffusion models enable fast, customized data generation, improving model robustness and generalization. Our study, then, really is an extension of our own, where the augmentation of real data with AI-generated images is shown through the "1k-real + 1k-AI" configuration to produce overall superior performance. The

same consideration on quality metrics and semi-synthetic datasets finds its parallel in the multi-configuration approach extant in the study's methodology that incorporates real with synthetic data, which then redounds to model generalization and task-specific performance benefits. As to the generative technologies discussion in the paper, they feature how these technologies can develop tailor-made datasets for certain cases, including speed bumps, potholes, and traffic lights, lending credence to the study's finding that balanced integration of synthetic and real data, such as a 1:3 ratio, can improve performance.

Classification of AI-generated images with a Convolutional Neural Network (CNN) that is trained on CIFAKE, a dataset of images consisting of 120,000 images evenly divided between real (from CIFAR-10) and artificial images, is introduced in the study conducted by [11]. The study achieved a high classification accuracy of 92.98%, demonstrating the value of synthetic images for training due to their diversity, scalability, and low cost. It also uses Explainable AI methods like Grad-CAM to interpret feature importance. However, its reliance on the CIFAR-10 dataset and focus on binary classification limit its generalizability to more complex, real-world, and multi-class scenarios. Nonetheless, subtle imperfections in AI-generated images can infuse biases; further corroboration of such findings in this study remains rather preliminary and calls for further exploration. The ethical considerations of using synthetic images were not addressed in detail, although they were considered. The paper overall highlights AI-generated images' merits and feasibility when adopting machine-learning modeling. Still, it stresses the balancing act that they need to integrate complementary real data sources along with themselves.

In another study [12], several challenges of AI-generated synthetic images have been identified and further evaluated comprehensively. Besides, several machine learning models have been assessed for their ability to differentiate real images from fake ones. Again, in this study the CIFAKE dataset has been used, which contains 120,000 labeled images (60,000 real and 60,000 AI-generated). The study compared traditional models like Support Vector Machines (SVM) with advanced deep learning architectures such as ResNet, VGGNet, and DenseNet. The study's findings showed that deep learning models perform noticeably better than traditional ones, with ResNet having a high ROC-AUC score of 0.9958 and DenseNet achieving the highest accuracy of 97.74%. Such findings evidently show the superior capability of deep learning models for complex classification tasks. It states the importance of diversity and quality in the dataset, as in the CIFAKE dataset. The CIFAKE dataset is a strong benchmark for testing model performance and highlights the role of synthetic images in ensuring digital media integrity. It emphasizes the effectiveness of AI-generated images in training and underscores the need for ongoing research into the ethical and technical aspects of synthetic data use.

The research [13] explores the difference in performance between real and AI-generated synthetic images in computer vision. Thus, a study of the relative capabilities of these two approaches in real-world applications has been made once more on the CIFAKE dataset, comprising 60,000 real images from CIFAR-10 and 60,000 AI-generated pictures, under two models: EfficientNet-V2 B0 and another CNN. The CNN attained an accuracy level of 75.26% at moderate F1-score values for both real and synthetic classes, while EfficientNet-V2 B0 surpassed all this performance, proving to distinguish better real versus AI-generated images. The findings showed that although AI-generated pictures can perform competitively in certain scenarios, they have some limitations due to domain shifts, insufficient diversity, and biases in the generative training data. The study highlights a performance gap between synthetic and real images, emphasizing the need for improved AI image generation. It provides a strong evaluation framework for model comparison and offers valuable insights into the challenges and potential of using synthetic data.

A comprehensive analysis of AI-generated synthetic images is presented in the thesis [6], which has addressed several issues regarding training models for machine learning related to data availability and quality, as well as the privacy aspect. The thesis comprises five chapters dealing with super-fast-growing synthetic data and dataset collection methods, followed by model selection and result analysis with suggestions for further research. The thesis evaluates the performance of AI-generated datasets across three datasets individually, achieving a maximum accuracy of 84% when tested on real images but highlighting the limitations of relying solely on synthetic data. It concludes that AI-generated images can enhance diversity and reduce bias, but cannot fully replace real-world datasets. This supports our study's hypothesis for a balanced approach that combines real and synthetic data for better model outcomes. Importantly, our study correlates directly with the work mentioned above, as we employed the same datasets but merged them with the application of data augmentation techniques that the thesis did not utilize. We extend this thesis framework using augmentation and merging datasets to remedy its deficiencies and demonstrate potential improvement in generalization and performance on image classification tasks.

The reviewed papers show how AI-generated datasets both pose challenges and create new opportunities for image classification and distinguishing real from synthetic images in machine learning. Specifically, in [4, 11], CNNs have shown solid performance in classifying AI-generated images, with accuracies of 88% and 92.98%, highlighting

the importance of dataset diversity and balance. The works [12, 13] have also compared advanced deep learning architectures like DenseNet and EfficientNet-V2 B0. Therefore, these models outperform others, with DenseNet achieving 97.74% accuracy and EfficientNet favored for its higher F1 score. In addition [5], presents some arguments on how synthetic datasets can scale and save costs, recommending their use in augmenting more real data to achieve better generalization and robustness. The thesis [6] is directly pertinent to this work, where it examines how AI-generated datasets perform status-wise in isolation. It finds that all limitations in generalization might emerge if only one dataset is used, and thus, it recommends merging datasets and augmentation. This research feeds from such studies, including merging three datasets through augmentation with an optimized realistic-synthetic ratio (for instance, 1:3), resulting in much better model performance and closing the domain gaps identified in the literature.

Methodology

Dataset Selection

In our study, we used three Kaggle datasets to examine how real and AI-generated images impacted the task of image classification. As shown in Table 1, all three datasets relate to traffic scenarios and road safety, including potholes, speed bumps, and traffic lights. Again, each dataset contains real and AI-generated images to give a detailed result on how machine learning models may perform when trained on a mixed dataset.

1. **Generated Speed Bumps**
   This dataset consists of images of roads with speed bumps and without speed bumps under two categories, named 'real' and 'AI-generated'. Real images have been collected from publicly available sources. The synthetically generated images are created using tools like Gemini, Microsoft Copilot, Microsoft Image Creator, Canva, and FreePik. The data contains a diverse set of images that would represent all kinds of road conditions and scenarios. The total dataset consists of 1.4k fake and 1.7k real images [14].

Table 1 Detailed description of the dataset utilized

Label No	Dataset Name	Real Images	Generative Images
Label 1	Speed Bumps	1,700	1,400
Label 2	Street Potholes	1,400	1,500
Label 3	Traffic Lights (Red/Green)	1,000	1,000

2. **Generated Street Potholes**
   This dataset comprises street images with and without potholes and is classified as positives (having potholes) and negatives (without potholes). Like the speed bumps dataset, the current dataset consists of both 'real' and 'AI-generated' images, where the AI-generated images were generated using tools such as Midjourney, Gemini, Microsoft Copilot, and Microsoft Designer Image Creator. The dataset captures from various angles, lighting, and even visibility of potholes to ensure diversity. The dataset comprises 2.9k images. In total, 1.5k are fake, and 1.4k are real [15].
3. **Generated Traffic Lights**
   This dataset contains images of traffic lights categorized as **red** or **green**, with a balanced mix of real and AI-generated images. The real images were sourced from various online datasets, and the synthetic images were created using AI tools such as Midjourney, Gemini, Microsoft Copilot, and Microsoft Designer Image Creator to simulate diverse lighting and environmental conditions. The dataset is of size 2k images total, 1k fake, and 1k real [16].

Figure 1 shows sample photos from the merged datasets. The datasets were designed to offer the realistic and synthetic images needed to evaluate different configurations of pure real and AI-generated datasets. Therefore, these datasets serve as an excellent platform to test the hypothesis that a mix of real and synthetic data strengthens the generalization and robustness of machine learning models for image classification tasks. Finally, the diversity and scalability of the datasets ensured that various edge cases were analyzed, thus adding reliability to the experiment results.

Dataset Preparation

In our study, we have merged the three datasets that focus on speed bumps, potholes, and traffic lights. As mentioned earlier, each dataset was categorized into real and synthetic (AI-generated) subsets. The combined dataset was subsequently augmented to enhance visual diversity and improve model generalization. A variety of data augmentation techniques were systematically applied. Each image in the dataset underwent one augmentation operation, ensuring a uniform and consistent application of the augmentation process across all samples. Several data augmentation techniques were applied in our study, such as:

- Rotation: The images are randomly rotated at a specific angle to give different perspectives.

Fig. 1 Samples from the used dataset (real + AI-generated)

- Width and height shifts: Generate random shifts along with the image in the horizontal and vertical axes to simulate a small change in camera position.
- Zooming: Zoom to vary image sizes and bring out the differences in parts of the image.
- Shear: Shear Transformations were applied to capture the geometric distortion.
- Horizontal Flip: Flip the image horizontally, creating a mirrored version, thus improving the symmetry of the data.
- New Neighbor: The new nearest neighbor to fill the pixels produced during transformation, making the image visually coherent.

The final dataset was divided into five classes:

a. pothole.
b. speed bumps.
c. normal roads.
d. go (green) traffic lights.
e. stop (red) traffic lights.

Figure 2 shows examples of the original dataset and augmented images.

We added data to help us change the viewpoint of the car and motorcycle cameras. It will help us test our models better, as seen in Fig. 3. With the help of rotations, translations, shearing, zooming, and flipping, we modified the camera view angles according to scenarios that a vehicle's camera might experience in real situations. This method enables checking the system's robustness concerning varied camera and environmental perspectives. As a result, the augmented dataset provides a comprehensive representation of diverse camera view conditions, enabling thorough evaluation and enhancing the system's performance across different scenarios.

Training Configurations

Seven configurations were tested:

a. 1k real images per class.
b. 1k real + 1k synthetic images per class.
c. 2k synthetic images per class.
d. 2k real images per class.
e. 2k real + 2k synthetic images per class.
f. 3k real images per class.
g. 4k real + 4k synthetic images per class.

The validation data was consistently made up of real images only (1k per class) to evaluate generalization on real-world data.

Model Architecture

In our study, we evaluated multiple pre-trained deep learning models to identify the most effective architecture for classifying road-related images. Among the tested models—ResNet50, VGG16, MobileNetV2, and SVM using extracted features—DenseNet201 emerged as the most

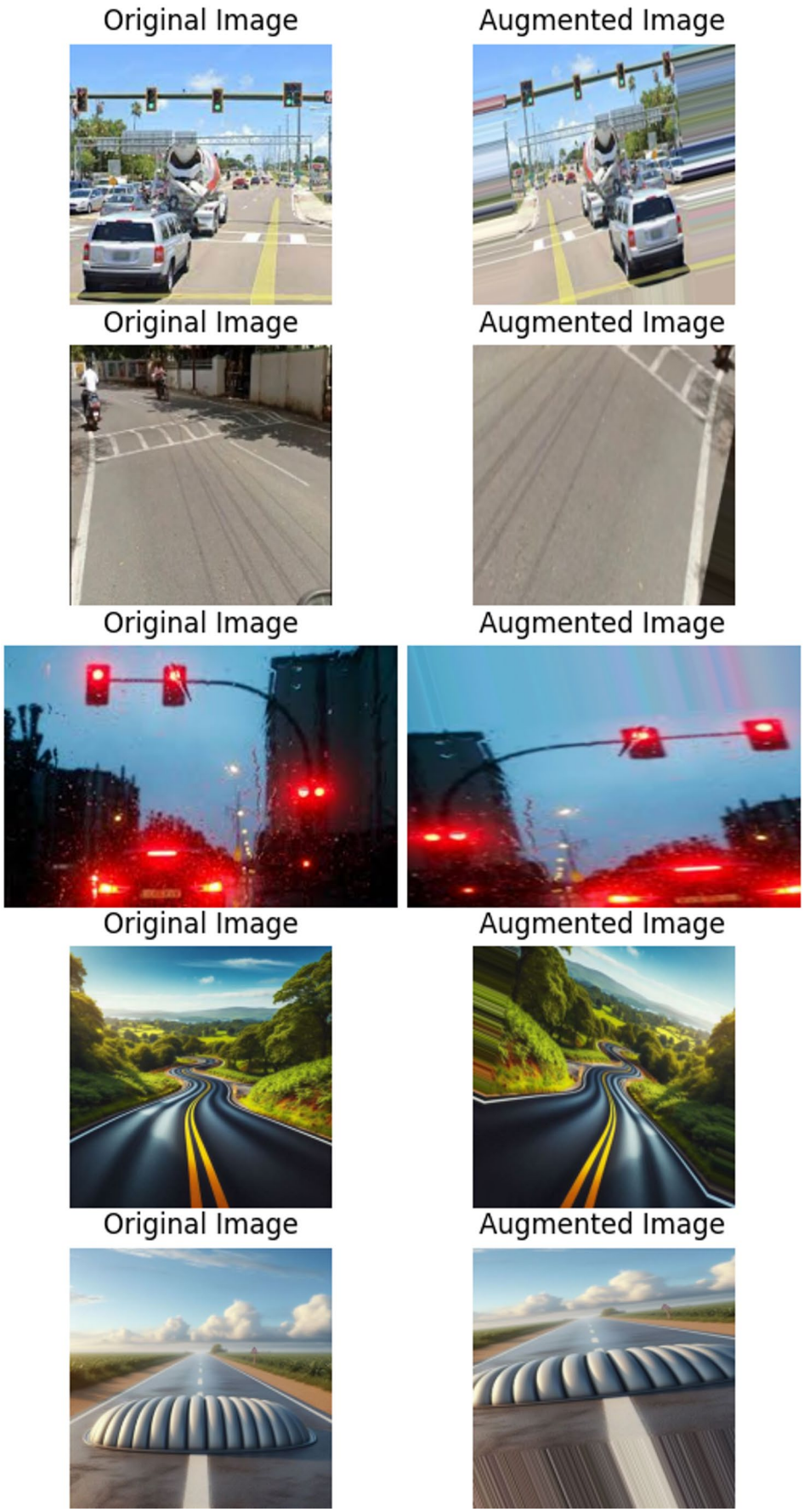

Fig. 2 Examples of the original dataset and augmented images

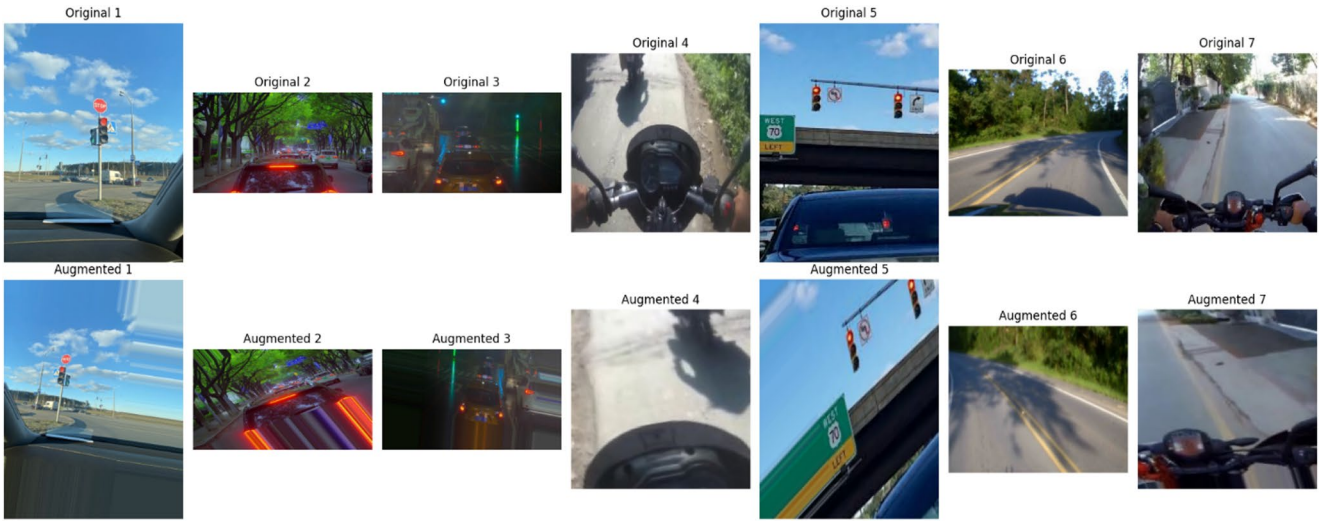

Fig. 3 Augmentation for vehicle camera view

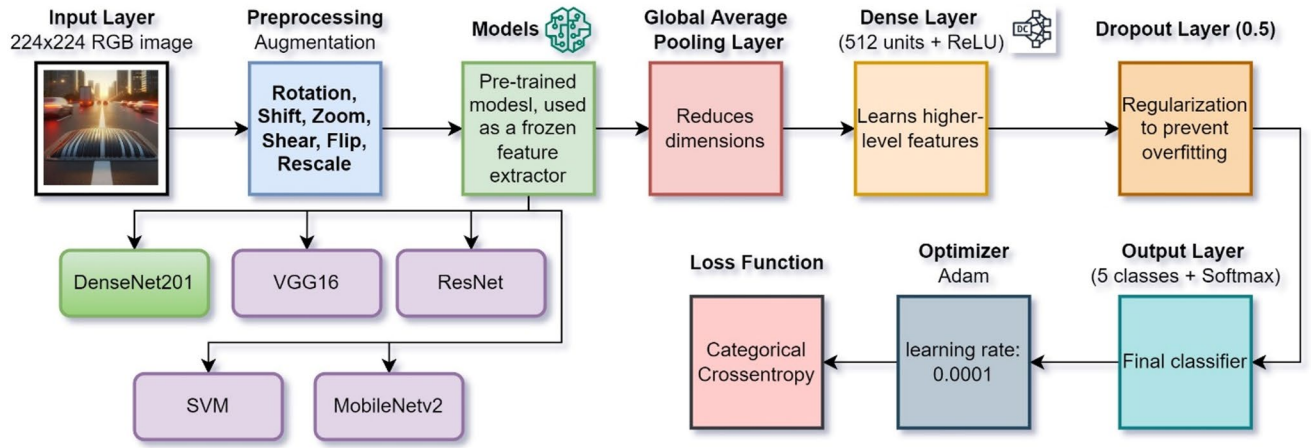

Fig. 4 Proposed approach's model architecture

accurate and robust. Pre-trained on ImageNet, DenseNet201 stands out for its efficient feature reuse through dense connections and its ability to mitigate vanishing gradient issues using dense blocks and transition layers. These characteristics make it particularly suitable for complex image recognition tasks (PyTorch, 2018). Figure 4 shows a block diagram of your model architecture. In our application, DenseNet201 demonstrated superior performance in detecting potholes, speed bumps, traffic lights, and other critical road features, confirming its suitability for traffic scene analysis and road safety applications [17]. Also, the following modifications have been applied:

- Global Average Pooling Layer: This was included to significantly reduce the spatial dimensions of the feature maps into a single feature vector for each feature map, thus minimizing the risk of overfitting.
- Fully Connected Layer: It includes a dense layer with 512 neurons and ReLU activation, enabling the network to learn high-level features.
- Output Layer: Lastly, the model includes a 5-class softmax output layer to classify images into five categories (e.g., potholes, speed bumps, normal roads, stop and go traffic lights).
- The model was trained using the Adam optimizer with a learning rate of 0.0001, which combines the benefits of momentum and adaptive learning rates.
- As the Loss Function 'categorical cross-entropy' has been used, which is appropriate for multi-class classification problems.

Accuracy, precision, recall, and F1-score were calculated to evaluate model performance. As mentioned, validation accuracy was consistently measured using 1k real images per class.

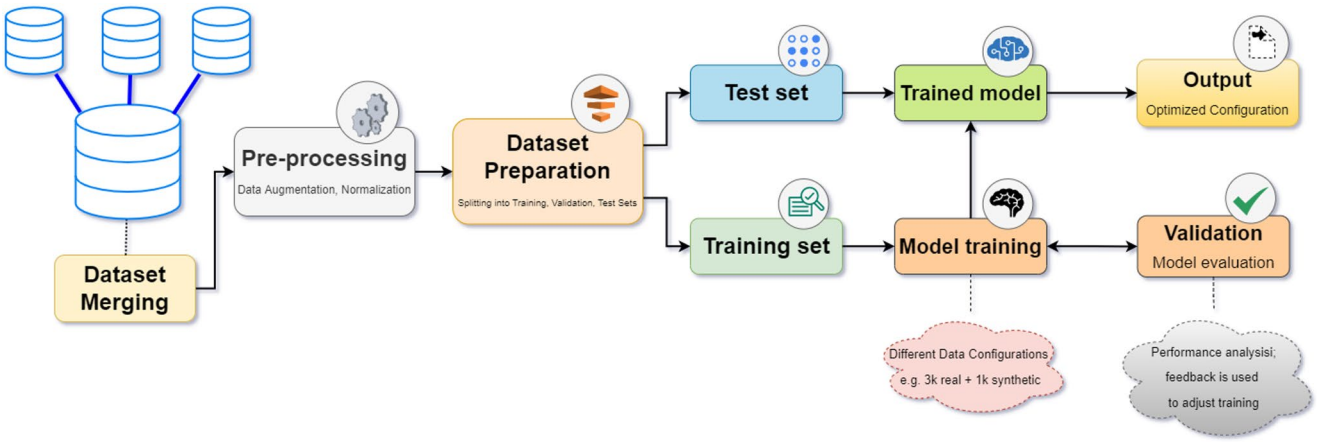

Fig. 5 The workflow of the proposed system

Table 2 Proposed system's training configurations

Config No	Real Images per Class	Synthetic Images per Class	Description
A	1,000	0	Real images only
B	1,000	1,000	Equal mix of real and synthetic
C	0	2,000	Synthetic images only
D	2,000	0	Increased real images only
E	2,000	2,000	Balanced set with higher volume
F	3,000	0	Larger real-only dataset
G	4,000	4,000	Maximum balanced real and synthetic

Table 3 The comparison of model performance across different configurations

Configuration	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
1k Real	97.38%	90.48%	0.0942	0.2849
1k Real + 1k AI	97.50%	90.48%	0.0831	0.2880
2k AI	96.77%	54.76%	0.1093	1.2743
2k Real	96.62%	94.29%	0.0939	0.1258
2k Real + 2k AI	95.35%	94.73%	0.1216	0.1179
3k Real	96.14%	97.36%	0.0987	0.0714
4k Real + 4k AI	79.05%	81.77%	0.3372	0.3047

Figure 5 demonstrates the workflow of the proposed system. The dataset was created by merging three datasets. After the pre-processing (e.g., data augmentation and normalization), the dataset is split into training, validation, and test sets. Then, models are trained according to seven different configurations. The trained models were then evaluated on a validation set and tested on a test set to assess the validity of the proposed system. The optimized configuration is then obtained after analyzing the performance.

Results and Discussion

Performance across Configurations

Table 2 shows the summarized training configurations as described earlier in Sect. 3.3.

Table 3 summarizes the model's performance across configurations. Validation accuracy improved with increasing proportions of real data, peaking at 97.36% for 3k real images. Combining 1k real and 1k synthetic images yielded slightly higher training accuracy (97.50%) compared to 1k real alone (97.38%), indicating that limited synthetic data can effectively complement real data.

Incidentally, the point was made in reference to the 97.50% training accuracy (1k real + 1k synthetic images) vs. the 97.36% validation accuracy (3k real images), likely not meant as an overall assessment of generalization. The objective was to focus on the effect of synthetic data on the model. So, the results imply that injecting synthetic data enhances the training procedure. Training with the combined real and synthetic data (1k + 1k) produced a better-trained accuracy (97.50%) than running on real data only (97.38%) of the model. The validation accuracy (97.36%) of a different configuration (3k real images) suggests using more real data to generalize better. However, over-reliance on synthetic data (e.g., 2k AI, 4k AI) resulted in significant performance degradation due to domain gaps.

Table 4 outlines the main training parameters and experimental environment used in this study. The model was trained for 20 epochs with a batch size of 32, using the Adam optimizer and categorical cross-entropy as the loss function. Although early stopping was not utilized, the best-performing model was saved based on validation accuracy to ensure optimal performance. All input images were resized to $224 \times 224 \times 3$, and data augmentation techniques were applied to improve model generalization. The

Table 4 Key training parameters and experimental setup

Parameter/Setting	Value/Description
Training Epochs	20
Batch Size	32
Early Stopping	Not used (training monitored via val_accuracy, with best model saved manually)
Optimizer	Adam (learning rate = 0.0001)
Loss Function	Categorical Cross-entropy
Data Augmentation	Rotation, Width/Height Shift, Shear, Zoom, Horizontal Flip, Fill Mode: Nearest
Image Input Size	224 × 224 × 3
Validation Strategy	1,000 real images per class (held-out set)
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, Confusion Matrix
Hardware Used	Intel Core i7 CPU, 16GB RAM, Windows 11
Software Stack	Python 3.11, TensorFlow 2.x, Keras, Matplotlib, NumPy, scikit-learn

experiments were conducted on a machine powered by an 11th Gen Intel® Core™ i7-1165G7 @ 2.80 GHz processor with 16 GB RAM.

Graphical Analysis

Figure 6 presents graphs from all configurations of the training and validation accuracies. The training accuracy remains high throughout all configurations with very little difference, whereas the validation accuracies have a more erratic performance. For example, configuration 2k-AI has a very high drop for validation accuracy, which reflects either that overfitting might be occurring or that learning or training issues may occur. However, there are configurations like 3k-Real that achieve the maximum validation accuracy, meaning that it is a balanced system in terms of training and generalization.

Figure 7 depicts the training and validation loss of the model having different configurations. It is evident that for 2k-AI, the validation loss shoots up dramatically, thus demonstrating its very poor generalization. In comparison, configurations like 3k-Real and 4k-Real + 1k-AI all show lower validation loss, which reflects the improvement of convergence and generalization because the training loss has consistently been very low across configurations, which in turn corroborates the high training accuracy reported here.

Results and Comparative Insights

Table 5 presents a performance comparison between the proposed DenseNet201-based model and several widely used deep learning architectures, including ResNet50, VGG16, MobileNetV2, and an SVM classifier trained on deep features. As shown, DenseNet201 achieves the highest training accuracy (97.50%) and validation accuracy (97.36%), along

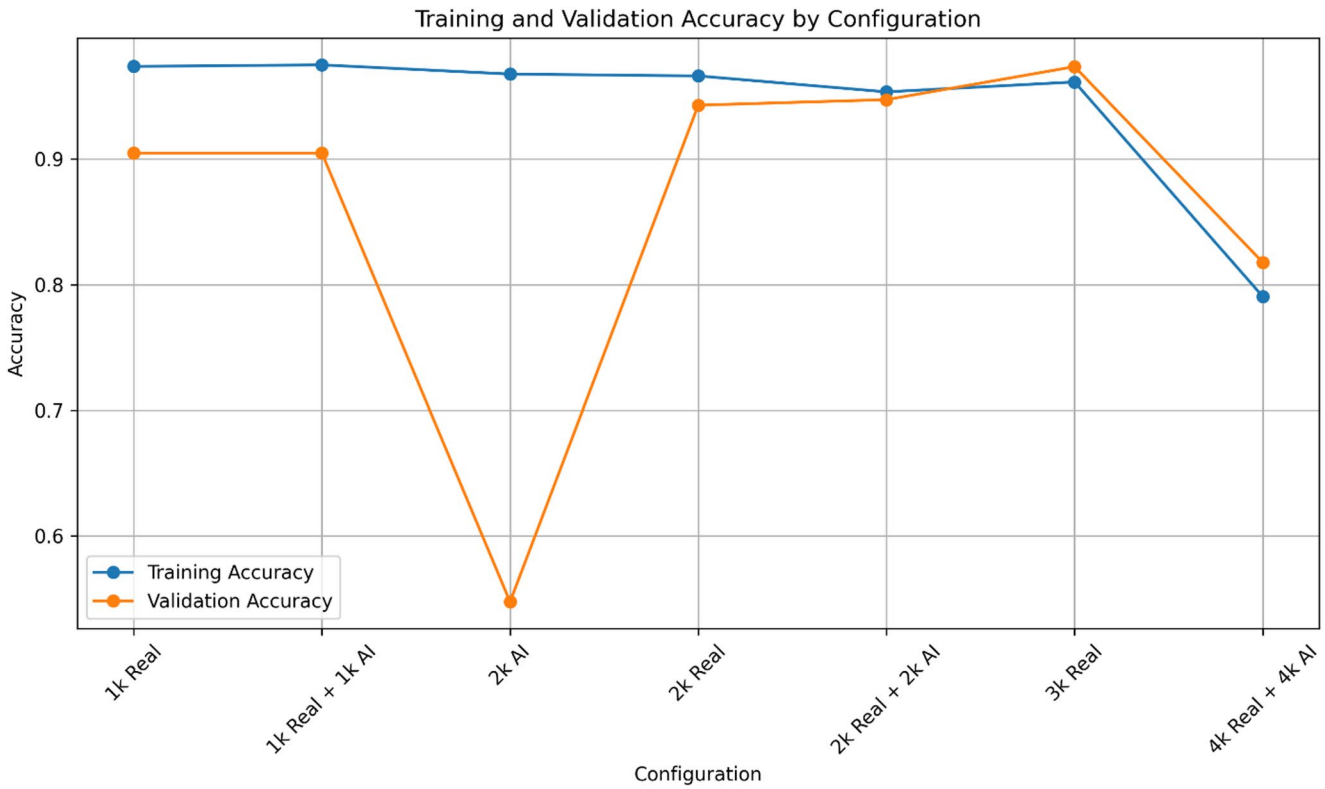

Fig. 6 Training and validation accuracy curves of the proposed scheme for several configurations

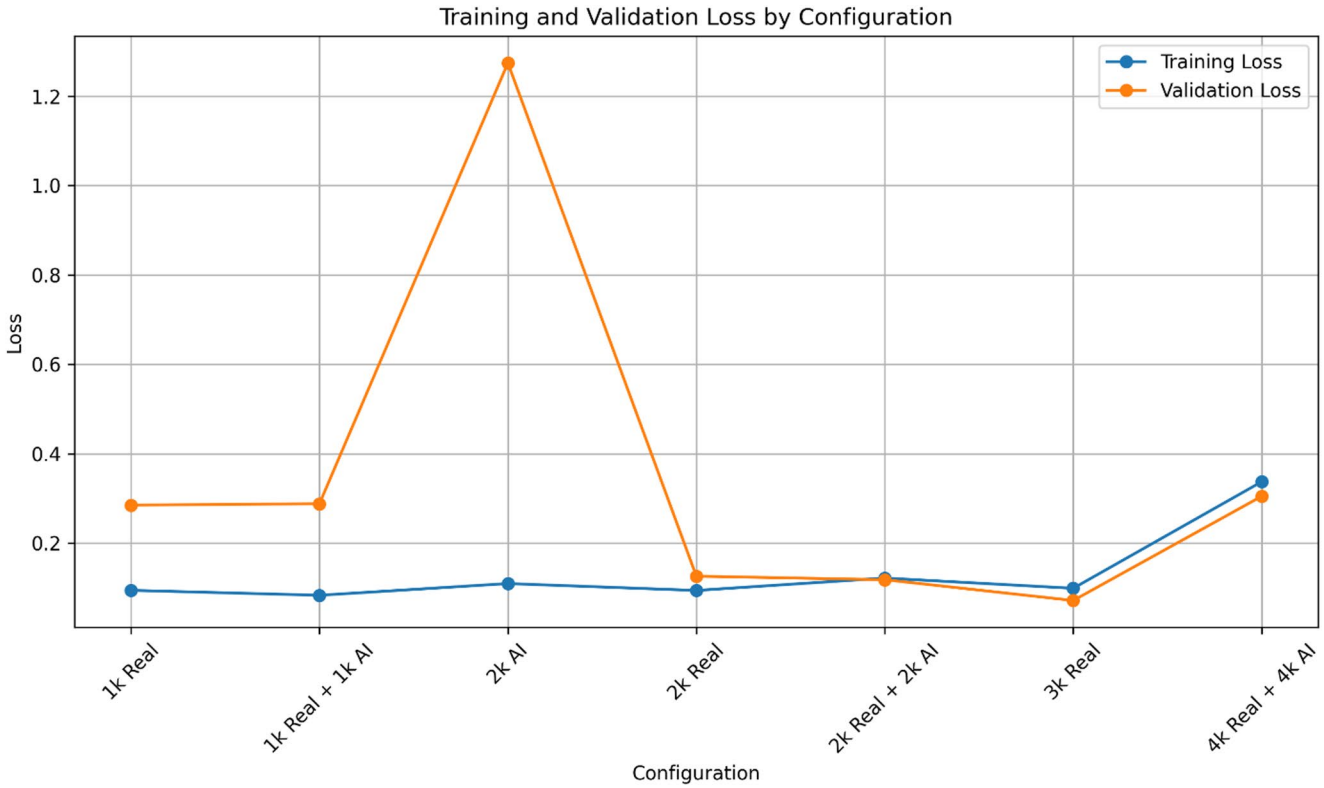

Fig. 7 Loss curve of the proposed model when trained and test with several configurations

Table 5 Performance comparison of different classification models

Model	Training Accuracy (%)	Validation Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DenseNet201	97.50	97.36	97.4	97.3	97.35
ResNet50	94.60	93.85	93.9	93.5	93.7
VGG16	93.40	91.20	91.5	91.0	91.25
MobileNetV2	92.80	90.70	90.8	90.6	90.7
SVM (on DenseNet features)	89.00	86.50	86.9	85.7	86.3

with superior precision, recall, and F1-score. This confirms its effectiveness in learning complex features and generalizing well across diverse traffic-related image classes. While other models perform reasonably well, DenseNet201 consistently outperforms them, validating its selection as the core architecture in our study.

A comparative analysis of our work is presented in Table 6, along with five of the most recent studies, and a thesis regarding datasets, models, and results. Studies, such as [11, 12, 13], only leveraged CIFAKE, a dataset of 120,000 images strictly with a 50/50 split between real and AI-generated, but have been limited in using smaller or dedicated synthetic datasets like [4] or [5]. However, our research incorporated three diverse datasets (speed bumps, potholes, traffic lights) and made an equal 4 K real and AI-generated split, thus allowing more extensive testing. In terms of models, the papers covered various architectures, from custom CNNs, GANs, and SVMs to those that included state-of-the-art models such as EfficientNet and DenseNet for all but one paper. The thesis involved several architectures of CNNs combined with dropout regularization. Uniquely among these, our work employed DenseNet201 pre-trained on ImageNet with Adam optimization, powerful data augmentation strategies, and real-to-synthetic ratio testing. The results of the study included the highest accuracy, which was 97.74% using DenseNet [12], while others reported between 84% and 93% among models of other papers. Our model achieved an impressive 97.36% validation accuracy using 3,000 real training images, along with a high training accuracy of 97.50%, which mirrored great performance at an extremely optimized 1:3 real-to-AI ratio. This speaks to the prowess of our dataset preparation, which, together with the supplement of model enhancements, brings forth state-of-the-art results.

Key Insights

a. **Balanced Integration**: A 1:3 imagery ratio turned out to be the best configuration in balancing elasticity and robustness in terms of the model.

Table 6 Comparative insights

Aspect	(Ruchira Purohit, 2024) [4]	(Zuhao Yang, 2023) [5]	(JORDAN J. BIRD, 2024) [11]	(Yuyang Wang, 2023) [12]	(Shivani Atul Bhinge, 2023) [13]	(Aml Yasser, 2024) [14–16]	Our Work
Datasets	Google Images (1 K images) and a small self-made dataset.	Fully and semi-synthetic datasets (e.g., DiffusionDB, HPD v2, ForgeryNet, DeepArt).	CIFAKE: 120,000 images (60k real from CIFAR-10, 60k AI-generated).	CIFAKE: 120,000 images (60k real from CIFAR-10, 60k AI-generated).	CIFAKE: 120,000 images (60k real, 60k AI-generated).	Three datasets focused on specific domains were tested individually.	Merged three datasets (speed bumps, potholes, and traffic lights) with 4 K real and 4 K AI-generated images.
Models	Custom CNN with four convolutional layers.	GANs, diffusion models for synthetic data generation.	CNN and Grad-CAM for XAI.	ResNet, DenseNet, VGGNet, and SVM.	EfficientNet-V2 B0 and a CNN. EfficientNet-V2 B0.	Various CNN architectures with dropout and regularization.	DenseNet201 with Adam optimizer, incorporating data augmentation and testing multiple real-to-synthetic ratios.
Results	Accuracy: 88% (Google dataset). Self-made dataset achieved 81%.	Synthetic data improves performance, robustness, and generalization, especially for object detection and segmentation tasks.	Achieved 92.98% accuracy using the CIFAKE dataset. Highlighted the importance of dataset diversity and balance.	DenseNet achieved the highest accuracy (97.74%) among tested models. SVM underperformed compared to deep learning models.	EfficientNet-V2 B0 demonstrated superior accuracy and F1 scores. AI-generated data performed well but lagged behind real data due to domain shifts.	Training exclusively on synthetic data yielded suboptimal results. The best model achieved 84% accuracy on real images.	Achieved the highest validation accuracy of 97.36% with 3k real images. A 1:3 ratio of real to AI-generated images showed strong performance (97.50% training accuracy).

b. **Data Augmentation**: These techniques greatly improved model generalization, addressing the limitations of the thesis, which relies on only real and raw datasets.
c. **Quality of Synthetic Data**: Despite their utility, synthetic images generate domain shifts within their confines, testifying the need for careful curation.

Discussion and Future Directions

The confusion matrix for the models trained on 3k real and 2k real + 2k AI-generated images is shown in Figs. 8 and 9. The two matrices show the performance of the models in classifying all the classes. The main errors in the 3k real data model happen in the go and stop classes. A few images that are go were classified as stop and vice versa. The second model trained on 2k real + 2k AI data also has false positives on go-stop classes. There are other false positives, too, like a speed bump instance being classified as a stop. The diagonal values were also high. Thus, the models are good but may require some fine-tuning.

Our research proves the usefulness of combining real and synthetic data in different image classification tasks, such as traffic and road safety scenarios. We studied how different data configurations affect model performance by utilizing real data from speed bumps, potholes, and traffic lights combined with AI-synthesized synthetic images. Seven training configurations were used on the DenseNet201 architecture with further data preprocessing techniques for diversity and robustness enhancement.

Results highlight that a hybrid dataset including real and synthetic images has a significantly higher advantage than a dataset using just one type of data alone. The experiments confirmed that the optimum real-to-synthetic-image ratio is 1:3 for improved generalization and classification accuracy. This further establishes the importance of synthetic data as a scalable, cost-effective means by which augmented datasets stand to benefit, particularly in cases of data scarcity and domain gaps.

The success of such an approach carries some weight of implication. The hybridization method with synthesized data offers practical application solutions for developing and sustaining performance in machine learning tasks from typical problems of overfitting and limited data. The effort had the added benefit of using tools like Gemini, Microsoft Copilot, and MidJourney to generate synthetic images, achieving better diversity and edge-case coverage for overall model robustness [18].

The research inspires performing further research applying hybrid datasets in other areas of computer vision, like object detection, image segmentation, and anomaly detection. Future studies may also include research into domain adaptation techniques and advanced generative AI models that would eventually narrow the performance gap in the coming years.

Fig. 8 Confusion matrix of 2k real + 2k AI-data model

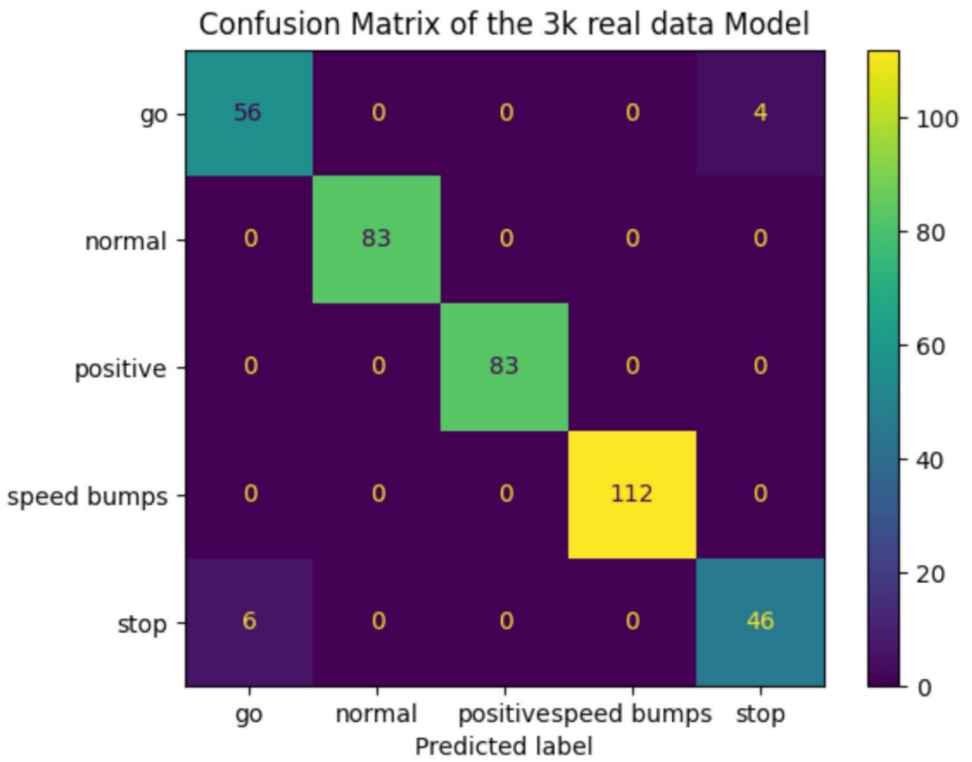

Fig. 9 Confusion matrix of 3k real data model

Conclusion

The study illustrates the efficacy of combining synthetic and real-world images to enhance image classification performance. Through experimentation with various ratios of real and AI-generated images across three traffic-related datasets (speed bumps, potholes, and traffic lights), we determined that a balanced hybrid dataset produces optimal performance. The results demonstrate that a real-to-synthetic ratio of 1:3 enhances generalization and accuracy, as reflected in the 97.36% validation accuracy obtained with 3,000 real images. The implementation of DenseNet201, combined with the Adam optimizer and data augmentation techniques, significantly improved model robustness. The findings indicate that although synthetic data cannot entirely substitute real-world data because of domain discrepancies, it acts as a valuable complement, particularly in situations where real data is scarce or expensive to acquire.

This study highlights the practical advantages of synthetic datasets in mitigating prevalent machine learning issues, including data scarcity and overfitting. Future research should investigate domain adaptation techniques and advanced generative AI models to reduce the performance disparity between synthetic and real-world data. Furthermore, applying this methodology to additional computer vision tasks, including object detection, segmentation, and anomaly detection, may yield further insights regarding the effectiveness of synthetic data in machine learning.

In summary, this research demonstrates the efficacy of hybrid datasets as a cost-efficient and scalable method for improving machine learning performance in image classification, facilitating wider implementation in AI-driven applications.

Acknowledgements This work received no funding.

Data Availability The dataset can be obtained via an email to the corresponding author.

Declarations

Conflict of Interest The authors declare no conflict of interest.

References

1. Ktena I, Wiles O, Albuquerque I, et al. Generative models improve fairness of medical classifiers under distribution shifts. Nat Med. 2024;30:1166–73. https://doi.org/10.1038/s41591-024-02838-6
2. Li G, Song Y, Zheng M. SAU: a dual-branch network to enhance. Long-tailed recognition via generative models; 2024.
3. He R, Sun S, Yu X et al. Is synthetic data from generative models ready for image recognition? 2022.
4. Purohit R, Sane Y, Vaishampayan D et al. AI vs. human vision: a comparative analysis for distinguishing AI-generated and natural images. In: 2024 4th International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2024. Institute of Electrical and Electronics Engineers Inc. 2024.
5. Yang Z, Zhan F, Liu K, et al. AI-generated images as data source. The dawn of synthetic era; 2023.
6. Aml Yasser. Studying the effect of generative image datasets on deep learning algorithms performance. German University; 2024.
7. Azizi S, Kornblith S, Saharia C, et al. Synthetic data from diffusion. Models improves ImageNet classification; 2023.
8. Alimisis P, Mademlis I, Radoglou-Grammatikis P, et al. Advances in diffusion models for image data augmentation: a review of methods, models. Evaluation metrics and future research directions; 2025.
9. Resmini N, Lomurno E, Sbrolli C, Matteucci M. Your image generator is your new private dataset. 2025.
10. Nguyen L-C, Nguyen-Tri Q, Khanh BT et al. Provably improving generalization of few-shot models with synthetic data. 2025.
11. Bird JJ, Lotfi A. CIFAKE: image classification and explainable identification of AI-generated synthetic images. IEEE Access. 2024;12:15642–50. https://doi.org/10.1109/ACCESS.2024.3356122
12. Wang Y, Hao Y, Cong AX. Harnessing machine learning for discerning. AI-Generated Synthetic Images; 2024.
13. Bhinge SA, Nagpal P. Quantifying the performance gap between real and AI-generated synthetic images in computer vision. 2023.
14. Aml Yasser. Generated speed bumps. 2024. In: https://www.kaggle.com/datasets/amlyasser/generated-speed-bumps
15. Aml Yasser. Generated street potholes. 2024. In: https://www.kaggle.com/datasets/amlyasser/generated-speed-bumps-imaged
16. Aml Yasser. Generated traffic lights. 2024. In: https://www.kaggle.com/datasets/amlyasser/generated-traffic-light
17. Charisma RA. Transfer learning with Densenet201 architecture model for potato leaf disease classification. In: 2023 International Conference on Computer Science, Information Technology and Engineering. 2023.
18. Zhao Y, Zhong Z, Zhao N et al. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. 2022.