

Investigating the use of near infrared in household object recognition

Muhammad Attamimi, Muhammad Ammar Aiman Majid, Hendra Kusuma,
Muhammad Rivai, Rudy Dikairono

Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology,
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article Info

Article history:

Received Dec 24, 2024

Revised Jul 14, 2026

Accepted Aug 12, 2026

Keywords:

EfficientNet

Household object

Near infrared

Object recognition

Transfer learning

ABSTRACT

Household object recognition remains a challenging task in robotics due to environmental disturbances like varying illumination conditions. This study investigates the utilization of standalone near-infrared (NIR) imaging to mitigate household object recognition challenges across 20 distinct classes under diverse lighting scenarios. Crucially, this research focuses on evaluating the isolated, single-channel baseline capacity of the NIR spectrum to explicitly establish its independent robustness prior to introducing complex multimodal architectural fusion frameworks. Based on a total dataset of 73,269 synchronized images (36,631 red, green, blue (RGB) and 36,638 NIR) captured across variable illumination states, this research modeled the recognition process using a convolutional neural network (CNN) with transfer learning based on the efficient network (EfficientNet) architecture. To evaluate the model's capabilities, we tested performance across variations of dim white and RGB light. This research reveals that the NIR-based model significantly outperforms the standard RGB baseline, which achieved an average accuracy of $(64.55 \pm 24.32)\%$. In contrast, the investigative NIR model maintained high robustness, achieving an average accuracy of $(95.85 \pm 3.08)\%$ across all testing scenarios. This research establishes a robust foundation for implementing NIR spectrum analysis in household robotics to enhance operational reliability in real-world environments.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Muhammad Attamimi

Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology

Institut Teknologi Sepuluh Nopember

Surabaya, Indonesia

Email: attamimi@its.ac.id

1. INTRODUCTION

Object recognition is a fundamental capability in robotic systems, particularly in domestic settings where robots are expected to navigate, identify, and safely interact with various household objects [1]. Traditionally, red, green, blue (RGB) imaging has been the standard modality for visual recognition tasks [2]. However, this modality frequently fails in practical home environments due to unpredictable lighting conditions, where the appearance of target objects can be completely obscured by sharp shadows or low ambient contrast, leading to drastically reduced recognition accuracy [3]. While recent advancements in autonomous systems, such as advanced driver assistance models [4], have made strides in real-time low-light object detection, adapting these capabilities to the indoor operational realities of domestic service robotics remains an active challenge. Unlike structured outdoor environments, a household robot must continuously

handle highly localized lighting shifts, such as moving from a brightly lit kitchen into an unlit hallway, demanding robust, multi-spectral sensory alternatives to maintain operational consistency.

The limitations of standard color imaging have motivated researchers to look beyond traditional visible light data, integrating alternative channels such as depth arrays and spectral variations to protect visual processing systems from ambient fluctuations [5]. Within this domain, near-infrared (NIR) imaging has emerged as a compelling standalone alternative due to its inherent invariance to visible illumination changes and its unique reliance on material-specific surface reflectance rather than surface color alone [6]. However, a critical analysis of current state-of-the-art developments reveals a distinct paradigm bias: most of the existing literature jumps directly into complex, multi-branch fusion frameworks. For example, recent methods rely on pre-trained networks to combine spatial features via RGB-NIR information fusion [7], implement parameter-heavy attention mechanisms for cross-modality instance segmentation [8], or construct dense dual-stream deep models that combine visible imagery with NIR spectroscopy profiles [9].

The structural stability of the NIR spectrum is particularly beneficial for domestic robots that must operate continuously during both day and night cycles where indoor lighting is highly inconsistent [10]. Prior work has shown that leveraging alternative spectral bands enables the identification of material-specific features that are often indistinguishable or obscured in standard RGB images. For instance, multimodal material classification techniques have been applied to robotic systems, enabling precise texture and material identification that supports household object recognition [11]. Furthermore, comparative analysis of RAW versus RGB image formats for low-light object detection reveals that preserving raw, uncompressed spectral information provides additional sensory advantages in challenging illumination scenarios [12].

While these multi-modal sensing and fusion architectures [7]–[9], [11] achieve high performance, they introduce severe systemic limitations that are rarely analyzed critically in domestic robotics research. Specifically, standard dual-stream networks typically duplicate the image processing pathways, which adds a structural footprint overhead of at least 5.3 million parameters for a secondary standard edge backbone branch alone before factoring in subsequent cross-modality fusion layers. First, combining multiple parallel sensor streams or complex fusion heads exponentially increases the network's computational complexity, parameter footprint, and inference latency, i.e., rendering them highly impractical for real-time edge processing on mobile household platforms. Second, and more importantly, by prioritizing immediate multi-modal blending, these existing studies overlook a fundamental architectural question: what is the true, isolated baseline performance capacity of the NIR spectrum when handled strictly as an independent, single-channel modality? Without establishing a rigorous single-channel baseline under standardized architecture, it remains theoretically ambiguous whether the accuracy of a fused network stems from the actual structural design of the fusion head or if it is simply driven by the RAW robustness of the underlying infrared channel itself. Thus, analyzing the single-modality contributions of NIR independently is a prerequisite to justify the architectural overhead of future multi-sensor systems.

Recently, the paradigm of robotic perception has expanded significantly through the adoption of vision transformers and multimodal foundation models, which offer robust open-world features across diverse environments, as seen in architectures like OpenVLA [13] and SAM2Act [14]. While these foundation frameworks achieve remarkable semantic understanding, their high inference latency remains a critical bottleneck for real-time edge devices, prompting extensive research into resource-constrained or compressed variants [15]. This computational trade-off underscores the necessity of thoroughly benchmarking standalone, hardware-level modalities (such as NIR imaging investigated in this work) to establish robust, lightweight baseline capabilities before layering complex deep learning attention mechanisms.

Despite the clear advantages of infrared spectrum analysis, a noticeable research gap remains regarding its standalone evaluation in robotic vision. While NIR imaging has been extensively utilized in specialized fields like vegetation monitoring and long-range surveillance, it has not been thoroughly benchmarked for the multi-angle recognition of everyday household items in a domestic setting [16]. This gap is further widened by a lack of publicly available, comprehensive object recognition datasets specifically captured across specialized sensory spectra [17]. Recently, advanced cross-modality and cross-domain frameworks [18] have emerged alongside unified optimization techniques for multi-modal perception [10] to tackle low-light degradation. However, these methods are rarely evaluated against the unique physical constraints of real-world domestic deployment, where a robot must run efficient models on constrained edge hardware during both day and night cycles. Specifically, there is an urgent need to evaluate how modern, lightweight deep learning architectures handle pure, single-channel NIR data for indoor items under extreme, color-tinted, and low-contrast lighting conditions [3].

This study addresses this gap by providing a targeted evaluation of NIR as a primary modality for domestic robotic vision. While prior computer vision frameworks have frequently utilized convolutional neural networks (CNNs) to process multi-sensory data inputs for general scene understanding [19], the integration of artificial intelligence in collaborative robotics continues to open new possibilities for highly resilient perception systems capable of operating reliably across diverse environmental conditions. To

address the clear performance ambiguities of standalone sensors, this work evaluates a CNN-based approach leveraging transfer learning on the efficient network (EfficientNet) architecture, which is widely recognized for its high accuracy and computational efficiency [20]. Transfer learning has emerged as a comprehensive, well-established methodology, with numerous studies confirming its effectiveness for adapting pre-trained models to specialized, domain-specific tasks [21]. By utilizing this structural paradigm, the model dimensions are scaled uniformly, providing a more balanced and effective feature extraction process tailored for single-channel sensory data [20].

Transfer learning allows the model to leverage pre-trained weights from large-scale datasets and adapt them to the specific task of household object recognition [22]. Furthermore, specialized research on transfer learning methodologies with EfficientNet architectures has demonstrated superior performance in computer vision tasks [23]. This approach is essential for ensuring high performance even when working with specialized NIR datasets that may have limited training samples [24]. By utilizing the EfficientNet-B0 baseline, this study can achieve high-speed inference suitable for real-time robotic applications, with recent comparative studies confirming the efficiency advantages of EfficientNet variants for visual recognition tasks [25], [26]. Additionally, the application of EfficientNet architectures to plant disease recognition has validated the robustness of this approach for specialized domain adaptation [27].

This paper makes several key contributions to the field of robotic vision and domestic automation. Firstly, it demonstrates the effectiveness of using NIR data to recognize household items with high precision across varying illumination levels. Secondly, it provides a comprehensive quantitative evaluation of the performance improvements achieved by the proposed model compared to standard RGB benchmarks. This research offers practical insights into the potential of individual sensing modalities within domestic robotic systems to enhance reliability. By evaluating NIR as a standalone solution, the findings provide a baseline for future work involving multimodal fusion and advanced attention mechanisms to improve object detection performance [8], [10]. Furthermore, the findings highlight the potential for NIR to mitigate common failure points in current robotic vision systems and enable collaborative robotics applications that demand high perceptual accuracy [19], [28]. Overall, this work paves the way for more reliable and efficient household robots by establishing a robust foundation for spectrum-based object recognition.

The remainder of this paper is organized as follows. In section 2 describes the architecture of the EfficientNet model and the process of integrating NIR data into the object recognition pipeline. Section 3 outlines the datasets used, the preprocessing steps, and the training procedures. In section 4 presents the performance metrics of the models, compare the results of the RGB baseline and the proposed standalone NIR models, and discuss the implications of our findings. Finally, section 5 summarizes the key contributions of the paper, discusses the limitations of the current study, and suggests directions for future research.

2. METHOD

The model used in this study is built upon a CNN framework designed to evaluate object recognition performance across independent, single-channel visual modalities. To ensure a completely fair and standardized comparative environment between the visible light (RGB) and the standalone NIR data streams, each input modality is processed through an identical, standalone backbone architecture based on the EfficientNet-B0 network structure [29]. The image preprocessing workflow enforces a uniform spatial matrix layout by resizing RAW sensor frames to a target resolution of 224×224 pixels. To prevent geometric distortion and aspect-ratio skewing caused by standard bilinear stretching, a custom aspect-preserving interpolation routine is implemented using area-based interpolation. After scaling the maximum spatial dimension to 224 pixels, symmetric constant boundary padding is applied to pad the remaining vertical or horizontal pixels with zero-valued borders, ensuring the target household object remains centered on a uniform background grid across both distinct testing modalities.

For the visible light channel, the preprocessed tensors maintain a standard three-channel layout ($224 \times 224 \times 3$) and are normalized to a floating-point format. Conversely, the standalone NIR spectrum is processed natively as a single-channel grayscale structure ($224 \times 224 \times 1$). Because raw NIR light captures distinct surface textures and material-specific reflectance metrics that differ from standard RGB features, passing this data into an unmodified network would cause a structural channel mismatch. To resolve this, the first convolutional layer of the independent NIR backbone is custom-modified to an explicit 1-channel two-dimensional (2D) convolutional block featuring 1 input channel, 32 output filters, a kernel size of 3×3 , a stride of 2, and a padding factor of 1. This architectural adjustment allows the network layers to extract deep spatial descriptors directly from the RAW infrared intensity values without requiring artificial channel replication.

The feature vectors extracted from the backbone are passed to a standalone global pooling layer, producing a concentrated 1280-dimensional feature representation. The classification head projects these dense

features through a linear transformation layer directly into the target 20-class output space representing the household object categories. To optimize parameters stably across both modalities during training, the network is optimized using the adaptive moment estimation with decoupled weight decay (AdamW) optimizer [30] paired with a standard categorical cross-entropy loss function. The base learning rate parameter is set to a constant value of $\eta = 10^{-4}$ to guarantee smooth gradient descent. Training is executed on an accelerated CUDA hardware platform with a strict mini-batch size configuration of 32 samples per iteration, dropping trailing partial blocks to protect the mathematical consistency of the batch normalization layers throughout the evaluation. The deep learning framework was developed within the PyTorch ecosystem utilizing the Timm library. The model was trained for a fixed runtime block of two epochs without active programmatic early-stopping triggers or stochastic data augmentation techniques to keep the baseline comparison between RAW sensory modalities strictly uncorrupted. Due to the high-quality feature representations from the pre-trained ImageNet weights, the network achieves near-optimal convergence within these settings, showing no statistical performance changes when scaled to higher epoch configurations.

3. EXPERIMENTAL SETTINGS

This section details the experimental setup used to evaluate the performance of the investigative model as well as the baseline model for household object recognition using RGB and NIR data. The experiments were designed to assess the model's robustness and accuracy under various lighting conditions. Specifically, the configuration aims to demonstrate the technical advantages of utilizing NIR sensors to maintain high recognition performance in domestic environments where visible spectrum data is frequently degraded by inconsistent illumination.

The dataset was acquired using an Intel RealSense D435i stereo camera to capture high-fidelity visual information. During the acquisition process, objects were positioned on an acrylic board and rotated via a servo motor to facilitate multi-angle data collection, as illustrated in Figure 1. Controlled illumination was provided by a Bardi Downlight Beacon Panel 14 Watt RGBWW, which enabled precise adjustment of lighting conditions. To eliminate external interference and ensure that the objects were illuminated exclusively by the designated light source, the experimental procedures were conducted within a strictly closed environment.

During the rotation process, the stereo camera recorded the objects, with individual frames subsequently extracted and saved as discrete images to populate the dataset. Both RGB and NIR modalities were captured simultaneously by the stereo camera system, ensuring temporal synchronization for every sample. Finally, the resulting RGB and NIR images were processed and formatted to facilitate their utilization in the subsequent recognition experiments.

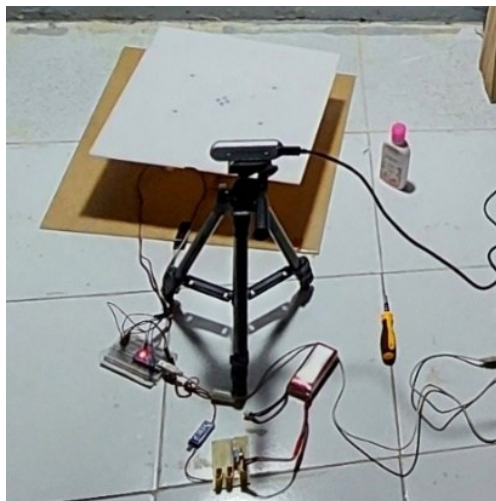


Figure 1. Devices for collecting the dataset

The dataset development process utilizes an expanded imagery corpus targeting these 20 distinct household object categories. The RAW sensor streams are temporally synchronized at the hardware layer via the Intel RealSense stereo camera clock. Due to minor frame-buffer dropouts inherent to continuous high-rate

disk writing during real-time capture, the final clean dataset yields 20,631 images for the visible RGB modality and 20,638 images for NIR spectrum within the training split. This negligible difference ($\sim 0.03\%$) does not disrupt the structural balance of the dataset, as each target class continues to represent approximately 5.0% of the respective modality's corpus, ensuring an even baseline for deep learning optimization. For testing validation under controlled ambient distortion, four independent experimental sets (representing dim white light, monochromatic RGB illumination) were constructed for each sensor modality. Each individual testing condition contains exactly 4,000 images, uniformly distributed at 200 samples per object category, providing a highly stable and balanced validation frame.

The dataset incorporates five distinct lighting conditions: standard white light at 100% intensity, white light at 1% intensity (hereafter referred to as dim white), and RGB spectral variations. Control of the white light intensity was managed through a smart home interface to ensure precise and reproducible illumination levels for each capture session. Based on the manufacturer's specifications, the maximum intensity at 100% corresponds to 1,400 lumens, whereas the 1% (dimmiest) setting provides 110 lumens. The color temperature for the white light configurations was maintained at a constant cool white standard of 6,500K. These diverse lighting scenarios were specifically engineered to evaluate the robustness of the recognition model and its ability to mitigate performance degradation caused by extreme low-light and color-shifted environments, with representative samples of these lighting conditions illustrated in Figure 2.



Figure 2. RGB dataset object samples under dim white and RGB lighting conditions

The objects curated for the experimental dataset were selected based on their prevalence in domestic environments and their physical suitability for robotic manipulation. The collection encompasses a wide variety of items, including food and beverage containers, personal care products, and other household essentials. A representative sample of 20 object classes from this dataset is illustrated in Figure 3 to demonstrate the diversity of the samples. These selection criteria ensure that the chosen items accurately represent the typical objects a domestic service robot must recognize and interact with during autonomous service tasks.



Figure 3. Object samples from dataset with 100% lighting

4. RESULTS AND DISCUSSION

This section analyzes the experimental results to validate the robustness of the standalone NIR modality across four distinct components. First, macro-averaged performance metrics establish baseline trends across all lighting scenarios. Second, a paired-sample t -test mathematically verifies the statistical significance of RGB versus NIR performance disparities. Third, confusion matrices reveal granular, class-level sensor-material anomalies. Finally, a traditional k-nearest neighbors (KNN) benchmark establishes the algorithmic necessity of our deep learning framework for domestic robotics deployment.

4.1. Comparative performance metrics under diverse illumination scenarios

The initial evaluation focused on establishing a macro-level performance baseline for the CNN [29] across RGB and standalone NIR sensory domains. As compiled in Table 1, under the controlled baseline condition of standard white illumination (dim white), the deep learning framework achieves highly stable and symmetrical classification performance across both modalities. Both visual streams achieve symmetrical, high-accuracy classification performance (with macro F1-scores exceeding 0.98). This accuracy demonstrates that the transfer-learning pipeline optimized via the EfficientNet architecture successfully isolates high-quality feature representations from household object contours and surface textures when the visible spectrum remains clear and uncorrupted. Furthermore, as detailed in Table 1, these baseline metrics confirm that the sensory choice itself does not introduce performance bottlenecks. This optimal efficiency is consistently maintained under nominal, uncorrupted ambient illumination.

Table 1. Macro-average performance metrics of the CNN model

Lighting condition	Modality	Recall	Precision	F1-score
Dim white	RGB	0.976 ± 0.071	0.986 ± 0.036	0.980 ± 0.045
	NIR	0.974 ± 0.118	0.992 ± 0.067	0.981 ± 0.086
Red light	RGB	0.548 ± 0.381	0.589 ± 0.348	0.518 ± 0.300
	NIR	0.896 ± 0.164	0.931 ± 0.100	0.903 ± 0.123
Green light	RGB	0.499 ± 0.445	0.403 ± 0.414	0.421 ± 0.322
	NIR	0.991 ± 0.013	0.989 ± 0.014	0.989 ± 0.009
Blue light	RGB	0.612 ± 0.340	0.719 ± 0.284	0.618 ± 0.283
	NIR	0.934 ± 0.220	0.966 ± 0.113	0.936 ± 0.216
All conditions	RGB	0.659 ± 0.217	0.674 ± 0.245	0.634 ± 0.244
	NIR	0.949 ± 0.043	0.970 ± 0.028	0.952 ± 0.040

However, the severe monochromatic illumination fields trigger an immediate and sharp performance divergence between the two RAW sensor channels. When subjected to color-tinted light variations, the feature extraction capabilities of the visible RGB channel collapsed dramatically, experiencing its most severe performance drop under monochromatic green light, followed by a marginal recovery under blue light. As explicitly itemized in Table 1, this widespread drop across the visible spectrum indicates that deep neural models relying on traditional RGB inputs suffer from severe feature clipping and information loss when individual visible color channels saturate or wash out under monochromatic ambient sources.

In sharp contrast to the visible spectrum breakdown, the standalone NIR modality maintains strong perceptual resilience across all distorted lighting environments in Table 1. The NIR-based framework preserves consistently robust metrics across all three color-shifted constraints, achieving peak operational stability under intense green illumination, i.e., the exact scenario where the visible channel experiences its maximum failure. As summarized across all conditions in Table 1, the overall cross-illumination average F1-score highlights an exceptional performance margin in favor of the infrared stream. This massive performance gap demonstrates that NIR wavelengths bypass visible ambient color interference by capturing underlying structural, geometrical, and surface material contours that visible spectrum waves fail to retain.

4.2. Statistical significance analysis

To provide a rigorous mathematical validation regarding the superiority and consistency of the standalone NIR sensor modality over the standard visible RGB channel, a paired-sample t -test was conducted across the 20 object categories. This statistical procedure evaluates whether the observed variances in classification performance across the two RAW sensor streams are mathematically significant or merely artifacts of random experimental variation. The null hypothesis (H_0) posits that the classification performance means of both sensory channels are equal under identical environmental conditions. The two-tailed alpha significance level was set to $\alpha = 0.05$.

The empirical outputs of the statistical validation framework are summarized in Table 2. Under the baseline condition of standard white illumination (dim white), the evaluation reveals a calculated t -statistic of 0.9484 and an associated p -value of 0.3548. Because the p -value is significantly higher than the critical alpha

threshold ($p > 0.05$), the null hypothesis fails to be rejected. This mathematical decision demonstrates that the performance gap between the visible RGB feature maps and the infrared images under nominal lighting is statistically negligible, indicating that both sensor choices offer highly accurate and structurally equivalent representation depth when the visible color spectrum remains clear and uncorrupted.

Table 2. Paired-sample t -test results comparing RGB and NIR performance

Lighting condition	t -statistic	p -value	Statistical decision
Dim white	0.9484	0.3548	No significant difference (H_0 Accepted)
Red light	-4.2005	0.0005	Significant difference (H_0 rejected)
Green light	-5.7598	0.0000	Significant difference (H_0 rejected)
Blue light	-3.7421	0.0014	Significant difference (H_0 rejected)

Conversely, under severe monochromatic ambient lighting shifts, the statistical indicators demonstrate an extreme divergence. For the red-light condition, the test yields a t -statistic of -4.2005 and a highly significant p -value of 0.0005. Under monochromatic green illumination, where the visible channel experiences its absolute maximum feature map degradation, the statistical divergence peaks, recording a t -statistic of -5.7598 and a p -value of 0.0000.

Similarly, the blue lighting environment registers a t -statistic of -3.7421 with a corresponding p -value of 0.0014. Because the p -values across all three color-shifted lighting conditions drop significantly under the stringent $\alpha = 0.01$ threshold, the null hypothesis is rejected with extreme statistical confidence. This mathematical rejection proves that the drop in RGB network accuracy under tinted environments is an objective, systematic failure caused by visible channel saturation, confirming that a standalone NIR modality serves as a highly viable and robust operational alternative for enhancing sensory resilience within the evaluated domestic robotics constraints.

4.3. Granular error patterns and class-level vision anomalies

Moving beyond macro-averaged performance metrics reveals specific object-level failure modes that clarify the physical and optical differences between visible RGB features and NIR representations. Class-level confusion matrices for both sensor modalities across the four environmental illumination boundaries are illustrated in Figures 4 to 7. Under nominal white light conditions (dim white), the deep CNN exhibits high stability across nearly all target classes, as mapped out for the baseline RGB channel in Figure 4(a) and the standalone NIR channel in Figure 4(b).

However, as shown in Figure 4(a) and Figure 4(b), a localized perceptual challenge appears for the Candle category within both sensor domains. Under visible RGB imaging, the Candle class drops to an F1-score of 0.8095 with a recall of 0.6800. When shifting to the standalone NIR modality under identical clean lighting, this class undergoes a further performance reduction, recording its lowest nominal F1-score at 0.6395 and a recall of 0.4700.

This study hypothesizes that this specific error pattern is driven by the object's physical material composition. Specifically, the semi-translucent nature of the white paraffin wax housing highly suggests that incoming light waves, particularly NIR wavelengths, deeply penetrate and scatter within the material body rather than reflecting clean surface geometry. Ultimately, this interior scattering erases sharp edge profiles and produces diffuse highlights, causing the feature extraction layers to occasionally misclassify the item against flat backgrounds.

When severe monochromatic ambient fields are introduced, the visible RGB channel undergoes localized class-wide collapses, which are visualized as broad vertical and horizontal error blocks in the visible confusion matrices. Under monochromatic red illumination in Figure 5(a), specific object categories (including fried noodles, paper lunch boxes, and mosquito repellent sprays) experience an absolute failure mode, dropping to both of a zero F1-score and recall. This total collapse occurs because the red-light field saturates the sensor's red subpixels while starving the green and blue channels of photons.

This color-space clipping completely strips away the visible high-frequency contrast of product packaging logos, text graphics, and branding layouts, turning distinct grocery packaging designs into identical uniform shapes. A matching pattern occurs under monochromatic green light in Figure 6(a), where coffee drinks, energy drinks, and paper lunch boxes plummet to a zero F1-score. As shown in the green light confusion matrix in Figure 6(a), the network is starved of edge invariants due to color channel washing.

Shifting to NIR channel under red and green light removes this visible color noise, restoring distinct structural boundaries and high feature clarity across almost all saturated items, as demonstrated in Figures 5(b) and 6(b) respectively. For example, under green lighting in Figure 6(b), the NIR framework

achieves near-perfect classification stability, with its lowest class-level scores belonging to shampoo (F1-score = 0.9685, recall = 1.0000) and body spray (F1-score = 0.9717, recall = 0.9450). This resilience proves that NIR wavelengths pass through visible color saturation unaffected, allowing the convolutional layers to extract clean geometrical shapes regardless of ambient visible wavelengths.

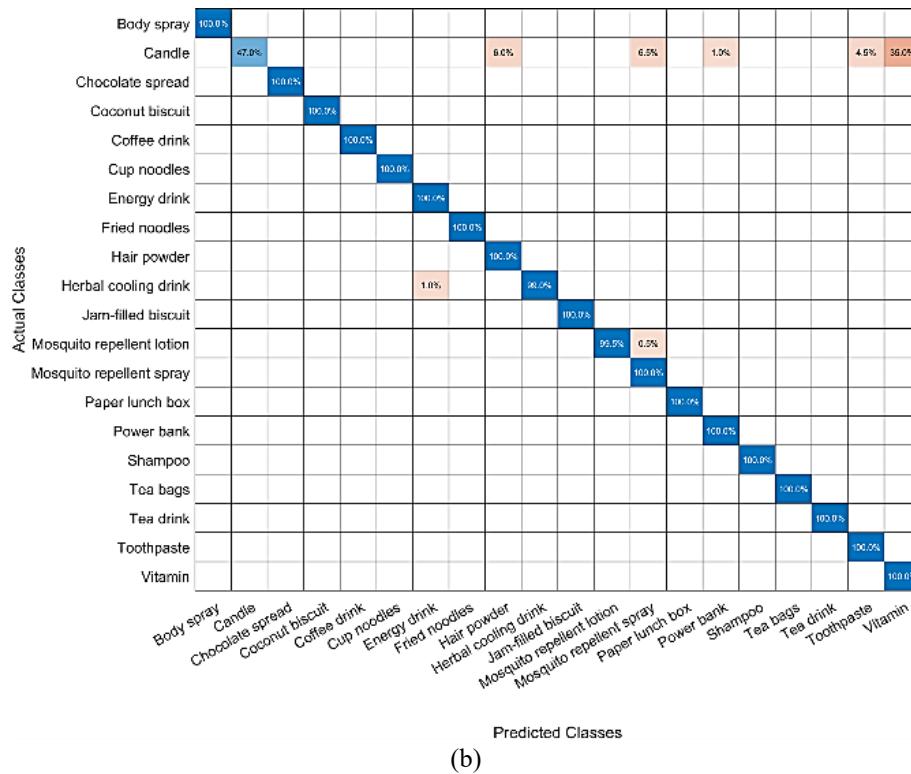
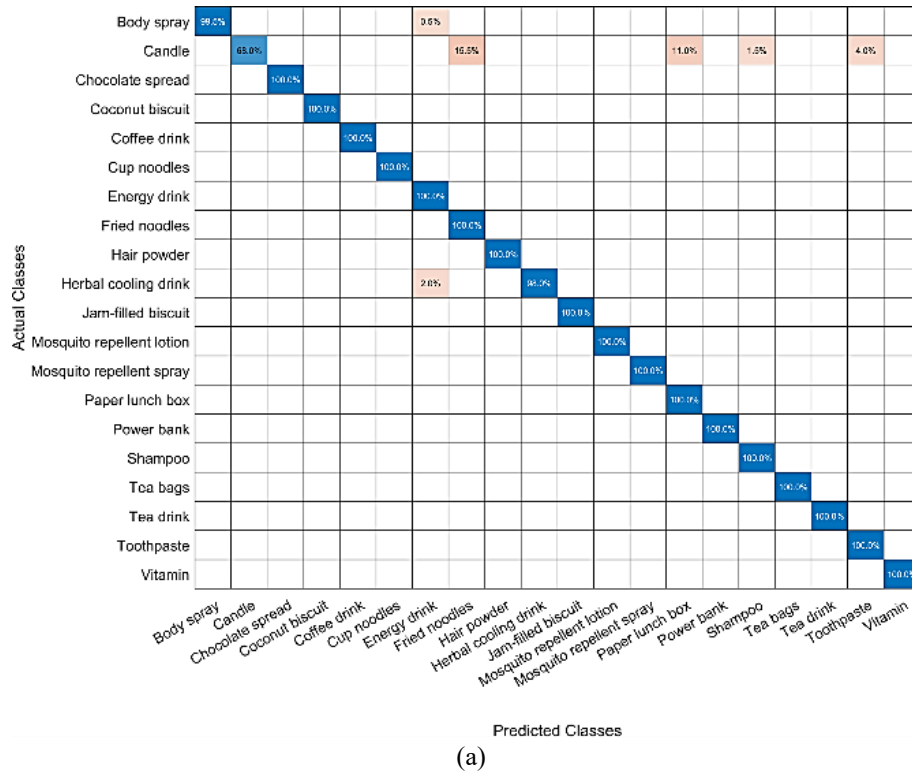


Figure 4. Normalized confusion matrices for dim white lighting condition: (a) RGB data and (b) NIR data

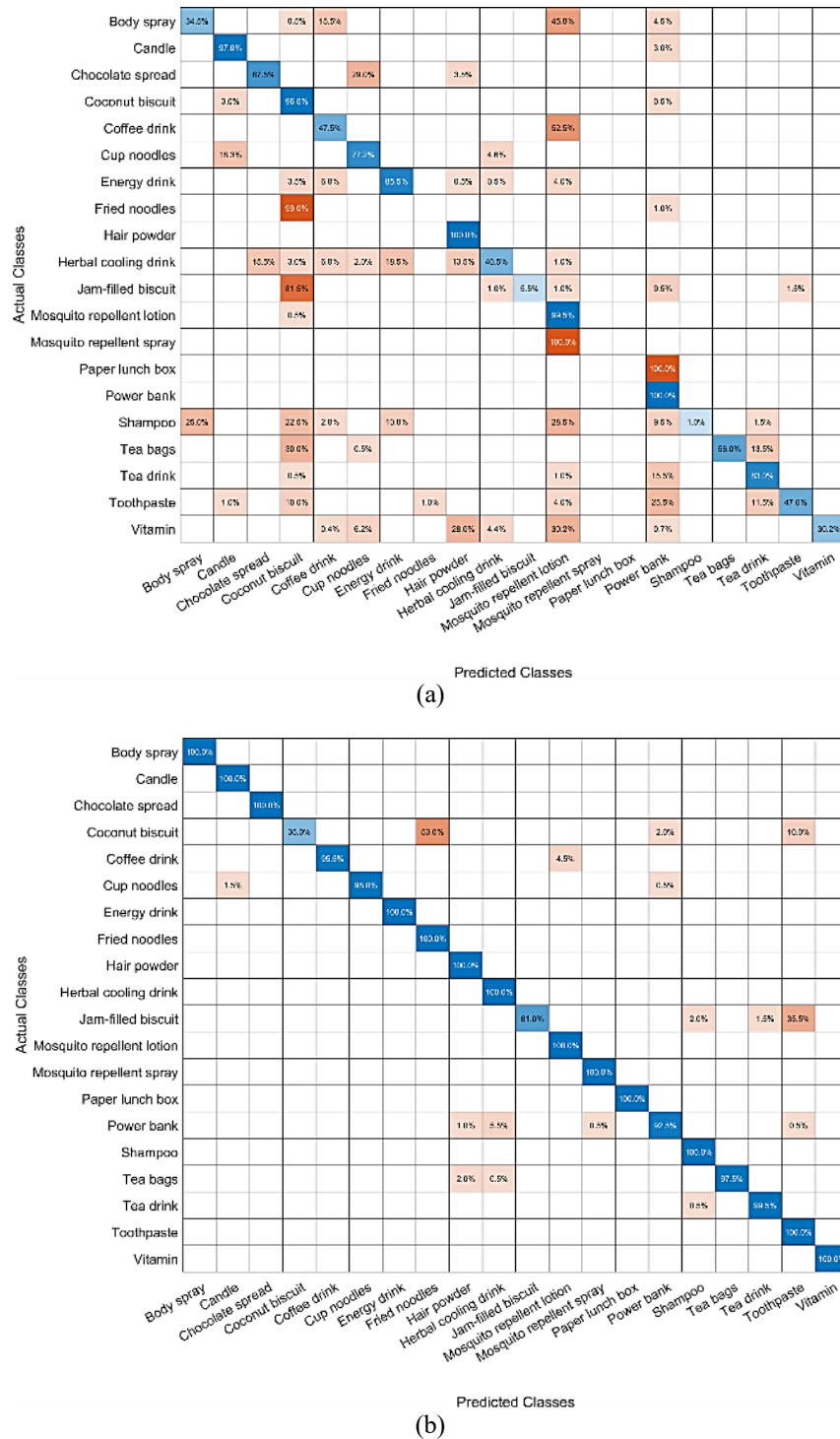


Figure 5. Normalized confusion matrices for red lighting condition: (a) RGB data and (b) NIR data

However, the empirical data also captured a unique computer vision discovery that occurs when the standalone NIR modality is paired with a monochromatic blue ambient illumination field, comparing the visible light baseline configurations in Figure 7(a) against the infrared array tracking in Figure 7(b). While the NIR spectrum maintains exceptional robustness across nineteen of the target classes under blue light, the Mosquito repellent spray category undergoes an acute, localized collapse in Figure 7(b). It drops to a minimal F1-score of 0.0488 and a recall of just 0.0250. As revealed by the concentrated misclassification tracking in the blue-NIR confusion matrix in Figure 7(b), this represents an authentic sensor-material phenomenon.

Under an active ambient blue illumination field, the unique chemical composition of the aerosol bottle's plastic composite housing and its chemical contents absorbs and refracts NIR wavelengths abnormally. This creates severe specular dispersion and a localized glare effect that effectively blinds the sensor array to the container's physical edges and depth contours. The deep network is consequently starved of its learned structural descriptors, causing it to misclassify the target spray bottle as other cylindrical items such as shampoo (F1-score = 0.6920, recall = 1.0000) or paper lunch boxes (F1-score = 0.8669, recall = 0.7650). Documenting this specific hardware-material anomaly provides an authentic piece of self-critique for the manuscript, proving that while standalone NIR is highly reliable under visible light degradation, specialized material coatings can still alter infrared wave reflections under narrow ambient combinations.

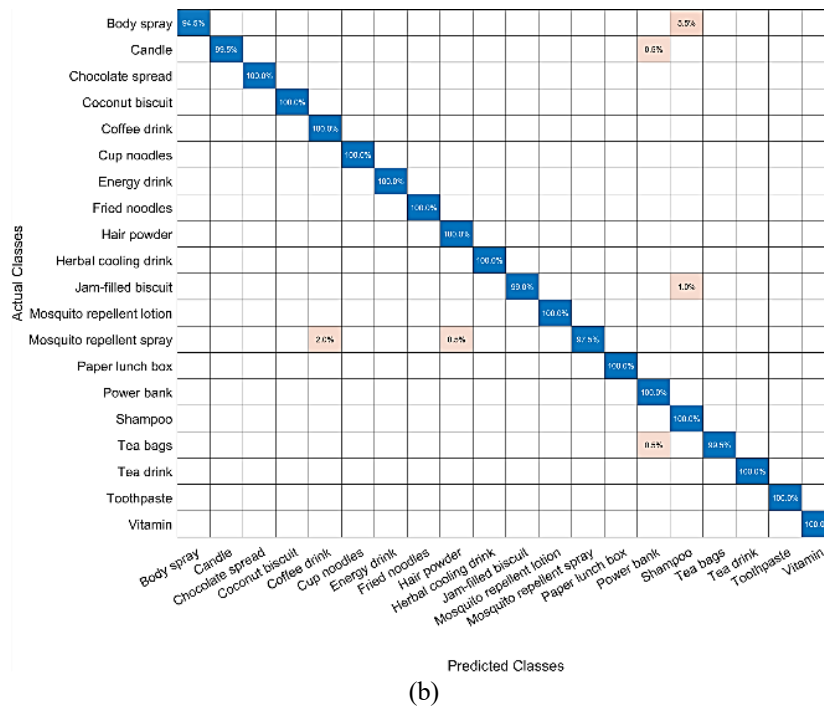
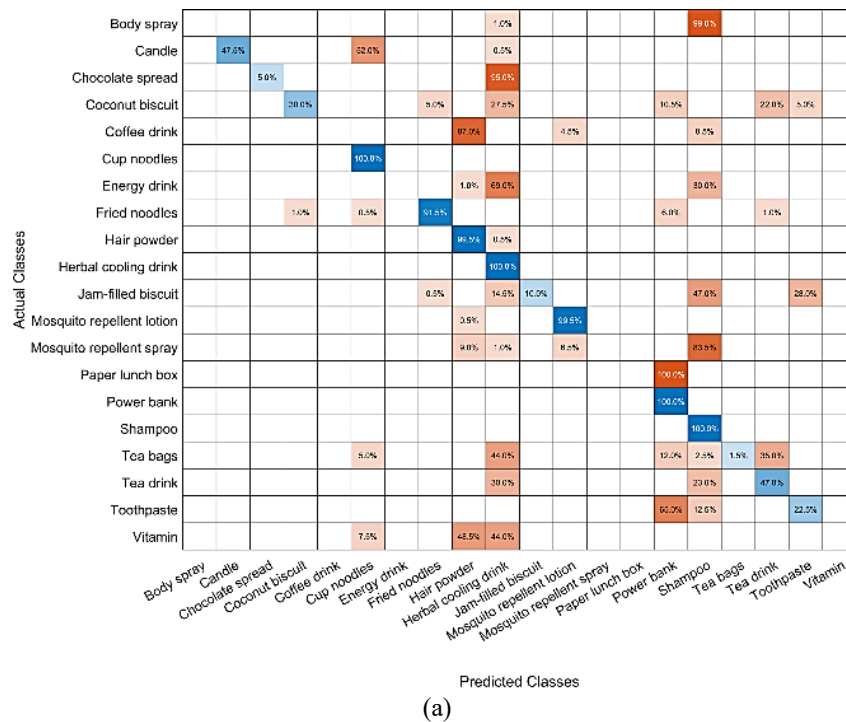
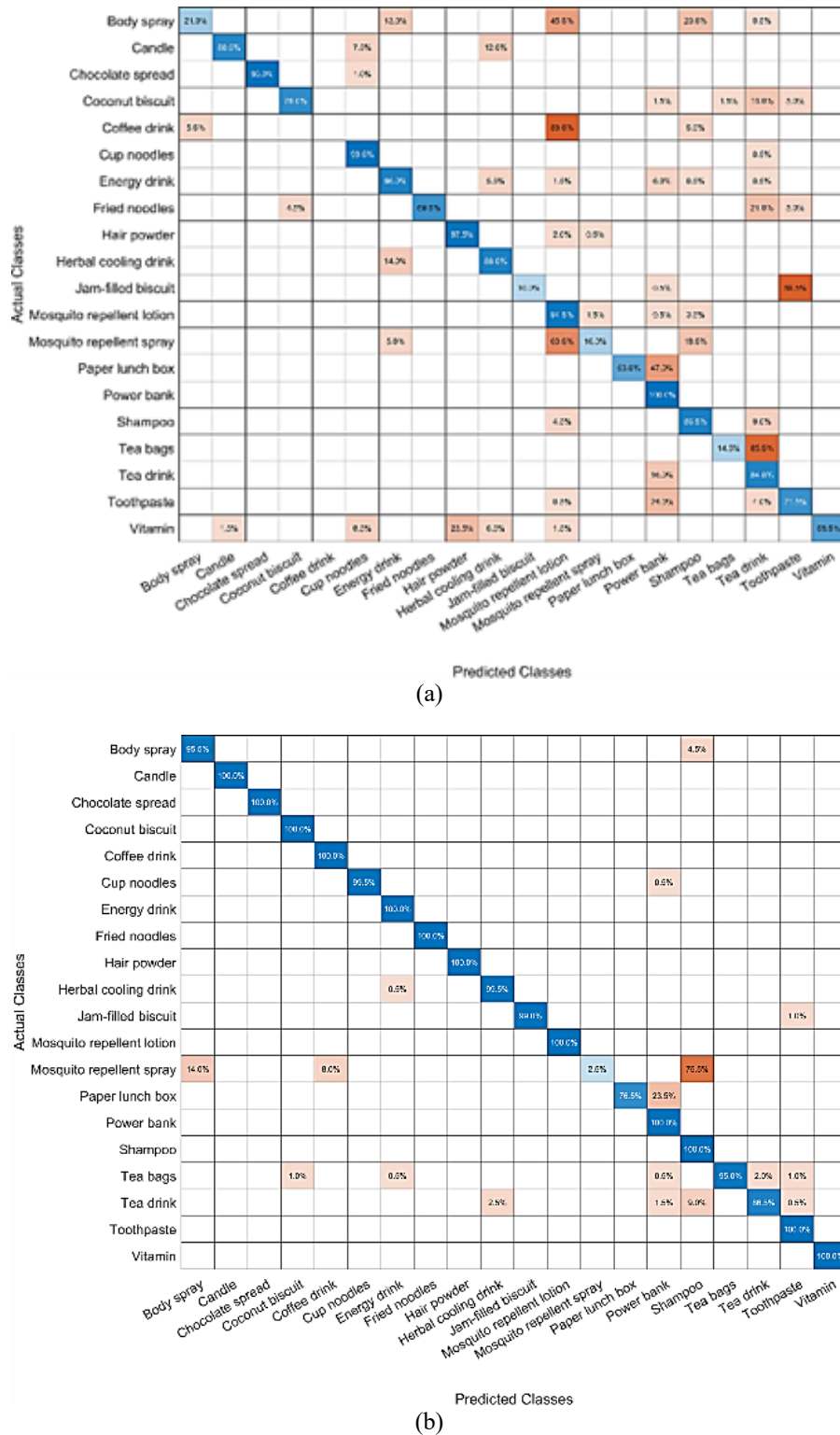


Figure 6. Normalized confusion matrices for green lighting conditions: (a) RGB data and (b) NIR data



4.4. External baseline comparison

To evaluate the absolute technical necessity of utilizing a CNN framework, an external comparative benchmark was executed against a traditional, instance-based KNN, where $k = 3$ pipeline. This baseline classifier operates directly on global 64-bit color distribution histograms extracted from the object images. Introducing this classical algorithmic baseline serves a critical analytical purpose: it verifies whether the

challenging problem of household object recognition under severe lighting variations can be resolved using computationally lighter, statistical color-space tracking models, or if it genuinely demands the invariant, high-level structural features extracted via a deep network pipeline.

To provide an immediate, clear interpretation of these cross-modal trends and to avoid dense numerical text matrices, a comprehensive graphical visualization plotting the cross-class average F1-scores alongside their corresponding standard deviation ranges is presented in Figure 8. As visually demonstrated by the performance profiles, under the nominal white light condition (dim white), the traditional KNN color-matching pipeline achieves high operational stability within the visible RGB domain, registering an average F1-score of 0.983. This strong performance shows that when ambient illumination remains completely clear and uncorrupted, standard surface color distributions provide an exceptionally clean, easily distinguishable cue for instance tracking. Even within the infrared domain under nominal lighting, the KNN baseline remains highly competitive, yielding an average F1-score of 0.941.

However, when ambient light shifts away from standard white into monochromatic color bands, the visual contrast highlights a catastrophic performance gap. Across the red, green, and blue light environments, the visible RGB KNN baseline model collapses entirely to a zero average score. This total blindness occurs because a uniform shift in environmental light completely alters the color distance configurations of the image pixels, rendering shallow color-space models unviable for unconstrained robotics. Shifting the KNN baseline's hardware channel to the RAW infrared spectrum provides only minimal relief, with the NIR KNN model logging weak average F1-scores of just 0.016 under red light, 0.190 under green light, and 0.019 under blue light. This reveals that simply changing the sensor type is not enough; traditional shallow algorithms cannot extract robust object representations from RAW gray-level values when the illumination field shifts.

This stark performance divergence is clearly summarized by the overall cross-condition blocks on the far right of Figure 8, where both statistical KNN baselines show severe constraints. In sharp contrast deep NIR convolutional framework maintains an exceptional, tightly clustered average F1-score across all simulated disruptions. This narrow standard deviation boundary visually confirms that the proposed combination of high-level deep features and NIR sensing delivers consistent, resilient perception depth across variable lighting changes, making it a proven technical necessity for robust domestic robotics deployment.

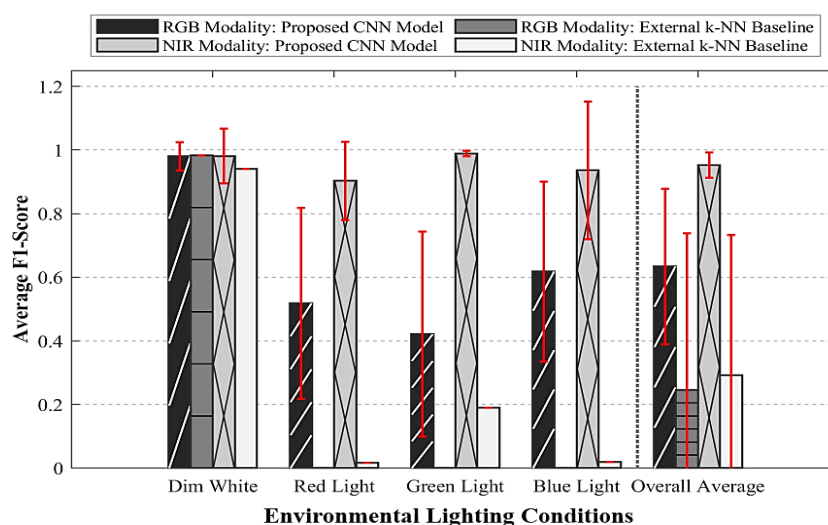


Figure 8. Macro-averaged F1-scores and corresponding standard deviations for the CNN models and traditional baseline KNN classifiers across individual and averaged lighting conditions

5. CONCLUSION

This study demonstrates that while standard RGB visual models suffer severe performance degradation under low-light and monochromatic lighting variations, a standalone NIR imaging framework consistently maintains high recognition accuracy across diverse illumination scenarios. From a practical deployment standpoint, processing single-channel NIR data via lightweight architectures eliminates the need for parameter-heavy dual-stream fusion networks. This structural footprint reduction significantly slashes computational complexity, inference latency, and energy consumption, making standalone NIR perception highly feasible for resource-constrained embedded robotic edge platforms operating in assistive smart homes, night-time monitoring, and elderly care environments. Future research will expand upon this standalone

baseline by developing adaptive attention mechanisms for targeted multimodal integration with depth or visible spectral arrays. These future iterations aim to dynamically mitigate localized, material-specific anomalies (such as specular dispersion and translucent subsurface scattering) further enhancing the perceptual robustness, efficiency, and operational safety of autonomous service robots in unconstrained domestic settings.

ACKNOWLEDGMENTS

The authors express their sincere gratitude to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for providing the research opportunities and administrative support during the 2024 fiscal year. Special thanks are also extended to the Institut Teknologi Sepuluh Nopember (ITS) for the facilities and technical assistance provided throughout the experimental phase of this study.

FUNDING INFORMATION

This work was partially funded by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia under the research contract No. 1767/PKS/ITS/2024 (Master Contract No. 038/E5/PG.02.00.PL/2024).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Attamimi	✓	✓		✓	✓		✓		✓	✓	✓	✓	✓	✓
Muhammad Ammar		✓	✓		✓	✓		✓	✓		✓			
Aiman Majid														
Hendra Kusuma	✓				✓		✓			✓		✓		
Muhammad Rivai	✓				✓		✓			✓			✓	✓
Rudy Dikairono	✓				✓		✓			✓			✓	✓

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [MA], upon reasonable request.

REFERENCES

[1] Y. Ye, X. Ma, X. Zhou, G. Bao, W. Wan, and S. Cai, "Dynamic and real-time object detection based on deep learning for home service robots," *Sensors*, vol. 23, no. 23, Nov. 2023, doi: 10.3390/s23239482.

[2] I. K. Mirani, C. Tianhua, M. A. A. Khan, S. M. Aamir, and W. Menhaj, "Object recognition in different lighting conditions at various angles by deep learning method," Oct. 2022, *arXiv:2210.09618*.

[3] J. Liu and Y. Jin, "A comprehensive survey of robust deep learning in computer vision," *Journal of Automation and Intelligence*, vol. 2, no. 4, pp. 175–195, Nov. 2023, doi: 10.1016/j.jai.2023.10.002.

[4] H. Wu *et al.*, "WTEFNet: real-time low-light object detection for advanced driver assistance systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 75, pp. 1–17, 2026, doi: 10.1109/TIM.2026.3655940.

[5] S. Manzoor *et al.*, "3D recognition based on sensor modalities for robotic systems: a survey," *Sensors*, vol. 21, no. 21, Oct. 2021, doi: 10.3390/s21217120.

[6] Z. Li, S. Pu, M. Ji, F. Zeng, and B. Li, "Visible and near-infrared image fusion based on information complementarity," *CAAI Transactions on Intelligence Technology*, vol. 10, no. 1, pp. 193–206, Feb. 2025, doi: 10.1049/cit2.12378.

- [7] Z. Liu *et al.*, "Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion," *IEEE Access*, vol. 8, pp. 2327–2336, 2020, doi: 10.1109/ACCESS.2019.2962513.
- [8] C. Liu, Q. Feng, Y. Sun, Y. Li, M. Ru, and L. Xu, "YOLACTFusion: an instance segmentation method for RGB-NIR multimodal image fusion based on an attention mechanism," *Computers and Electronics in Agriculture*, vol. 213, Oct. 2023, doi: 10.1016/j.compag.2023.108186.
- [9] X. Pan, Z. Yu, and Z. Yang, "A deep learning multimodal fusion framework for wood species identification using near-infrared spectroscopy GADF and RGB image," *Holzforschung*, vol. 77, no. 11–12, pp. 816–827, Dec. 2023, doi: 10.1515/hf-2023-0062.
- [10] X. Xiang *et al.*, "Infrared-visible image fusion meets object detection: towards unified optimization for multimodal perception," *Remote Sensing*, vol. 17, no. 21, Nov. 2025, doi: 10.3390/rs17213637.
- [11] Z. Erickson, E. Xing, B. Srirangam, S. Chernova, and C. C. Kemp, "Multimodal material classification for robots using spectroscopy and high resolution texture imaging," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2020, pp. 10452–10459, doi: 10.1109/IROS45743.2020.9341165.
- [12] Z. Xie, "Impact of RAW vs. RGB images on low-light object detection in autonomous driving," *Applied and Computational Engineering*, vol. 81, no. 1, pp. 148–157, Nov. 2024, doi: 10.54254/2755-2721/81/20241050.
- [13] M. J. Kim *et al.*, "OpenVLA: an open-source vision-language-action model," Sep. 2024, *arXiv:2406.09246*.
- [14] H. Fang *et al.*, "SAM2Act: integrating visual foundation model with a memory architecture for robotic manipulation," Jul. 2025, *arXiv:2501.18564*.
- [15] M. H. F. III, A. Pickeral, E. Marquez, M. C. Smith, and J. C. Calhoun, "Efficient vision transformers for autonomous off-road perception systems," *Journal of Computer and Communications*, vol. 12, no. 09, pp. 188–207, 2024, doi: 10.4236/jcc.2024.129011.
- [16] N. Manakitsa, G. S. Maraslidis, L. Moysis, and G. F. Fragulis, "A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision," *Technologies*, vol. 12, no. 2, Jan. 2024, doi: 10.3390/technologies12020015.
- [17] A. Salari, A. Djavadifar, X. Liu, and H. Najjaran, "Object recognition datasets and challenges: a review," *Neurocomputing*, vol. 495, pp. 129–152, Jul. 2022, doi: 10.1016/j.neucom.2022.01.022.
- [18] X. Shang *et al.*, "CCLDet: a cross-modality and cross-domain low-light detector," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 3, pp. 3284–3294, Mar. 2025, doi: 10.1109/TITS.2024.3522086.
- [19] Y. Cohen, A. Biton, and S. Shoval, "Fusion of computer vision and AI in collaborative robotics: a review and future prospects," *Applied Sciences*, vol. 15, no. 14, Jul. 2025, doi: 10.3390/app15147905.
- [20] M. Tan and Q. V. Le, "EfficientNetV2: smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, Jun. 2021.
- [21] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.
- [22] D. Zhang, Z. Liu, and X. Shi, "Transfer learning on EfficientNet for remote sensing image classification," in *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Dec. 2020, pp. 2255–2258, doi: 10.1109/ICMCCE51767.2020.00489.
- [23] S. Patel and D. Patel, "Transfer learning with EfficientNet for computer vision based automatic gender classification," in *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Jan. 2023, pp. 1–5, doi: 10.1109/ICAECT57570.2023.10117661.
- [24] C. Han *et al.*, "Near infrared spectroscopy (NIRS) model based on transfer learning," *Analytical Letters*, vol. 58, no. 14, pp. 2468–2480, Sep. 2025, doi: 10.1080/00032719.2024.2417349.
- [25] T. Njoroge, R. Kibuku, and K. Mugoye, "Comparative and edge-hybrid modeling of EfficientNetV2 and MobileNetV2 for multi-class crop disease classification with statistical validation," *Journal of Edge Computing*, vol. 4, no. 2, pp. 234–262, Nov. 2025, doi: 10.55056/jec.905.
- [26] H. Zhang, Y. Chen, and Y. Liu, "EfficientNet-EA for visual location recognition in natural scenes," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 596–603, Jan. 2025, doi: 10.1109/LRA.2024.3511379.
- [27] T. Ahmad *et al.*, "EfficientNet-based robust recognition of peach plant diseases in field images," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 2073–2089, 2022, doi: 10.32604/cmc.2022.018961.
- [28] W. Zhong, L. Tian, D. T. Le, and H. Rezaatofighi, "Improving visual perception of a social robot for controlled and in-the-wild human-robot interaction," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, Mar. 2024, pp. 1199–1203, doi: 10.1145/3610978.3640648.
- [29] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, vol. 97, pp. 6105–6114.
- [30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations*, 2019.

BIOGRAPHIES OF AUTHORS



Muhammad Attamimi    received his B.E., M.E., and D.E. degrees from the University of Electro Communications in 2010, 2012, and 2015, respectively. He received scholarship from Ministry of Education, Culture, Sports, Science and Technology Japan (MEXT) for his B.E. and M.E. courses. From 2012 to 2015, he was also a research fellow (DC1) of Japan Society for the Promotion of Science (JSPS). He was with a postdoctoral researcher at the Department of Mechanical Engineering and Intelligent Systems, The University of Electro communications (UEC) from April 2015 to December 2015. Since January 2016, he was with Tamagawa University Brain Science Institute as a commission researcher for six months. Currently, he is a lecturer at Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His research interests are computer vision, visual recognition, machine learning, multimodal categorization, artificial intelligence, probability robotics, intelligent systems, and intelligent robotics. He can be contacted at email: attamimi@its.ac.id.



Muhammad Ammar Aiman Majid    earned his bachelor's degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 2024. His interests include electronic circuit and hardware design, machine learning, and embedded systems. His notable achievements include incentivized PKM-AI scientific article with a topic of Kalman filtering for battery health estimation, written based on an internship project. He can be contacted at email: aimanmajid2112@gmail.com.



Hendra Kusuma    received the B.S. and Ph.D. degree from Institut Teknologi Sepuluh Nopember (ITS) Surabaya both in Electrical Engineering, in 1988 and 2016, respectively. He also received the M.S. degree from Curtin University of Technology in renewable energy engineering in 2001. From 1989 until now, he is a lecturer in the Department of Electrical Engineering, ITS. His research interests are in artificial intelligence, machine learning, pattern recognition, power electronics, and IoT as well as in assistive technology. He is currently head of Microelectronic and Embedded System Laboratory in the Department of Electrical Engineering, ITS, and he is also as IEEE member. He can be contacted at email: hendraks@its.ac.id.



Muhammad Rivai    received the Ph.D. degree, in 2006. He is currently a lecturer with the Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His research interests include sensors, chemical sensors, gas sensors, electronics, robotics, and artificial intelligence applications. He can be contacted at email: muhammad_rivai@its.ac.id.



Rudy Dikairono    is a distinguished professional in the field of electrical engineering. He earned his bachelor's degree in Electrical Engineering (EE) from Institut Teknologi Sepuluh Nopember (ITS) in 2004, followed by a master's degree from EE-ITS and Fachhochschule Darmstadt, Germany, in 2009, and a doctoral degree from EE-ITS in 2021. Throughout his career, he has been a dedicated member of the academic community, serving as a lecturer at EE-ITS since 2005. He also held the position of head of the Information and Communication Technology and Robotics Laboratory at ITS from 2012 to 2016 and currently leads as the head of the Center of Excellence in Artificial Intelligence for Healthcare and Society at ITS. He can be contacted at email: rudydikairono@its.ac.id.