Comparative Evaluation of Transfer Learning Models for Household Appliance Image Classification

Fuzy Yustika Manik*, Anandhini Medianty Nababan, and T. M. Rezha Taufiqurrahman

Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

Email: fuzy.yustika@usu.ac.id (F.Y.M.); nababan.anandhini@usu.ac.id (A.M.N.); rezhataufik1234@gmail.com (T.M.R.T.)

*Corresponding author

Manuscript received January 25, 2026; revised April 13, 2026; accepted May 26, 2026; published September 11, 2026

***Abstract*—Household appliances contribute substantially to residential energy demand, yet appliance-level monitoring remains difficult due to the cost and limited availability of sensing infrastructure. This study presents a systematic comparative evaluation of deep transfer learning models for visual recognition of household appliance positioning, which is a foundational component for future energy-aware systems. A mixed-source dataset of 1500 images (1100 online and 400 real-world field samples) across 11 appliance categories was constructed. Three architectures, DenseNet121, ResNet50, and EfficientNet-B0, were evaluated under identical experimental settings. To ensure robustness and reproducibility, performance was assessed using accuracy, precision, recall, F1-score, an external real-world test set, and 5-fold stratified cross-validation with results reported as mean ± standard deviation. Experimental results show that DenseNet121 achieves the best performance, with 97.0% accuracy, 96.0% validation accuracy, and an F1-score of 98.4%, outperforming EfficientNet-B0 and ResNet50. Cross-validation yields a mean accuracy of 97.0% ± 0.29%, indicating strong stability and generalization on medium-scale datasets. In addition, a prototype application demonstrates how visual recognition can be integrated into a user-facing system that provides contextual energy-related information to enhance user awareness. Rather than directly measuring or predicting electrical consumption, this work establishes a robust visual recognition framework and provides empirical evidence that DenseNet121-based transfer learning offers an effective and stable solution for household appliance image classification. The findings provide a reliable technical foundation for the future development of vision-based energy-aware applications.**

***Keywords*—deep learning, DenseNet121, energy-aware systems, household appliance recognition, image classification, transfer learning**

I. INTRODUCTION

Household appliances are among the major contributors to electricity consumption in both developed and developing countries [1, 2]. Inefficient usage can increase operational costs and place additional pressure on energy infrastructure and natural resources [3, 4], while rising appliance demand driven by economic growth may further intensify environmental impacts and carbon emissions [1, 5, 6]. Unmanaged appliance usage patterns can also contribute to instability in power systems [7]. However, appliance-level monitoring remains challenging in practice because many households lack access to smart plugs or watt meters due to cost and infrastructure limitations [8]. Consequently, low-cost alternative approaches that can support appliance identification and monitoring without requiring additional hardware are increasingly needed.

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have demonstrated strong performance in visual recognition tasks [9, 10]. CNN-based approaches have been successfully applied across various domains, including remote sensing [11] and medical imaging [12]. However, CNN models trained from scratch typically require large datasets and are prone to overfitting when data availability is limited [13, 14]. Transfer learning has therefore emerged as an effective strategy, enabling models to leverage features learned from large-scale datasets such as ImageNet and adapt them to new tasks with improved generalization [13, 15]. Prior studies further confirm the effectiveness of transfer learning in a range of image classification tasks, including equipment recognition and fault detection [15, 16].

Most existing research on household energy monitoring focuses on sensor-based approaches, such as Non-Intrusive Load Monitoring (NILM) and smart meter analysis [17, 18]. While these methods provide valuable contributions, they require specialized infrastructure and differ fundamentally from image-based approaches. Image-based appliance recognition using deep learning, therefore, remains relatively underexplored, particularly with respect to systematic architectural comparison and robust validation protocols.

Motivated by this research gap, this study presents a comparative evaluation of deep transfer learning models for household appliance image classification. Three widely used architectures, DenseNet121, ResNet50, and EfficientNet-B0, are evaluated under consistent experimental conditions using a mixed-source dataset. Model performance is assessed using 5-fold cross-validation and statistical stability indicators. Rather than directly estimating energy consumption, this work aims to establish a robust visual recognition foundation that can support the development of future intelligent energy-aware systems.

This study contributes to the field by constructing a household appliance image dataset that combines online sources with real-world field data across 11 appliance categories. It further provides a systematic comparative evaluation of 3 deep transfer learning architectures under consistent experimental settings. To ensure robustness and reduce evaluation bias, a comprehensive evaluation framework based on *k*-fold cross-validation and statistical stability analysis (mean and standard deviation) is employed. The experimental results provide empirical evidence that DenseNet121 achieves superior accuracy and stability on medium-scale datasets, indicating its suitability for household appliance image classification. Overall, the findings offer a reliable technical foundation for the future development of vision-based household appliance recognition systems and intelligent energy-aware applications. This study distinguishes itself by combining cross-validation, external real-world testing, ablation analysis, and statistical significance testing within a unified evaluation framework for household appliance recognition.

II. Related Work

Most studies on household energy monitoring rely on sensor-based techniques, particularly Non-Intrusive Load Monitoring (NILM) using smart meter data. Manca and Massidda [19] proposed a deep learning-based NILM model using low-resolution data, while Gan *et al.* [20] introduced self-attention mechanisms for improved load identification. Liu *et al.* [21] combined supervised and unsupervised neural networks for appliance identification, and Ullah *et al.* [22] explored clustering and autoencoders for energy profiling. Other works have further extended these approaches using CNN-, Long Short Term Memory (LSTM)-, and hybrid-based architectures [23–28].

Although these methods provide valuable insights, they rely on electrical signal measurements and specialized infrastructure, which limits their applicability in environments where such devices are unavailable. In contrast, image-based appliance recognition depends solely on visual information and can be deployed using widely available devices such as smartphones and cameras, making it more accessible for broader adoption.

Deep learning techniques, particularly CNNs, have demonstrated strong performance in image classification tasks across various domains [9, 10]. However, models trained from scratch typically require large datasets and are prone to overfitting when data availability is limited [13, 14]. Transfer learning has therefore been widely adopted to improve generalization by leveraging pretrained representations learned from large-scale datasets such as ImageNet [13, 15]. Prior studies have confirmed the effectiveness of transfer learning in specialized image recognition applications, including industrial inspection and equipment classification [15, 16].

Recent reviews and empirical studies have further explored the evolution and application of deep learning and transfer learning in image processing. Trigka and Dritsas [29] provided a broad survey of deep learning techniques and their implications for model efficiency, robustness, and evaluation in modern image processing tasks. Gu and Lee [30] formalized the concept of real-world feature transfer learning and demonstrate its effectiveness on medical image classification tasks using pretrained models. Moreover, hybrid frameworks that integrate multiple pretrained architectures have shown improved performance in applied contexts, such as delamination detection in structural health monitoring [31]. Comprehensive explorations of Deep Learning (DL) techniques for super-resolution and visual data enhancement also underscore the importance of rigorous evaluation and architectural choices when applying these methods across varied domains [32]. Recent systematic reviews further discuss the state of the art in deep learning for image analysis across multiple applications, reinforcing the need for robust evaluation practices in comparative studies [33].

Despite these advances, systematic comparative studies evaluating multiple transfer learning architectures for household appliance image classification under robust validation protocols remain limited. Most existing studies either focus on sensor-based approaches or evaluate only a single model without rigorous statistical validation. This study addresses this gap by benchmarking multiple architectures using *k*-fold cross-validation and stability analysis, thereby providing a more reliable and comprehensive evaluation framework. This work does not directly measure or estimate energy consumption but provides a visual recognition foundation that can support future integration with energy monitoring systems.

III. Methodology

This section describes the overall experimental procedure, including data acquisition and preprocessing, feature learning and classification, model training configuration, and model evaluation. Three Convolutional Neural Network (CNN) architectures are evaluated in this study: DenseNet121, ResNet50, and EfficientNet-B0. A high-level overview of the proposed experimental framework is illustrated in Fig. 1.

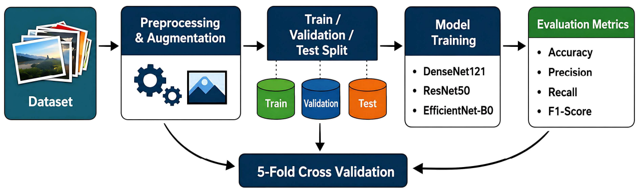


Fig. 1. Overview of the proposed experimental framework.

Fig. 1 illustrates the complete workflow of this study. The process begins with dataset collection from online sources and real-world environments, followed by preprocessing and data augmentation. The processed data are then used to train different deep learning models. Finally, the trained models are evaluated using quantitative metrics and cross-validation to ensure robustness.

A. Data Acquisition and Preprocessing

Representative samples of the dataset used in this study are shown in Fig. 2.

Fig. 2. Sample images of household appliance classes showing variations in lighting, background, viewpoint, and image quality.

A dataset of 1500 images was constructed from online sources (1100 images) and real-world field data (400 images). 11 appliance classes were included: air conditioner, oven, washing machine, laptop, hair dryer, iron, lamp, refrigerator, rice cooker, television, and vacuum cleaner. From these, 1400 images were used for training and validation (80/20 split), while 100 real-world images were reserved as an external test set to evaluate cross-domain generalization. All images were resized to 256 × 256 pixels and normalized to the range [0, 1]. Data augmentation (rotation ±20°, flipping, brightness adjustment, zooming, and color jitter) was applied only to the training data [14, 34, 35]. A 5-fold stratified cross-validation scheme was employed, and performance is reported as mean ± standard deviation [36, 37].

To ensure transparency and reproducibility, the detailed dataset composition and split protocol per class are summarized in Table 1. The table reports the number of images collected from online sources and field environments, along with their distribution between the cross-validation pool and the strictly held-out external test set.

Table 1. Dataset composition and split per class

Class	Online Images	Field Images	Total Images	External Test	CV Pool
Air conditioner	100	37	137	9	128
Oven	100	37	137	9	128
Washing machine	100	37	137	9	128
Laptop	100	37	137	9	128
Hair dryer	100	36	136	9	127
Iron	100	36	136	9	127
Lamp	100	36	136	9	127
Refrigerator	100	36	136	9	127
Rice cooker	100	36	136	9	127
Television	100	36	136	9	127
Vacuum cleaner	100	36	136	10	126
Total	**1100**	**400**	**1500**	**100**	**1400**

The external test set consists exclusively of field images and was never used during training or cross-validation. The CV pool (1400 images) was used for 5-fold stratified cross-validation, ensuring that class proportions were preserved across folds. This protocol minimizes data leakage and strengthens the reliability of the reported evaluation results.

B. Feature Learning and Classification

The architecture used for classification is illustrated in Fig. 3.

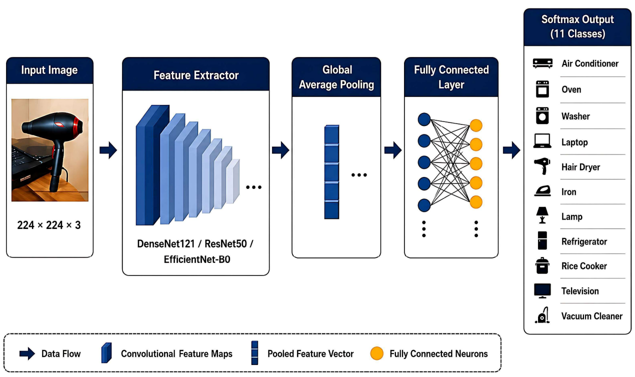


Fig. 3. Architecture of the classification models.

As shown in Fig. 3, all models share a common architecture comprising a feature-extraction backbone, a global average pooling layer, and a fully connected classification layer with a softmax activation. In addition to transfer learning models, a baseline CNN trained from scratch is included for comparison.

C. Baseline CNN Architecture

To provide a fair comparison with transfer learning models, a baseline CNN architecture was designed and trained from scratch. The architecture consists of four convolutional blocks, each including a convolutional layer (3 × 3 kernel), batch normalization, Rectified Linear Unit (ReLU) activation, and max pooling. The number of filters increases progressively from 32 to 256. The feature extraction stage is followed by a fully connected layer with 512 neurons and a dropout rate of 0.5 to reduce overfitting. The final classification layer uses a softmax activation to produce probability distributions over 11 appliance classes. This baseline model allows evaluation of the effectiveness of transfer learning compared to learning features from scratch on a medium-scale dataset.

D. Transfer Learning Models

Three pretrained architectures were evaluated in this study: DenseNet121, ResNet50, and EfficientNet-B0. These models were initialized with ImageNet-pretrained weights to leverage learned visual representations. Transfer learning was implemented by freezing the backbone layers during the initial 20 epochs to retain pretrained features. Afterward, all layers were unfrozen and fine-tuned to adapt the models to the target dataset. This strategy stabilizes early training while allowing deeper task-specific optimization in later stages.

E. Model Training Configuration

All models were trained using the same hyperparameter configuration to ensure fair comparison. The training settings are summarized in Table 2.

Table 2. Model training configuration

Parameter	Configuration
Optimizer	Adam
Learning rate	0.0001 (Reduce LR On Plateau)
Batch size	32
Maximum epochs	100
Loss function	Categorical cross-entropy
Transfer learning	Backbone frozen for first 20 epochs, then full fine-tuning

To improve model generalization and prevent overfitting, several regularization strategies were applied, including dropout with a rate of 0.5 on the fully connected layer, L2 regularization with a weight decay of 1×10^{-4}, and early stopping with a patience of 10 epochs. These strategies help stabilize the training process and ensure that the model does not merely memorize the training data.

F. Experimental Setup and Evaluation Strategy

All experiments were conducted on a workstation equipped with an NVIDIA RTX 3060 GPU (12GB VRAM), Intel Core i7 processor, and 32 GB RAM. The models were implemented using Python with the TensorFlow/Keras framework. Compute Unified Device Architecture (CUDA) acceleration was enabled to improve training efficiency. The evaluation procedure used in this study is illustrated in Fig. 4.

As shown in Fig. 4, model performance was evaluated using multiple evaluation metrics, including accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, while precision and recall evaluate class-wise performance. The F1-score provides a balanced measure between precision and recall. In this study, precision, recall, and F1-score are reported as weighted averages to account for class imbalance across appliance categories. To ensure robustness, 5-fold cross-validation was used, and results are reported as mean ± standard deviation. In addition, an external real-world test set was used to evaluate generalization performance on unseen data. To further validate performance differences, a paired *t*-test was conducted to assess statistical significance between models. An ablation study was also performed to evaluate the impact of transfer learning and data augmentation. Results show that removing transfer learning significantly reduces performance, highlighting its importance in this study.

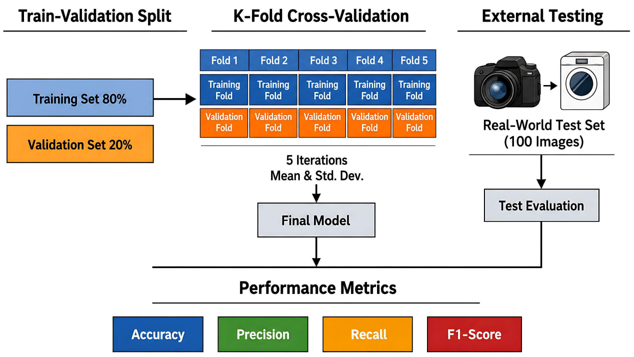


Fig. 4. Evaluation strategy.

IV. Result and Discussion

This section presents the experimental results and discusses the performance of the proposed models. The analysis focuses on comparing baseline CNN and transfer learning approaches, evaluating architectural performance, assessing generalization capability, and interpreting class-wise behavior using confusion matrix analysis.

A. Comparison between CNN without Transfer Learning and Transfer Learning

The training behavior of the baseline CNN model trained from scratch is illustrated in Fig. 5.

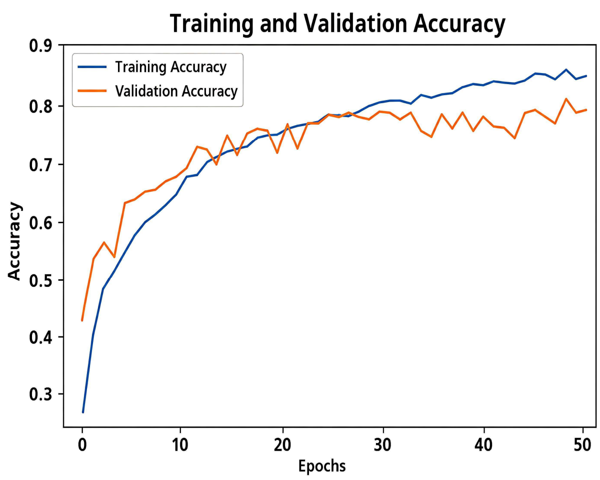
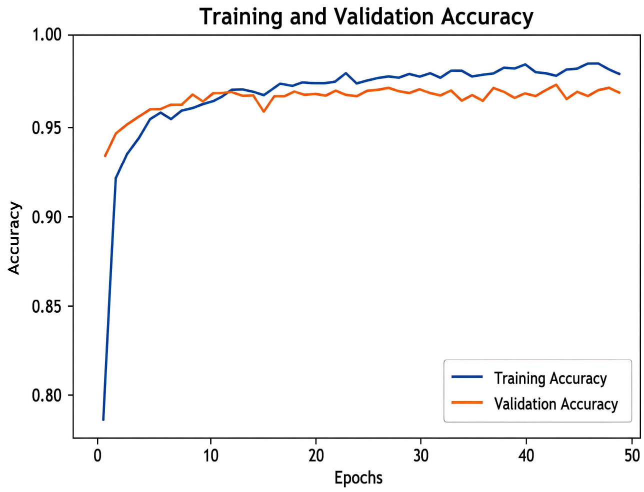
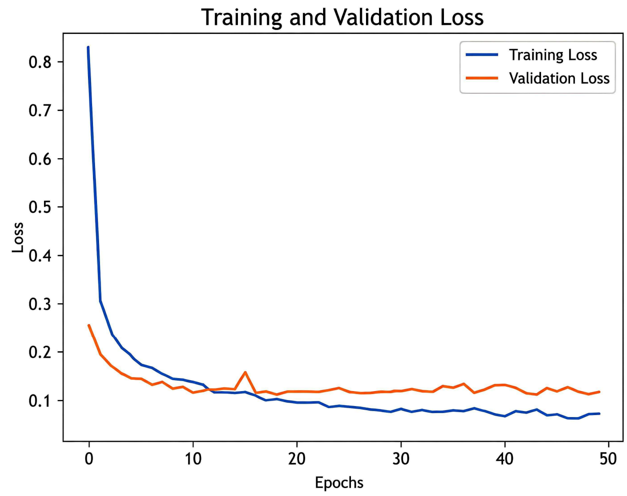


Fig. 5. Training and validation accuracy for the CNN model trained without transfer learning.

Fig. 5 shows that the baseline CNN achieved approximately 85% training accuracy and 80% validation accuracy. The gap between training and validation curves indicates a tendency toward overfitting, suggesting that the model struggles to generalize when trained solely on the available dataset without pretrained features. This confirms that training from scratch is less suitable for medium-scale datasets, as the model must learn all visual features from the beginning. In contrast, the learning dynamics of the transfer learning-based model are illustrated in Fig. 6.



(a)



(b)

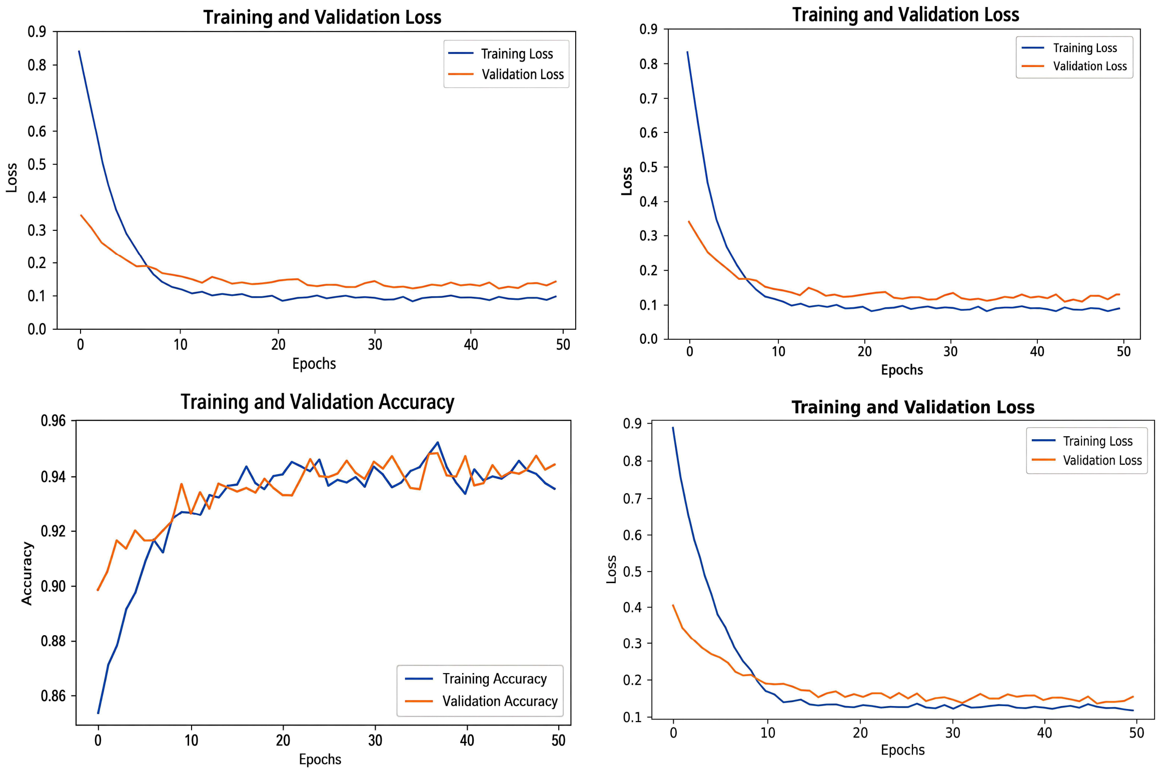
Fig. 6. Training and validation accuracy and loss curves for the DenseNet121 model using transfer learning: (a) training and validation accuracy, (b) training and validation loss.

As shown in Fig. 6, the DenseNet121-based model achieved approximately 97% training accuracy and 96% validation accuracy, with training and validation curves closely aligned. This behavior indicates improved learning stability and stronger generalization capability. The use of pretrained ImageNet features enables the model to capture more effective visual representations, resulting in faster convergence and reduced overfitting. To further validate the consistency of this behavior across different architectures, the training and validation performance of EfficientNet-B0 and ResNet50 are presented in Fig. 7. These figures illustrate the accuracy and loss curves over 50 training epochs, providing additional insight into convergence behavior and model stability.

As shown in Fig. 7, EfficientNet-B0 exhibits rapid convergence during the early training stages, with both training and validation accuracy increasing steadily and remaining closely aligned. This indicates good generalization performance, though minor fluctuations are observed, suggesting slightly lower stability than DenseNet121. The corresponding loss curves show a consistent decrease, further confirming the optimization's effectiveness.

In contrast, ResNet50 exhibits slower convergence and greater variability in both the accuracy and loss curves. The gap between training and validation performance is relatively large, indicating greater sensitivity to data variations and a greater tendency toward overfitting. The loss curves also reflect less stable optimization, with more pronounced fluctuations throughout the training process.

Overall, these observations confirm that while EfficientNet-B0 provides competitive performance with good generalization, DenseNet121 remains the most stable architecture, and ResNet50 exhibits comparatively lower robustness in this task. These results further emphasize the importance of architectural design for achieving stable, generalizable learning performance.



(a) (b)

Fig. 7. Training and validation accuracy and loss curves for (a) the EfficientNet-B0 and (b) the ResNet50 model using transfer learning.

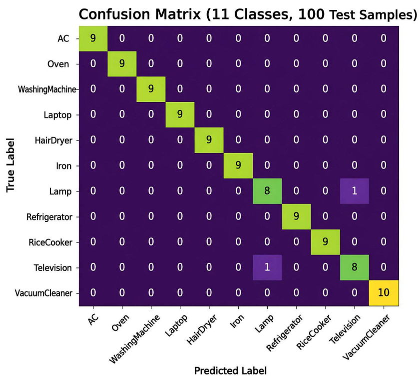
B. Confusion Matrix and Error Analysis

The class-wise performance of the evaluated models was further analyzed using a confusion matrix, as shown in Fig. 8.

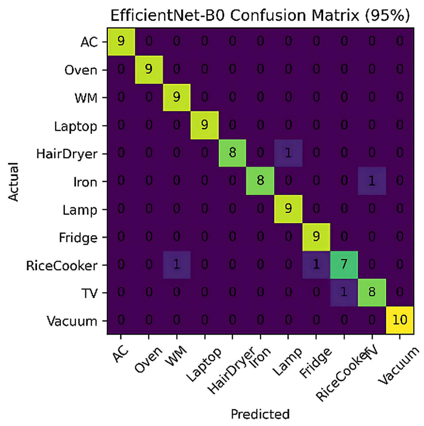
The confusion matrices provide detailed insight into each model's classification behavior across appliance categories. DenseNet121 achieves the highest performance, correctly classifying 97 out of 100 test samples, with a strong diagonal concentration, indicating high classification consistency. EfficientNet-B0 achieves competitive performance, correctly classifying 95 samples. While its diagonal distribution remains dominant, a slightly higher number of misclassifications is observed compared to DenseNet121, indicating a minor reduction in classification stability. ResNet50 shows comparatively lower performance, correctly classifying 93 out of 100 samples.

The confusion matrix reveals a more dispersed pattern of misclassifications, suggesting that the model is more sensitive to variations in visual features and less effective in distinguishing certain appliance categories. This comparative analysis confirms that architectural design plays a critical role in classification robustness, particularly in scenarios involving visually similar object categories.

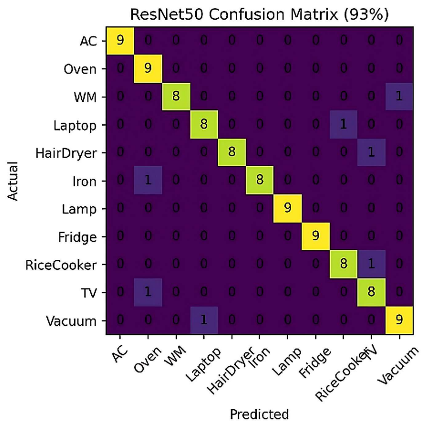
Across all evaluated models, misclassifications are predominantly observed among visually similar classes, particularly between lamp and television. These errors are likely caused by similarities in object shape, illumination patterns, and background conditions that are common in real-world environments. To further illustrate these limitations, representative misclassified samples are shown in Fig. 9.



(a)



(b)



(c)

Fig. 8. Comparative confusion matrices of (a) DenseNet121, (b) EfficientNet-B0, and (c) ResNet50.

(a)

(b)

(c)

Fig. 9. Examples of misclassified samples: (a) a laptop misclassified as a television, (b) a vacuum cleaner misclassified as a hair dryer, and (c) a television misclassified as a lamp.

As shown in Fig. 9, misclassification cases occur primarily between objects with overlapping visual characteristics. For instance, laptops may be misclassified as televisions due to their screen-like appearance and rectangular structure. Similarly, vacuum cleaners can be confused with hair dryers because of their elongated shape and handheld design. In addition, televisions may be misclassified as lamps because of their similar lighting patterns and reflective surfaces. These examples highlight the inherent challenge of distinguishing objects with similar visual features, particularly in unconstrained environments. This analysis provides deeper insight beyond overall accuracy, revealing class-specific behavior and the practical limitations of the models.

These observations further highlight the challenge of distinguishing objects with overlapping visual characteristics. While the confusion matrix provides a quantitative overview of model performance, the misclassification examples offer qualitative insights into specific failure cases and model limitations. This behavior is consistent with earlier findings: DenseNet121 demonstrates more stable performance, while EfficientNet-B0 and ResNet50 show greater variability.

C. Performance Comparison across Architectures

The quantitative performance comparison among DenseNet121, EfficientNet-B0, and ResNet50 is presented in Table 3. The evaluation metrics, including precision, recall, and F1-score, are reported as weighted averages to account for class imbalance across appliance categories.

The results in Table 3 indicate that DenseNet121 consistently outperforms the other architectures across all evaluation metrics. This performance advantage stems from DenseNet's dense connectivity, which encourages feature reuse and improves gradient flow. EfficientNet-B0 shows competitive performance with slightly lower accuracy but offers better computational efficiency, making it suitable for resource-constrained applications. ResNet50 demonstrates stable but lower performance, which aligns with prior findings that residual networks typically require larger datasets to achieve optimal performance.

Table 3. Performance comparison of transfer learning architectures (mean±standard deviation, weighted metrics)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DenseNet121	97.0 ± 0.29	98.8 ± 0.31	98.1 ± 0.27	98.4 ± 0.25
EfficientNet-B0	95.0 ± 0.36	97.9 ± 0.40	96.9 ± 0.38	97.4 ± 0.36
ResNet50	93.0 ± 0.50	97.1 ± 0.52	95.6 ± 0.49	96.3 ± 0.47

D. Generalization and Stability Analysis

To further assess model stability and generalization beyond test-set evaluation, the models were evaluated using 5-fold cross-validation, as summarized in Table 4.

Table 4. Cross-validation results (mean ± standard deviation)

Model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean (%)	Std.Dev. (%)
DenseNet121 (TL)	96.8	97.1	97.3	96.5	97.2	97.0	0.29
EfficientNet-B0(TL)	94.5	95.2	95.0	94.3	95.1	95.0	0.36
ResNet50 (TL)	92.1	93.5	92.7	93.1	92.8	93.0	0.50

DenseNet121 achieved the highest average accuracy and the lowest standard deviation, indicating superior stability across different data splits. The low variance suggests that the model is less sensitive to changes in data distribution and possesses strong generalization capability. EfficientNet-B0 demonstrates relatively stable performance, while ResNet50 shows higher variability, suggesting greater sensitivity to data composition.

E. Statistical Significance Analysis

To further validate whether the observed performance differences between models are statistically significant, a paired t-test was conducted using the cross-validation results. This analysis aims to determine whether the performance improvements are consistent and not due to random variation across different data splits.

Table 5. Paired t-test results between models

Comparison	p-value	Significance
DenseNet121 vs EfficientNet	<0.001	Significant
DenseNet121 vs ResNet50	<0.001	Significant
EfficientNet vs ResNet50	<0.001	Significant

As shown in Table 5, all pairwise comparisons yield p-values lower than 0.001, indicating that the performance differences between the evaluated models are statistically significant. In particular, DenseNet121 significantly outperforms both EfficientNet-B0 and ResNet50, confirming that the observed improvements are consistent across different folds and not due to random variation. These results strengthen the reliability of the experimental findings and demonstrate that DenseNet121's superior performance is statistically significant. Furthermore, the consistent differences between EfficientNet-B0 and ResNet50 also indicate that architectural design plays a critical role in determining classification performance.

F. Model Complexity and Computational Analysis

In addition to classification performance, model complexity and computational cost were analyzed to evaluate the practical feasibility of each architecture for real-world deployment. This analysis is important because high-performing models with excessive computational requirements may be less suitable for implementation in resource-constrained environments such as mobile or edge-based systems. Therefore, several indicators were considered, including the number of parameters, model size, and computational cost. The quantitative comparison of model complexity is summarized in Table 6.

Table 6. Model complexity and computational cost comparison

Model	Parameters (Millions)	Model Size (MB)	Floating Point Operations (FLOPs) (G)	Relative Training Time
DenseNet121	8.0	32	2.9	Moderate
EfficientNet-B0	5.3	20	0.39	Fast
ResNet50	25.6	98	4.1	Slow

Table 6 shows that EfficientNet-B0 has the fewest parameters and the lowest computational cost, making it attractive for lightweight applications. However, its classification performance remains slightly below that of DenseNet121. In contrast, ResNet50 incurs the highest computational burden due to its large parameter count, yet its predictive performance does not surpass that of DenseNet121. DenseNet121 demonstrates a favorable balance between performance and efficiency. Despite having significantly fewer parameters than ResNet50, it achieves superior accuracy and stability. This confirms that the dense connectivity mechanism allows more efficient feature reuse and stronger gradient propagation without requiring excessive model capacity. From an application perspective, this balance makes DenseNet121 particularly suitable for practical deployment in intelligent household appliance recognition systems, where both accuracy and computational efficiency are essential. In addition to model complexity, training time, and inference time per image were measured. DenseNet121 requires moderate training time but provides efficient inference (~15 ms/image), making it suitable for near-real-time applications. EfficientNet-B0 shows the fastest inference (~10 ms/image), while ResNet50 has the highest computational cost (~22 ms/image).

G. Ablation Study

An ablation study was conducted to evaluate the contribution of key components in the proposed framework, particularly transfer learning and data augmentation. Three configurations were evaluated: (1) baseline CNN trained from scratch without augmentation, (2) baseline CNN with data augmentation, and (3) transfer learning with data augmentation. The objective of this analysis is to understand how each component influences overall classification performance and generalization capability.

Table 7. Ablation study results

Configuration	Accuracy (%)
CNN (No Augmentation)	80
CNN + Augmentation	85
DenseNet121 (Transfer Learning + Augmentation)	97

As shown in Table 7, the baseline CNN without augmentation achieves the lowest performance due to limited feature representation and a higher tendency to overfit. Including data augmentation improves accuracy by increasing data diversity and enhancing generalization. However, the most significant improvement is observed when transfer learning is applied. By leveraging pretrained feature representations from large-scale datasets, the model achieves substantially higher performance. Specifically, removing transfer learning reduces accuracy by approximately 12%, while removing data augmentation decreases performance by around 4%–5%. These findings confirm that transfer learning is the most critical factor in achieving high classification performance, while data augmentation further enhances model robustness.

Although this analysis focuses on DenseNet121 as the best-performing model, similar trends are expected across other architectures, as transfer learning and data augmentation are fundamental components in deep learning-based image classification. This analysis highlights the importance of combining effective architectural design with appropriate training strategies to achieve optimal performance.

H. System Implementation

To complement the quantitative evaluation, this study presents a prototype application demonstrating how the proposed visual recognition model can be integrated into a practical, user-oriented system. The prototype allows users to upload appliance images and obtain classification results through an interactive interface. The implementation environment and deployment-related interface of the developed system are illustrated in Fig. 10.

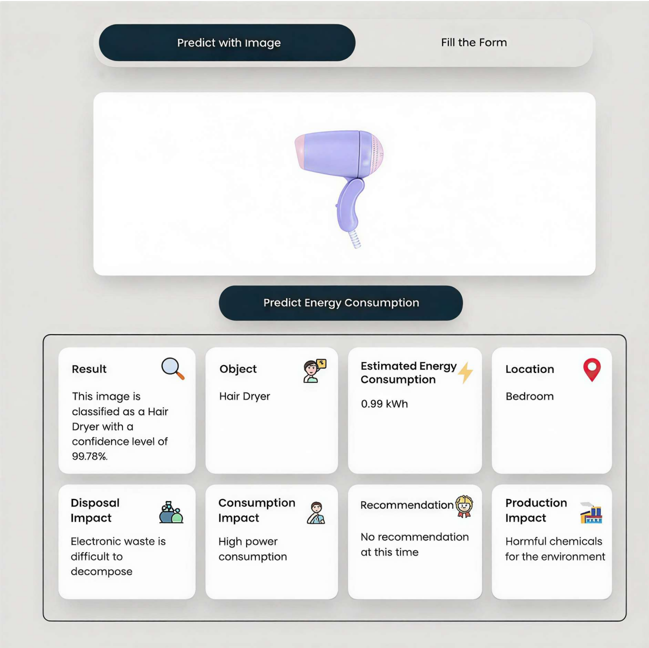


Fig. 10. Prototype system interface showing deployment and user interaction environment.

No.	Picture	Appliance	Energy Consumption
1.		Hair Dryer	0.99 KWh
2.		Air Conditioner	1.77 KWh
3.		Television	0.06 KWh
4.		Washing Machine	4.94 Wh
5.		Laptop	16.38 Wh
5.		Laptop	16.38 Wh

Fig. 11. Example output of the proposed system showing appliance classification and contextual information.

As shown in Fig. 10, the system has been implemented as a functional prototype that supports user interaction and submission of required information for operational use. This implementation demonstrates that the proposed model is not only evaluated in an experimental environment but is also feasible for integration into a real-world application framework. The availability of a functional interface supports the practicality and applicability of the proposed approach. It is important to note that this system does not directly measure or estimate actual energy consumption. Instead, it provides contextual information based on appliance categories to support user awareness. An example of the system output generated from a real appliance image is presented in Fig. 11.

Fig. 11 illustrates the system behavior when a user submits an appliance image. The system successfully identifies the object as a hair dryer and provides the corresponding predicted class. In addition, the interface displays contextual energy-related information associated with the recognized appliance category, such as typical power ranges and general usage characteristics, to support user awareness. The displayed energy-related information is derived from predefined appliance characteristics and typical power consumption ranges obtained from standard references, rather than from direct measurement. It is important to emphasize that this information is provided as an illustrative, user-awareness feature and does not represent actual or real-time energy consumption estimation. This prototype does not sense, measure, or predict real power usage from images. Instead, it demonstrates how visual appliance recognition can be embedded into a broader, user-friendly application that supports interpretability and practical deployment.

V. Conclusion

This study presented a comparative evaluation of deep transfer learning models for household appliance image classification as a foundational step toward vision-based energy-aware systems. Experiments conducted on a mixed-source dataset of 1500 images across eleven appliance categories demonstrate that DenseNet121 achieves the most consistent performance, with an accuracy of 97.0%, validation accuracy of 96.0%, and an F1-score of 98.4%, outperforming EfficientNet-B0 and ResNet50 under identical experimental conditions. Furthermore, 5-fold cross-validation yields a mean accuracy of 97.0% ± 0.29%, indicating strong stability and reliable generalization. The results confirm that transfer learning significantly improves model performance compared with training from scratch, achieving approximately 85% accuracy and exhibiting less overfitting. Rather than directly estimating energy consumption, this work establishes a robust visual recognition framework that serves as a technical foundation for future intelligent energy-monitoring applications. Despite these promising results, several limitations remain. The dataset is still moderate in size, and the model has not yet been extensively evaluated under more complex real-world conditions, such as extreme lighting variations or highly cluttered environments. In addition, the current system focuses on appliance identification rather than direct power consumption measurement. Therefore, future work should expand the dataset, evaluate performance across a broader range of environments, and integrate visual recognition with sensor-based or smart meter data to enable more comprehensive household energy monitoring systems.

Exploring lightweight architectures for deployment on edge devices is also a promising direction for real-time implementation. In addition, future work may explore other variants of EfficientNet (e.g., B1–B4) to further analyze performance–efficiency trade-offs.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Fuzy Yustika Manik conceptualized the study, developed the research framework, supervised the research activities, analyzed and interpreted the results, and led the writing of the manuscript. Anandhini Medianty Nababan contributed to methodology development, provided scientific and technical feedback, and critically revised the manuscript. T. M. Rezha Taufiqurrahman assisted with dataset preparation, software implementation, experimental execution, and data processing. All authors contributed to the discussion of results, reviewed the manuscript, and approved the final version for publication.

Funding

This research was financially supported by the Universitas Sumatera Utara through the TALENTA government collaborative research scheme for the fiscal year 2024. (Grant number: 18589/UN5.1.R/PPM/2024, development of an automatic detection system for household energy consumption using deep learning methods to support sustainable energy).

Acknowledgment

The author would also like to thank the village head and all the heads of the Binjai Timur sub-district who were involved in this study.

References

[1] R. Hischier *et al.*, "Environmental impacts of household appliances in Europe and scenarios for their impact reduction," *Journal of Cleaner Production*, vol. 267, 121952, 2020. doi: 10.1016/j.jclepro.2020.121952

[2] S. K. Firth, K. J. Lomas, and A. J. Wright, "Targeting household energy-efficiency measures using sensitivity analysis," *Building Research & Information*, vol. 38, no. 1, pp. 25–41, 2010. doi: 10.1080/09613210903236706

[3] M. Arun *et al.*, "Internet of things and deep learning-enhanced monitoring for energy efficiency in older buildings," *Case studies in Thermal Engineering*, vol. 61, 104867, 2024. doi: 10.1016/j.csite.2024.104867

[4] A. Alzoubi, "Machine learning for intelligent energy consumption in smart homes," *International Journal of Computations, Information and Manufacturing (IJCIM)*, vol. 2, no. 1, 2022. doi: 10.54489/ijcim.v2i1.75

[5] J. Kelly and W. Knottenbelt, "Neural nilm: Deep neural networks applied to energy disaggregation," in *Proc. 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, 2025, pp. 55–64. doi: 10.1145/2821650.2821672

[6] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 8, pp. 1819–1835, 2009. doi: 10.1016/j.rser.2008.09.033

[7] J. J. M. Escobar *et al.*, "A comprehensive review on smart grids: Challenges and opportunities," *Sensors*, vol. 21, no. 21, 6978, 2021. doi: 10.3390/s21216978

[8] S. Liu, Z. Xie, and Z. Hu, "Research on distributed smart home energy management strategies based on non-intrusive load monitoring (NILM)," *Electronics*, vol. 14, no. 9, 1719, 2025. doi: 10.3390/electronics14091719

[9] M. A. Hossain and M. S. A. Sajib, "Classification of images using Convolutional Neural Network (CNN)," *Global Journal of Computer Science and Technology*, vol. 19, no. 2, 2019. doi: 10.34257/GJCSTDVOL19IS2PG13

[10] B. B. Traore *et al.*, "Deep convolution neural network for image recognition," *Ecological Informatics*, vol. 48, pp. 257–268, 2018. doi: 10.1016/j.ecoinf.2018.10.002

[11] A. Bravo-Diaz, S. Moreno, and J. Lopatin, "Evaluating the ability of convolutional neural networks for transfer learning in Pinus radiata cover predictions," *Ecological Informatics*, vol. 82, 102684, 2024. doi: 10.1016/j.ecoinf.2024.102684

[12] N. A. Baghdadi *et al.*, "An automated diagnosis and classification of COVID-19 from chest CT images using a transfer learning-based convolutional neural network," *Computers in Biology and Medicine*, vol. 144, 105383, 2022. doi: 10.1016/j.compbiomed.2022.105383

[13] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021. doi: 10.1109/JPROC.2020.3004555

[14] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019. doi: 10.1186/s40537-019-0197-0

[15] K. A. A. Mahmoud *et al.*, "Transfer learning by fine-tuning pre-trained convolutional neural network architectures for switchgear fault detection using thermal imaging," *Alexandria Engineering Journal*, vol. 103, pp. 327–342, 2024. doi: 10.1016/j.aej.2024.05.102

[16] T. Le *et al.*, "Improving electric energy consumption prediction using CNN and Bi-LSTM," *Applied Sciences*, vol. 9, no. 20, 4237, 2019. doi: 10.3390/app9204237

[17] A. V. Jonnalagadda, H. A. Hashim, and A. Harris, "Comprehensive and comparative analysis between transfer learning and custom built VGG and CNN-SVM models for wildfire detection," in *Proc. 6th International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2024, pp. 1–7.

[18] K. Okazawa *et al.*, "Evaluation of deep learning-based non-intrusive thermal load monitoring," *Energies*, vol. 17, no. 9, 2012, 2024. doi: 10.3390/en17092012

[19] M. M. Manca and L. Massidda, "Deep learning based non-intrusive load monitoring with low resolution data from smart meters," *Communications in Applied and Industrial Mathematics*, vol. 13, no. 1, pp. 39–56, 2022. doi: 10.2478/caim-2022-0004

[20] T. Gan *et al.*, "Non-intrusive load monitoring based on self-attention," *Journal of Physics: Conference Series*, vol. 2221, no. 1, 012038, 2022. doi: 10.1088/1742-6596/2221/1/012038

[21] Y. Liu *et al.*, "Non-intrusive load monitoring based on unsupervised optimization enhanced neural network deep learning," *Frontiers in Energy Research*, vol. 9, 718916, 2021. doi: 10.3389/fenrg.2021.718916

[22] A. Ullah *et al.*, "Deep learning assisted buildings energy consumption profiling using smart meter data," *Sensors*, vol. 20, no. 3, 873, 2020. doi: 10.3390/s20030873

[23] S. Mari *et al.*, "An embedded deep learning NILM system: A year-long field study in real houses," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–5, 2023. doi: 10.1109/TIM.2023.3328085

[24] M. K. Akbar, M. Amayri, and N. Bouguila, "Deep learning based solution for appliance operational state detection and power estimation in non-intrusive load monitoring," in *Proc. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2023, pp. 59–65. doi: 10.1007/978-3-031-36822-6_5

[25] T. Y. Kim and S. B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, 2019. doi: 10.1016/j.energy.2019.05.230

[26] Z. Yue *et al.*, "Bert4nilm: A bidirectional transformer model for non-intrusive load monitoring," in *Proc. the 5th International Workshop on Non-Intrusive Load Monitoring*, 2020, pp. 89–93. 2020.https://doi.org/10.1145/3427771.3429390

[27] İ. H. Çavdar and V. Feryad, "Efficient design of energy disaggregation model with bert-nilm trained by adax optimization method for smart grid," *Energies*, vol. 14, 4649, 2021. doi: 10.3390/en14154649

[28] I. Laouali *et al.*, "Non-intrusive load monitoring of household devices using a hybrid deep learning model through convex hull-based data selection," *Energies*, vol. 15, no. 3, 1215, 2022. doi: 10.3390/en15031215

[29] M. Trigka and E. Dritsas, "A comprehensive survey of deep learning approaches in image processing," *Sensors*, vol. 25, no. 2, 531, 2025. doi:10.3390/s25020531

[30] C. Gu and M. Lee, "Deep transfer learning using real-world image features for medical image classification, with a case study on pneumonia X-ray images," *Bioengineering*, vol. 11, no. 4, 406, 2024. doi:10.3390/bioengineering11040406

[31] M. H. Yazdani, M. M. Azad, S. Khalid, and H. S. Kim, "A hybrid deep transfer learning framework for delamination identification in composite laminates," *Sensors*, vol. 25, no. 3, 826, 2025. doi:10.3390/s25030826
[32] Z. Liu, S. Jiang, S. Feng *et al.*, "Comprehensive review of deep learning approaches for single-image super-resolution," *Sensors*, vol. 25, no. 18, 5768, 2025. doi: 10.3390/s25185768
[33] M. Jakubec *et al.*, "Comparison of CNN-based models for pothole detection in real-world adverse conditions: overview and evaluation," *Applied Sciences*, vol. 13, no. 9, 5810, 2023. https://doi.org/10.3390/app13095810
[34] Y. Chen, Z. Yan, and Y. Zhu, "A comprehensive survey for generative data augmentation," *Neurocomputing*, vol. 600, 128167, 2024. doi: 10.1016/j.neucom.2024.128167
[35] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: A review," *Journal of Imaging*, vol. 9, no. 4, 81, 2023. doi: 10.3390/jimaging9040081
[36] H. L. Vu *et al.*, "Analysis of input set characteristics and variances on k-fold cross validation for a recurrent neural network model on waste disposal rate estimation," *Journal of Environmental Management*, vol. 311, 114869, 2022. doi: 10.1016/j.jenvman.2022.114869
[37] J. Allgaier and R. Pryss, "Cross-validation visualized: A narrative guide to advanced methods," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 1378–1388, 2024. doi: 10.3390/make6020065