Malaysian Journal of Computing, 10 (1): 2012-2025, 2025
Copyright © UiTM Press
eISSN: 2600-8238

EXPLORING TRANSFER LEARNING AND CONVOLUTIONAL AUTOENCODER FOR EFFECTIVE KITCHEN UTENSILS CLASSIFICATION

Hashim Rosli[1], Rozniza Ali[2*], Muhamad Suzuri Hitam[3], Ashanira Mat Deris[4] and Noor Hafhizah Abd Rahim[5]

[1,2*,3,4,5]*Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, 21030 Terengganu*

[1]p5812@pps.umt.edu.my, [2*]rozniza@umt.edu.my,
[3]suzuri@umt.edu.my, [4]ashanira@umt.edu.my,
[5]noorhafhizah@umt.edu.my

ABSTRACT

Effective classification of kitchen utensils is crucial for advancing assistive technologies and enhancing daily living for individuals with visual impairments. This study investigates the use of transfer learning and convolutional autoencoders to improve classification accuracy. We integrate pre-trained networks into an autoencoder framework to enhance feature extraction and image reconstruction. Models including ResNet50, DenseNet121, and their autoencoder variants were evaluated using precision, recall, accuracy, Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR). Results show that DenseNet121 outperforms ResNet50 with a classification accuracy of 72% and shorter training time. When combined with autoencoders, DenseNet121-Autoencoder achieves the highest classification accuracy of 76% and superior image reconstruction quality, as indicated by higher SSIM and PSNR scores. This improvement highlights DenseNet121's effectiveness in handling complex, high-dimensional classification tasks and noise reduction. The study underscores the model's potential for enhancing assistive technologies and sustainable learning by providing more accurate and reliable object recognition. This advancement supports greater independence for visually impaired users and promotes more inclusive learning environments.

Keywords*: Classification, Convolutional Autoencoder, Deep Learning, Images, Kitchen Utensils, Transfer Learning.*

Received for review: 03-01-2025; Accepted: 18-03-2025; Published: 01-04-2025
DOI: 10.24191/mjoc.v10i1.4533

1. Introduction

The accurate classification of kitchen utensils is a critical area of research within computer vision and artificial intelligence (AI). This study focuses on leveraging advanced AI techniques, including transfer learning and convolutional autoencoders, to develop a robust system for classifying a diverse array of kitchen utensils. Transfer learning allows the model to utilize pre-trained networks to enhance classification performance, while convolutional autoencoders are employed for both feature extraction and noise reduction. The role of autoencoders in reducing image noise is crucial, as it enhances the system's ability to differentiate between similar objects by providing cleaner and more accurate representations.

This capability is essential for precise identification and categorization, which can significantly impact the development of assistive technologies, particularly for individuals with visual impairments. By improving the accuracy of utensil classification, the functionality of assistive tools can be enhanced, supporting users in performing daily living tasks with greater independence and safety (Muhsin et al., 2024; Okolo et al., 2024). Effective classification systems can thus provide meaningful improvements in quality of life for visually impaired individuals by facilitating more reliable interaction with everyday objects (Aung et al., 2024).

This research specifically targets the refinement of a crucial component within AI-integrated assistive technology which is object recognition. Rather than developing an entire assistive technology system, the focus is on advancing a key element like accurate object recognition and classification. The integration of autoencoders to reduce noise and enhance feature extraction aims to improve object recognition systems, which are essential for the development of assistive technologies that offer precise and reliable support in real-world settings. For visually impaired users, AI-driven devices that can effectively identify and provide contextual information about kitchen utensils can lead to more adaptive and practical assistance. This enhanced support fosters greater autonomy, enabling users to perform tasks with minimal external assistance and reducing their reliance on caregivers or other forms of help (Dang et al., 2024; Okolo et al., 2024; Zafar et al., 2022). By refining object recognition capabilities through advanced techniques like noise reduction, this research aims to contribute to the broader goal of creating more effective and user-centered assistive technologies.

The implications of this research extend beyond the immediate scope of assistive applications to broader areas of sustainable living and practical skills development. Accurate AI-driven object recognition can integrate technology into various aspects of daily life, promoting sustainable practices by supporting individuals in home and vocational settings. The ability to recognize and interact with objects such as kitchen utensils can be part of a broader life skills development process, aiding individuals with disabilities in acquiring practical skills that enhance their daily living capabilities. This integration of AI tools into everyday tasks supports a more inclusive approach to living, fostering independence and adaptability in a variety of settings (Baker et al., 2021). Consequently, the research contributes to a vision of sustainable living that incorporates advanced technology to support practical learning and daily activities.

This study highlights the broader potential of AI-integrated assistive technologies in enhancing the quality of life for people with disabilities. By offering personalized, real-time assistance and improving noise reduction in image classification, these technologies have the potential to bridge gaps between abilities and disabilities, promoting greater independence and social inclusion. The advancements achieved through this research not only advance the field of assistive technology but also align with the broader goals of sustainable education and living. By driving inclusivity and adaptability, AI-powered tools can meet the diverse needs of all individuals, regardless of their physical abilities, thus supporting a vision of a more equitable and supportive environment for everyone.

2. Related Work

The integration of artificial intelligence (AI) into assistive technologies (AT) has the potential to greatly enhance the daily lives of individuals with disabilities, particularly those with visual impairments. Basic kitchen utensil classification plays a vital role in this context, offering a practical solution to one of the everyday challenges faced by visually impaired individuals. Historically, AT solutions for kitchen tasks have relied on tactile markers or audio cues, which, while useful, often lack the precision and adaptability that AI can provide. Recent advancements in computer vision and machine learning have enabled more sophisticated approaches, such as automated classification systems that can recognize and differentiate kitchen utensils with high accuracy (Dang et al., 2024). These systems not only facilitate more independent living but also contribute to broader applications in accessibility technology. By leveraging AI, particularly in classifying and detecting kitchen utensils, there is a significant opportunity to improve the functionality and inclusiveness of AT solutions, making them more responsive to the needs of users in real-world scenarios.

Recent advancements in kitchen utensil classification and detection using artificial intelligence demonstrate significant progress and highlight various methodological outcomes. Rosello et al. (Rosello et al., 2023) introduced Kurcuma, a dataset designed for unsupervised domain adaptation. Their approach achieved notable recognition rates of 76.58% with the EKUD dataset and 89.29% with the AKUD dataset, illustrating the dataset's effectiveness in improving model generalization across different environments. In comparison, Yusro et al. (Yusro et al., 2023) evaluated Faster R-CNN and YOLOv5 for detecting overlapping objects, finding that YOLOv5 achieved an accuracy of 89.12%. This high-performance underscores YOLOv5's suitability for real-time applications, although it has limitations with small or occluded objects. Meanwhile, Karungaru (Karungaru, 2019) leveraged fine-tuning and transfer learning, achieving an impressive recognition accuracy of 95%. This result highlights the potential of transfer learning in enhancing classification accuracy but also points to its dependence on the quality and diversity of the pre-trained models. Sáez-Pérez et al. (Sáez-Pérez et al., 2022) explored domain adaptation in robotics, showing that while adapting models to new environments improves performance, it is still constrained by real-world variability. Collectively, these studies illustrate the advancements in kitchen utensil detection and classification, revealing both significant achievements and ongoing challenges such as model adaptability, handling occlusions, and ensuring robustness across diverse scenarios.

Given the significant advancements in kitchen utensil classification and detection, addressing noise remains a critical challenge. Previous research in this area, including studies by Rosello et al. (Rosello et al., 2023), Yusro et al. (Yusro et al., 2023), and others, primarily focused on standard image conditions and did not specifically tackle the problem of noisy images. This gap highlights the need for methods that can effectively handle real-world issues such as noise. Autoencoders present a promising solution to this problem due to their ability to both denoise and extract features. They work by encoding an image into a lower-dimensional latent space, removing noise while preserving key features, and then decoding it back to an image. Recent research has shown that autoencoders are effective in reconstructing noisy images and enhancing feature extraction (J. He et al., 2018). For instance, a study on image deblurring demonstrated that autoencoder-based methods significantly improved image recognition accuracy by reconstructing cleaner images, with Vision-in-Transformer used as the feature extractor (Kang et al., 2023). Additionally, recent advancements in masked autoencoders have shown their scalability and robustness in learning representations from incomplete or occluded data, which can be advantageous in handling noisy inputs (K. He et al., 2022). The integration of autoencoders with advanced neural network architectures, such as convolutional layers, further enhances their ability to manage complex tasks, making them particularly suitable for classifying noisy kitchen utensil images. Thus, autoencoders offer a robust framework for improving classification accuracy by addressing noise and extracting essential features from images.

3. Methodology

In this study, we propose a methodology for classifying kitchen utensils using an autoencoder architecture that leverages a pre-trained network as the encoder. This approach combines the power of transfer learning with the flexibility of autoencoders to enhance classification accuracy, particularly in scenarios involving noisy and complex images. By utilizing a pre-trained model in the encoder, the autoencoder benefits from advanced feature extraction capabilities, while the decoder reconstructs the input, allowing for both effective noise reduction and precise classification. Figure 1 illustrates the overall methodology process, which begins with dataset preparation, followed by the encoding phase where features are extracted using the pre-trained network. The decoder then reconstructs the images, with the final step involving classification. This process is optimized through the careful tuning of hyperparameters to ensure the model's robustness and accuracy.

The steps of the methodology process are:

1. Data is gained from Edinburgh Kitchen Utensils Database and internet images.
2. Augmentation process ran on the data to gained balance images in all classes.
3. In image processing, each image was resized into 64x64 and normalized. Also, gaussian noise was injected into all these images to simulate disturbance or interference in real-world environments.
4. The data split into training and testing dataset using 70:30 ratio.
5. The training dataset used to train the autoencoder which include to train the classification and image reconstruction at the same time. The autoencoder used ResNet pretrained model with ImageNet weight.
6. The testing dataset are used to test the trained autoencoder.
7. Test classification result is evaluated through accuracy, recall, precision and confusion matrix and the reconstructed images are evaluated using SSIM and PSNR score.
8. Step 5-7 are repeated using DenseNet as pretrained network encoder.

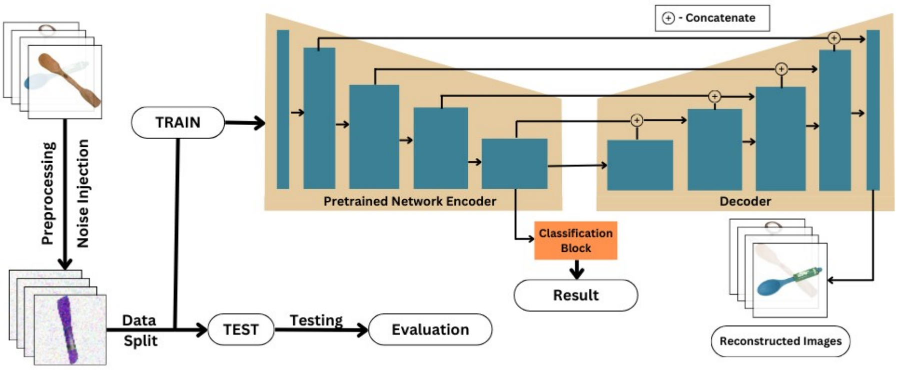


Figure 1. Methodology process of data preparation, training and testing of the Autoencoder

3.1 Dataset Preparation

The dataset used in this study combines the Edinburgh Kitchen Utensils Dataset (EKUD), a comprehensive collection of kitchen utensil images, with additional images sourced from the internet to enhance the dataset's diversity. The EKUD provides a strong foundation due to its structured nature but combining it with internet images ensures a more extensive variety of utensil appearances and conditions, crucial for training a robust classification model. To address the issue of dataset imbalance, where certain classes might be underrepresented, we applied a series of data augmentation techniques using the ImageDataGenerator from Python's Keras library. Specifically, images were flipped horizontally and vertically, rotated at various angles, and translated, simulating different perspectives and orientations of kitchen utensils which generated 100 images per classes in Table 1. This augmentation process not only increases the quantity of training data but also enhances the model's ability to generalize across different real-world scenarios, thereby reducing the risk of overfitting (Alomar et al., 2023).

Table 1. Number of Images before and after augmentation

Classes	EKUD	EKUD and Internet after Augmentation
Bread Knife	24	100
Bottle Opener	30	100
Can Opener	19	100
Dinner Fork	59	100
Dessert Spoon	33	100
Dinner Knife	51	100
Fish Slice	82	100
Ladle	54	100
Kitchen Knife	39	100
Masher	38	100
Potato Peeler	22	100
Pizza Cutter	16	100
Peeler	18	100
Serving Spoon	84	100
Soup Spoon	27	100
Spatula	53	100
Tongs	37	100
Teaspoon	105	100
Whisk	44	100
Wooden Spoon	62	100

Table 2. Noise Injection Parameter

Parameter	
Noise Type	Gaussian
Standard Deviation	30
Mean	0

Preprocessing of the dataset involved resizing all images to 64x64 pixels and normalizing the pixel values to a [0,1] range, ensuring consistency across the dataset and compatibility with the neural network architecture. The images were maintained in RGB format, as retaining all three-color channels is essential for capturing the full spectrum of utensil features, which can be critical for accurate classification (Velastegui et al., 2021). To simulate real-world conditions, Gaussian noise was injected into each image based on the parameter in Table 2. Gaussian noise, characterized by its bell-shaped probability distribution, is commonly used to model the random variations in intensity that occur in natural scenes, making it an ideal choice for simulating disturbances in this context (Miranda-González et al., 2023). The dataset was divided randomly into two portions which are 70% for training and 30% for testing. This allocation provides the model with a significant amount of data for learning while reserving an independent set for testing. This approach is essential to assess the model's performance and ability to generalize to unseen data.

3.2 Autoencoder Architecture

An autoencoder is a neural network designed to compress and reconstruct data efficiently. It works by using an encoder to transform the input into a compact, lower-dimensional form and a decoder to rebuild the original data from this reduced representation. In this study, transfer learning is applied to improve the encoder’s feature extraction capabilities. Pre-trained models like ResNet and DenseNet are used to build on knowledge gained from large datasets such as ImageNet, which are then fine-tuned for smaller, specific datasets. Research has shown that this approach can significantly enhance performance. Sevinc et al. (Sevinc et al., 2022) utilized a pre-trained ResNet50 as the encoder in an autoencoder for brain tumor MRI analysis. Their study showed that transfer learning significantly enhanced feature extraction and reconstruction quality, achieving a classification accuracy of 94.56%. The improved feature representation from the pre-trained model led to better performance in medical image analysis, demonstrating the effectiveness of transfer learning in this domain. Similarly, Prabira et al. (Pan et al., 2021) applied a pre-trained DenseNet201 model in an autoencoder for analyzing remote sensing images. They found that transfer learning with DenseNet201 accelerated image reconstruction and improved accuracy, achieving 92.7% in land cover classification. This result highlights how DenseNet201's deep architecture contributes to more effective feature extraction and faster processing compared to training from scratch.

The AutoCovNet study employed a pre-trained VGG16 model in the encoder for COVID-19 detection by using images of chest X-ray. Transfer learning significantly enhanced feature extraction and image reconstruction, achieving an impressive accuracy of 98.9% in detecting COVID-19. This study underscores the robustness of transfer learning in improving diagnostic performance and feature extraction in medical imaging applications (Rashid et al., 2021). These findings highlight the advantages of employing ResNet and DenseNet for feature extraction within autoencoders. In this study, both ResNet and DenseNet are utilized as pre-trained networks in the encoder phase of the autoencoder architecture. ResNet, introduced features deep residual learning with up to 152 layers that can help with accuracy and computational efficiency in image classification tasks which mentioned by (Das Gupta et al., 2023). It also employing residual connections to address the vanishing gradient problem and enhance model training (Alahmadi et al., 2023; Lippl et al., 2024). DenseNet utilizes dense connections between layers to improve feature propagation and gradient flow, with architectures extending up to 201 layers (Yang et al., 2023; Zhai et al., 2020). The deep and intricate nature of these networks equips them to extract robust features from input images, providing a solid foundation for the autoencoder's performance.

The decoder is responsible for reconstructing images from the encoded features and is composed of seven blocks. Each block consists of a single Conv2DTranspose layer, followed by batch normalization and ReLU activation layers. For the first five blocks, we use 2x2 strides, while the last two blocks employ 1x1 strides. The output from the first four blocks is combined using the Add() function from Keras, implementing skip connections similarly to U-Net architecture of multiple studies from (Latha H N and Rajiv R Sahay, 2020; Thai et al., 2023; Tripathi, 2021) to enhance the quality of the reconstructed images. Each skip connection from the encoder is processed through a Conv2D layer with a kernel size of 1x1 to reduce the number of filters, aligning them with the decoder's output. The final output layer of the decoder uses a Conv2D layer with a linear activation function to produce the reconstructed image. All convolutional layers in the decoder utilize 'same' padding to preserve the spatial dimensions of the feature maps. During training, the hyperparameter used was based on the Table 3 with the loss functions stated below:

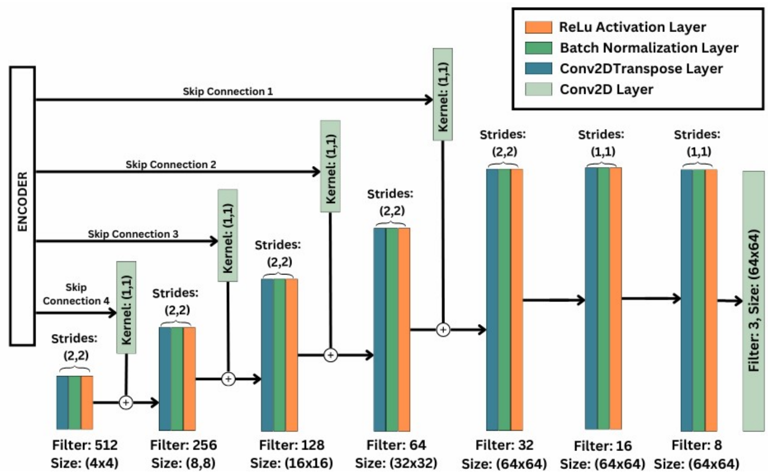


Figure 2. Decoder Layers

Table 3. Hyperparameter of Autoencoder Training

Hyperparameter	
Epochs	100
Batch Size	32
Optimizer	Adam
Learning Rate	0.0001

In our approach, we utilize Mean Squared Error (MSE) loss function, L_{MSE} in Equation (1) and Sparse Categorical Crossentropy (SCC) loss functions, L_{SCC} in Equation (2) from Keras TensorFlow to optimize both image reconstruction and classification tasks. This approach utilizing multi task learning based on some studies from (Chen et al., 2022; Xiang et al., 2024) which uses hybrid loss and in this case, reconstructions and classification loss. MSE loss measures the average squared difference between the predicted images as $\hat{x}$ and true images as x, emphasizing larger reconstruction errors by the number of pixels in the images, N. This makes it a robust metric for evaluating reconstruction quality. On the other hand, SCC loss is used for classification tasks where labels are integers, effectively handling any number of classes, C just by using sparse labels to produce the classification results. To address both tasks simultaneously, we combine these loss function which shown in Equation (3) with weights to balance their contributions, ensuring the model performs well in both reconstruction and classification. The combined loss function integrates MSE and SCC, with weights α and β set to 0.5 and 0.5 respectively, to optimize the model for accurate image reconstruction and effective classification. The losses equation that used in Keras Tensorflow and combined loss are:

$$L_{MSE}(\hat{x},x) = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 \tag{1}$$

$$L_{SCC}(\hat{y},y) = \sum_{i=1}^{C} y_i \log\,(\hat{y}_i) \tag{2}$$

$$L = \alpha\,\cdot L_{MSE}(\hat{x},x) + \,\beta\cdot L_{SCC}(\hat{y},y) \tag{3}$$

where,

x – Ground truth image

y – Ground truth label

N – Number of pixels in an image

$\hat{x}$ – Reconstructed (predicted) image

$\hat{y}$ – Predicted probability distribution

C – Number of classes

Lastly, the classification block, which is crucial for translating the encoded features into class predictions, incorporates a single max pooling layer. This layer reduces the spatial dimensions of the encoded features, effectively capturing the most prominent patterns and making the model more computationally efficient. The pooled features are then passed through a dense layer, where a softmax activation function is applied to generate probabilities for each of the 20 distinct classes. This final structure not only ensures that the model can generalize well across various input images but also enhances its accuracy in multi-class classification by emphasizing the most relevant features during the decision-making process. The combination of max pooling and the softmax-activated dense layer enables the model to effectively differentiate between the various classes, leading to more precise and reliable classifications.

4. Results and Discussion

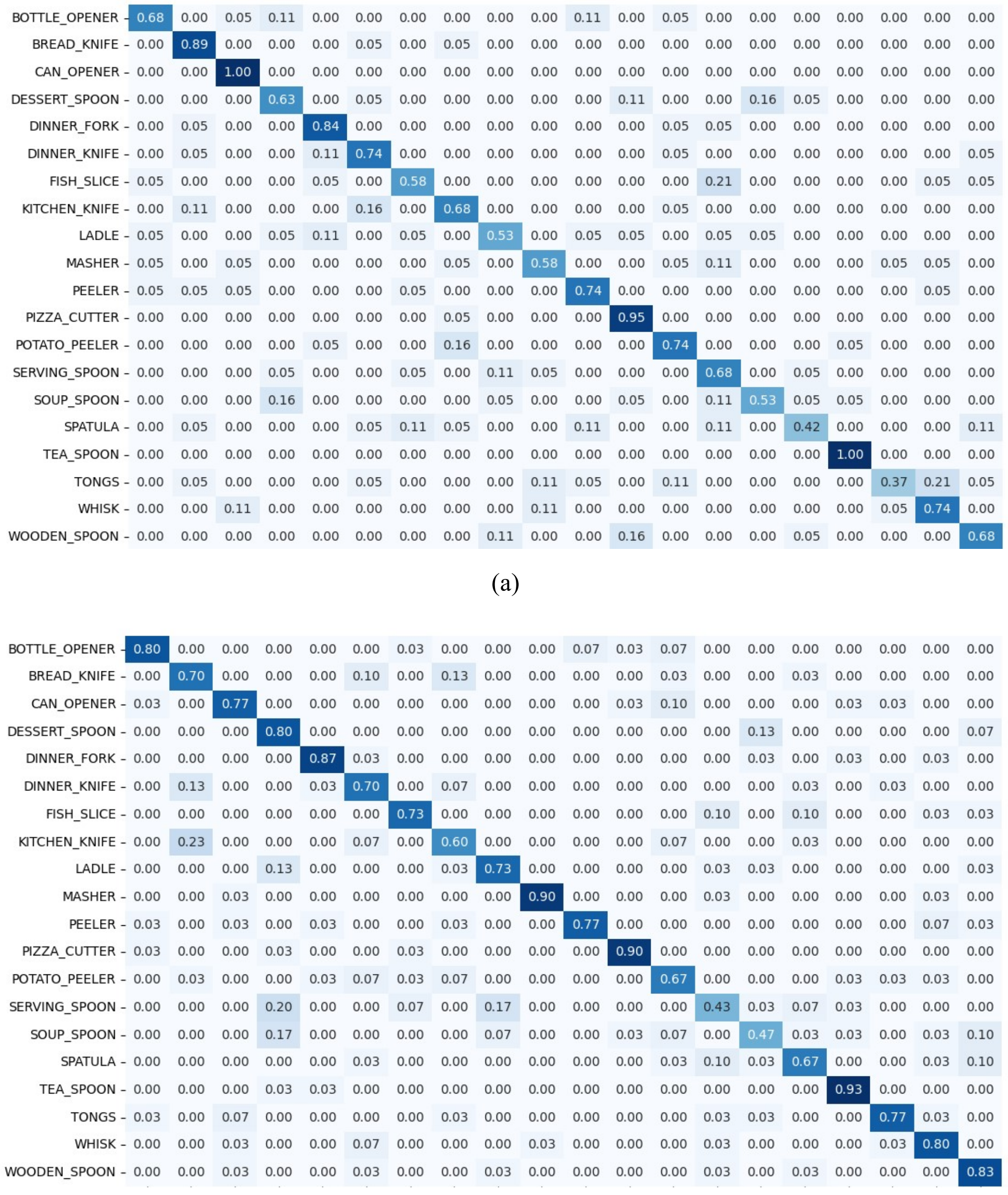
The experiment in this study involved training and testing various models, including ResNet50, DenseNet121, ResNet50-Autoencoder, and DenseNet121-Autoencoder. The autoencoders were designed based on the proposed method, utilizing pretrained networks as encoders and employing the same decoder structure outlined in Figure 2. All models were trained using identical hyperparameters and datasets to ensure consistency. The evaluation of these models was conducted based on precision, recall, accuracy, and training time. Additionally, for the autoencoder models, we assessed the quality of reconstructed images using Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR). Evaluating the reconstructed images is crucial because it reflects how well the autoencoder captures and reconstructs key features from the input, which directly impacts the model's ability to learn meaningful representations. Higher SSIM and PSNR scores indicate better reconstruction quality, leading to improved feature learning and, consequently, enhanced performance in classification tasks. The primary focus of the data was on noise reduction and feature learning, aiming to enhance overall model performance.

Table 4. Result performance of image reconstruction and classification

Model		**Image Reconstruction**		**Image Classification**		
Type	***Training Time (h)***	**SSIM**	**PSNR**	**Precision**	**Recall**	**Accuracy**
Resnet50	1.63	-	-	0.65	0.64	64%
DenseNet121	0.94	-	-	0.74	0.72	72%
ResNet50-Autoencoder	2.34	0.6876	73.1514	0.66	0.67	67%
DenseNet121-Autoencoder	1.87	0.7371	72.9398	0.76	0.75	76%

The results in Table 4 highlight clear performance differences among the models. DenseNet121, with the shortest training time of 0.94 hours, achieves the highest classification accuracy at 72%, compared to ResNet50's accuracy of 64%. This suggests that DenseNet121 is more effective at feature extraction within the given timeframe. When autoencoders are incorporated, both models show improved image reconstruction quality. The DenseNet121-Autoencoder achieves a higher Structural Similarity Index Measure (SSIM) of 0.7371 and a Peak Signal-to-Noise Ratio (PSNR) of 72.94, whereas the ResNet50-Autoencoder has an SSIM of 0.6876 and a PSNR of 73.15, indicating that DenseNet121-Autoencoder better preserves image details. Furthermore, the DenseNet121-Autoencoder also leads in classification accuracy with 76%, whereas the ResNet50-Autoencoder achieves 67%. This model also demonstrates superior precision and recall. Overall, these findings indicate that DenseNet121 combined with an autoencoder provides superior performance in both image reconstruction and classification tasks.

The confusion matrix results for the ResNet50-Autoencoder and DenseNet121-Autoencoder models, shown in Figure 3, reveal significant differences in classification performance. The ResNet50-Autoencoder often misclassifies items such as 'FISH_SLICE' as 'SERVING_SPOON' and TONGS' as WHISK.' This model struggles with categories that share similar features or are less represented in the dataset, leading to lower overall accuracy and precision. In contrast, the DenseNet121-Autoencoder exhibits more refined performance. It demonstrates fewer misclassifications and better accuracy, particularly for categories like 'CAN_OPENER' and 'MASHER,' which it identifies more correctly. The DenseNet121-Autoencoder is more effective at distinguishing between similar or overlapping categories, such as 'DINNER_FORK' and 'DINNER_KNIFE,' showing improved handling of these similar items. Additionally, it performs better with less frequent classes like 'WOODEN_SPOON,' resulting in fewer errors. Overall, the DenseNet121-Autoencoder's confusion matrix reflects a more balanced and accurate classification performance, indicating that its architecture and training are better suited for managing the complexities of distinguishing between similar kitchen utensils. The DenseNet121-Autoencoder produces reconstructions that more closely match the original images compared to the ResNet50-Autoencoder in Figure 4. While both autoencoders reduce noise, the DenseNet121-Autoencoder leaves fewer artifacts and provides cleaner, more accurate images. In contrast, the ResNet50-Autoencoder often results in more visible artifacts, indicating less effective noise reduction and detail preservation. This suggests that the DenseNet121-Autoencoder is superior in maintaining image fidelity.



BOTTLE_OPENER	0.68	0.00	0.05	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BREAD_KNIFE	0.00	0.89	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CAN_OPENER	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DESSERT_SPOON	0.00	0.00	0.00	0.63	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.16	0.05	0.00	0.00	0.00	0.00
DINNER_FORK	0.00	0.05	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00
DINNER_KNIFE	0.00	0.05	0.00	0.00	0.11	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.05
FISH_SLICE	0.05	0.00	0.00	0.00	0.05	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.05	0.05
KITCHEN_KNIFE	0.00	0.11	0.00	0.00	0.00	0.16	0.00	0.68	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LADLE	0.05	0.00	0.00	0.05	0.11	0.00	0.05	0.00	0.53	0.00	0.05	0.05	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00
MASHER	0.05	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.58	0.00	0.00	0.05	0.11	0.00	0.00	0.00	0.05	0.05	0.00
PEELER	0.05	0.05	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00
PIZZA_CUTTER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
POTATO_PEELER	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.05	0.00	0.00	0.00
SERVING_SPOON	0.00	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.11	0.05	0.00	0.00	0.00	0.68	0.00	0.05	0.00	0.00	0.00	0.00
SOUP_SPOON	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.11	0.53	0.05	0.05	0.00	0.00	0.00
SPATULA	0.00	0.05	0.00	0.00	0.00	0.05	0.11	0.05	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.42	0.00	0.00	0.00	0.11
TEA_SPOON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
TONGS	0.00	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.11	0.05	0.00	0.11	0.00	0.00	0.00	0.00	0.37	0.21	0.05
WHISK	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.74	0.00
WOODEN_SPOON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.16	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.68

(a)

BOTTLE_OPENER	0.80	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.07	0.03	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BREAD_KNIFE	0.00	0.70	0.00	0.00	0.00	0.10	0.00	0.13	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00
CAN_OPENER	0.03	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.10	0.00	0.00	0.00	0.03	0.03	0.00	0.00
DESSERT_SPOON	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.07
DINNER_FORK	0.00	0.00	0.00	0.00	0.87	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.03	0.00
DINNER_KNIFE	0.00	0.13	0.00	0.00	0.03	0.70	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.00
FISH_SLICE	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.10	0.00	0.00	0.03	0.03
KITCHEN_KNIFE	0.00	0.23	0.00	0.00	0.00	0.07	0.00	0.60	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.03	0.00	0.00	0.00	0.00
LADLE	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.03	0.73	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.03
MASHER	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03	0.00
PEELER	0.03	0.00	0.03	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.03
PIZZA_CUTTER	0.03	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
POTATO_PEELER	0.00	0.03	0.00	0.00	0.03	0.07	0.03	0.07	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.03	0.03	0.03	0.00
SERVING_SPOON	0.00	0.00	0.00	0.20	0.00	0.00	0.07	0.00	0.17	0.00	0.00	0.00	0.00	0.43	0.03	0.07	0.03	0.00	0.00	0.00
SOUP_SPOON	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.03	0.07	0.00	0.47	0.03	0.03	0.00	0.03	0.10
SPATULA	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.10	0.03	0.67	0.00	0.00	0.03	0.10
TEA_SPOON	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00
TONGS	0.03	0.00	0.07	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.77	0.03	0.00
WHISK	0.00	0.00	0.03	0.00	0.00	0.07	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.80	0.00
WOODEN_SPOON	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.83

(b)

Figure 3. Result normalized confusion matrix of (a) ResNet50-Autoencoder (b) DenseNet121-Autoencoder

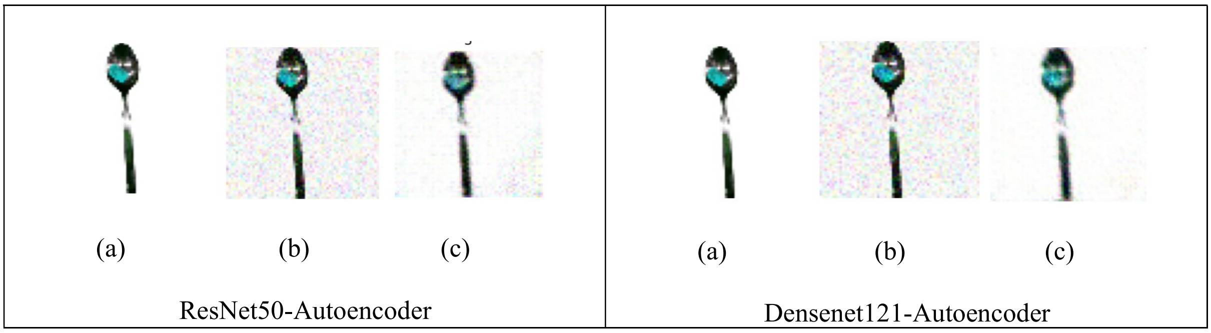


Figure 4. (a) Target images, (b) noise images and (c) reconstructed images

The results of this study underscore the superior performance of DenseNet121 over ResNet50 in the context of kitchen utensil classification. DenseNet121 achieved a classification accuracy of 72% and a training time of 0.94 hours, surpassing ResNet50's 64% accuracy. This advantage is primarily due to DenseNet121's dense connectivity, which enhances feature reuse and gradient flow, enabling the model to capture more detailed features and better differentiate between visually similar objects. This architectural efficiency is crucial in high-dimensional classification tasks, where distinguishing between 20 distinct utensil classes presents significant challenges. The integration of autoencoders into both models led to substantial improvements in image reconstruction quality and noise reduction. DenseNet121-Autoencoder demonstrated superior performance with higher SSIM and PSNR scores compared to ResNet50-Autoencoder. The autoencoder's role in reducing noise and enhancing feature representation was evident, as it allowed DenseNet121 to reconstruct images with greater fidelity and fewer artifacts. This effective noise reduction is critical for accurate object classification, as it improves the model's ability to learn meaningful features from cleaner, more accurate image data. DenseNet121-Autoencoder's enhanced classification accuracy of 76% reflects the benefits of combining robust reconstruction with improved noise handling, which contributes to more reliable feature differentiation.

The observed accuracy in this study, while lower than some existing research, is primarily influenced by the high number of classes involved, which totals 20 in this case. Multi-class classification tasks inherently become more challenging as the number of categories increases. This is because the model must differentiate between a greater variety of classes, some of which may be very similar to each other. As the number of classes grows, the potential for confusion between similar categories also increases, making accurate classification more difficult (Ali et al., 2023; Archana & Jeevaraj, 2024). In high-dimensional classification problems, distinguishing between items becomes more complex, leading to lower accuracy as models struggle with numerous similar categories. Despite this, the DenseNet121-Autoencoder performs well, indicating its architecture handles the complexity effectively. Its higher accuracy compared to other models suggests that its architecture is relatively well-suited for managing the difficulties associated with a large number of classes. To address the limitations posed by high-dimensional classification problems and improve accuracy, future research could explore advanced techniques. For instance, hierarchical classification could be employed to break down the classification task into a series of simpler, nested problems.

The findings of this research have significant implications for both assistive technology and sustainable learning. The DenseNet121-Autoencoder's advanced object recognition and noise reduction capabilities markedly enhance assistive tools for individuals with visual impairments, enabling more accurate identification and classification of kitchen utensils. This improved performance facilitates greater independence and safety by providing clearer, more reliable feedback, allowing users to perform daily tasks with minimal assistance. Moreover, these advancements in AI-driven object recognition contribute to sustainable education by integrating effective technology into practical skills training, promoting a more inclusive and adaptive learning environment for individuals with disabilities. Thus, the DenseNet121-Autoencoder not only improves the functionality of assistive technologies but also supports a more equitable and supportive approach to education and daily living.

5. Conclusion and Future Work

This research highlights the significant progress achieved in utilizing AI models for classifying kitchen utensils, with a focus on enhancing assistive technologies for visually impaired individuals. The comparative analysis of ResNet50, DenseNet121, and their autoencoder variants reveals that the DenseNet121-Autoencoder outperforms its counterparts in both image reconstruction and classification accuracy. The improved performance metrics, including the highest accuracy of 76% and superior SSIM and PSNR scores, demonstrate the efficacy of DenseNet121 in maintaining image detail and enhancing the reliability of utensil classification. These findings underscore the potential of AI-driven solutions to support daily living tasks for visually impaired users, emphasizing the role of advanced deep learning techniques in developing practical, assistive tools. Our study resonates with the work of (Afridi & Sher, 2024) and (Kiruthika Devi & Subalalitha, 2022), as both studies endeavours harness the power of advanced deep learning techniques to aid visually impaired individuals in performing daily tasks. By focusing on object recognition and navigation support, these studies collectively highlight the significant role of AI-driven solutions in enhancing users' independence and quality of life. This alignment further reinforces the relevance and impact of our research in this evolving field.

Building on the current findings, future work will concentrate on advancing object detection capabilities to address more intricate challenges, such as handling overlapping objects and varying noise levels. This involves exploring and implementing more sophisticated algorithms and model architectures that can better manage real-world complexities. Specific efforts will include enhancing noise reduction techniques, improving the robustness of feature extraction methods and developing advanced strategies for distinguishing overlapping objects. Additionally, investigating hybrid or ensemble approaches to integrate multiple models may further refine detection accuracy. These advancements aim to improve the performance of assistive technologies, making them more effective in practical applications and thereby increasing their utility and reliability for visually impaired individuals.

Acknowledgement

This research received financial support from the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS).

Funding Statement

This work was funded under the Fundamental Research Grant Scheme (FRGS), reference no. (FRGS/1/2023/ICT02/UMT/03/2), provided by the Ministry of Higher Education Malaysia.

Author Contribution

In this study, Author 1 led the experimental procedures and manuscript preparation. Author 2, serving as the project leader, also presented the research at a conference. Author 3 co-supervised the project. Authors 4 and 5 assisted with report formatting and conducted similarity checks to ensure originality.

Conflicts of Interest

The authors state that they have no conflicts of interest to disclose in relation to this study.

References

Afridi, Y. S., & Sher, M. (2024). *Visually: Assisting the Visually Impaired People Through AI-Assisted Mobility*. *6*(5), 9–17. https://www.researchgate.net/publication/382397849

Alahmadi, T. J., Rahman, A. U., Alkahtani, H. K., & Kholidy, H. (2023). Enhancing Object Detection for VIPs Using YOLOv4_Resnet101 and Text-to-Speech Conversion Model. *Multimodal Technologies and Interaction*, *7*(8). https://doi.org/10.3390/mti7080077

Ali, A. K., Abdullah, A. M., & Raheem, S. F. (2023). Impact the Classes' number on the convolutional neural networks performance for image classification. *International Journal of Advanced Science Computing and Engineering*, *5*(2), 119–128. https://doi.org/10.62527/ijasce.5.2.132

Alomar, K., Aysel, H. I., & Cai, X. (2023). Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *Journal of Imaging*, *9*(2). https://doi.org/10.3390/jimaging9020046

Archana, R., & Jeevaraj, P. S. E. (2024). Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, *57*(1). https://doi.org/10.1007/s10462-023-10631-z

Aung, M. M., Maneetham, D., Crisnapati, P. N., & Thwe, Y. (2024). Enhancing Object Recognition for Visually Impaired Individuals using Computer Vision. *International Journal of Engineering Trends and Technology*, *72*(4), 297–305. https://doi.org/10.14445/22315381/IJETT-V72I4P130

Baker, K., Parekh, A., Fabre, A., Addlesee, A., Kruiper, R., & Lemon, O. (2021). The Spoon is in the Sink: Assisting Visually Impaired People in the Kitchen. *ReInAct 2021 - Proceedings of the Conference on Reasoning and Interaction*, 32–39.

Chen, L., Saykin, A. J., Yao, B., & Zhao, F. (2022). Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood. *Computational and Structural Biotechnology Journal*, *20*, 5761–5774. https://doi.org/10.1016/j.csbj.2022.10.016

Dang, B., Ma, D., Li, S., Dong, X., Zang, H., & Ding, R. (2024). Enhancing Kitchen Independence: Deep Learning-Based Object Detection for Visually Impaired Assistance. *Academic Journal of Science and Technology*, *9*(2), 180–184. https://doi.org/10.54097/hc3f1045

Das Gupta, N., Rajoo, R., & Jacob, P. J. (2023). Driver Drowsiness Detection System Through Facial Expression Using Convolutional Neural Networks (Cnn). *Malaysian Journal of Computing*, *8*(1), 1375–1387.

He, J., Liu, L., Zhang, C., Zhao, K., Sun, J., & Li, P. (2018). Deep denoising autoencoding method for feature extraction and recognition of vehicle adhesion status. *Journal of Sensors*, *2018*. https://doi.org/10.1155/2018/5419645

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR52688.2022.01553

Kang, Q., Gao, J., Li, K., & Lao, Q. (2023). Deblurring Masked Autoencoder Is Better Recipe for Ultrasound Image Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-031-43907-0_34

Karungaru, S. (2019). Kitchen utensils recognition using fine tuning and transfer learning. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3376067.3376104

Kiruthika Devi, S., & Subalalitha, C. N. (2022). Deep learning based audio assistive system for visually impaired people. *Computers, Materials and Continua*, *71*(1), 1205–1219. https://doi.org/10.32604/cmc.2022.020827

Latha H N and Rajiv R Sahay. (2020). A Local Modified U-net Architecture for Image Denoising. *International Journal for Modern Trends in Science and Technology*, *6*(8S), 140–144. https://doi.org/10.46501/ijmtstciet27

Lippl, S., Peters, B., & Kriegeskorte, N. (2024). Can neural networks benefit from objectives that encourage iterative convergent computations? A case study of ResNets and object classification. *PLoS ONE*, *19*(3 March). https://doi.org/10.1371/journal.pone.0293440

Miranda-González, A. A., Rosales-Silva, A. J., Mújica-Vargas, D., Escamilla-Ambrosio, P. J., Gallegos-Funes, F. J., Vianney-Kinani, J. M., Velázquez-Lozada, E., Pérez-Hernández, L. M., & Lozano-Vázquez, L. V. (2023). Denoising Vanilla Autoencoder for RGB and GS Images with Gaussian Noise. *Entropy*, *25*(10). https://doi.org/10.3390/e25101467

Muhsin, Z. J., Qahwaji, R., Ghanchi, F., & Al-Taee, M. (2024). Review of substitutive assistive tools and technologies for people with visual impairments: recent advancements and prospects. In *Journal on Multimodal User Interfaces* (Vol. 18, Issue 1, pp. 135–156). https://doi.org/10.1007/s12193-023-00427-4

Okolo, G. I., Althobaiti, T., & Ramzan, N. (2024). Assistive Systems for Visually Impaired Persons: Challenges and Opportunities for Navigation Assistance. *Sensors*, *24*(11). https://doi.org/10.3390/s24113572

Pan, Y., Pi, D., Khan, I. A., Khan, Z. U., Chen, J., & Meng, H. (2021). DenseNetFuse: a study of deep unsupervised DenseNet to infrared and visual image fusion. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10339–10351. https://doi.org/10.1007/s12652-020-02820-3

Rashid, N., Hossain, M. A. F., Ali, M., Islam Sukanya, M., Mahmud, T., & Fattah, S. A. (2021). AutoCovNet: Unsupervised feature learning using autoencoder and feature merging for detection of COVID-19 from chest X-ray images. *Biocybernetics and Biomedical Engineering*, *41*(4), 1685–1701. https://doi.org/10.1016/j.bbe.2021.09.004

Rosello, A., Valero-Mas, J. J., Gallego, A. J., Sáez-Pérez, J., & Calvo-Zaragoza, J. (2023). Kurcuma: a kitchen utensil recognition collection for unsupervised domain adaptation. *Pattern Analysis and Applications*, *26*(4), 1557–1569. https://doi.org/10.1007/s10044-023-01147-x

Sáez-Pérez, J., Gallego, A. J., Valero-Mas, J. J., & Zaragoza, J. C. (2022). Domain Adaptation in Robotics: A Study Case on Kitchen Utensil Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13256 LNCS*, 366–377. https://doi.org/10.1007/978-3-031-04881-4_29

Sevinc, O., Mehrubeoglu, M., Guzel, M. S., & Askerzade, I. (2022). An Effective Medical Image Classification: Transfer Learning Enhanced by Auto Encoder and Classified with SVM. *Traitement Du Signal*, *39*(1), 125–131. https://doi.org/10.18280/ts.390112

Thai, D. H., Fei, X., Le, M. T., Zufle, A., & Wessels, K. (2023). Riesz-Quincunx-UNet Variational Autoencoder for Unsupervised Satellite Image Denoising. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, *61*. https://doi.org/10.1109/TGRS.2023.3291309

Tripathi, M. (2021). Facial image denoising using AutoEncoder and UNET. *Heritage and Sustainable Development*, *3*(2), 89–96. https://doi.org/10.37868/hsd.v3i2.71

Velastegui, R., Yang, L., & Han, D. (2021). The Importance of Color Spaces for Image Classification Using Artificial Neural Networks: A Review. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12950 LNCS*, 70–83. https://doi.org/10.1007/978-3-030-86960-1_6

Xiang, Q., Tang, Y., & Zhou, X. (2024). Multi-task learning with self-learning weight for image denoising. *Journal of Engineering and Applied Science*, *71*(1). https://doi.org/10.1186/s44147-024-00425-7

Yang, L., Chen, G., & Ci, W. (2023). Multiclass objects detection algorithm using DarkNet-53 and DenseNet for intelligent vehicles. *Eurasip Journal on Advances in Signal Processing*, *2023*(1). https://doi.org/10.1186/s13634-023-01045-8

Yusro, M. M., Ali, R., & Hitam, M. S. (2023). Comparison of Faster R-CNN and YOLOv5 for Overlapping Objects Recognition. *Baghdad Science Journal*, *20*(3), 893. https://doi.org/10.21123/bsj.2022.7243

Zafar, S., Asif, M., Ahmad, M. Bin, Ghazal, T. M., Faiz, T., Ahmad, M., & Khan, M. A. (2022). Assistive Devices Analysis for Visually Impaired Persons: A Review on Taxonomy. In *IEEE Access* (Vol. 10, pp. 13354–13366). https://doi.org/10.1109/ACCESS.2022.3146728

Zhai, S., Shang, D., Wang, S., & Dong, S. (2020). DF-SSD: An Improved SSD Object Detection Algorithm Based on DenseNet and Feature Fusion. *IEEE Access*, *8*, 24344–24357. https://doi.org/10.1109/ACCESS.2020.2971026