BJET
(Bayero Journal of Engineering and Technology)
https://bjet.ng/index.php/jet

Research Article
2026, Vol. 21 (1), 46-54

AN IN-DEPTH COMPARISON OF PLAIN CNN, FINE-TUNE VGG16, AND VISION TRANSFORMER MODELS IN OBJECT DETECTION

N. H. Adamu[1*], A. T. Balarabe[2]

[1,2] Department of Computer Science, Faculty of Computing, Sokoto State University, Sokoto, P.M.B. 2134, Sokoto, Nigeria

[*]Corresponding author email: najibhassanadamu@gmail.com

Received: 15-11-2025
Revised: 01-12-2025
Accepted: 07-01-2026
Published: 11-01-2026

Abstract: *Object detection is one of the most essential yet challenging tasks in computer vision, playing a crucial role in applications such as tumour detection, quality control and inventory management, security and surveillance, autonomous systems and robotics, crop monitoring and disease detection in agriculture, defect detection in the construction industry and items detection for home robots. With the rise of deep learning techniques, significant progress has been made through models like convolutional neural networks (CNNs) and, more recently, vision transformers (ViTs). In this research, we investigate how three object detection architectures- plain CNN, transfer learning, and ViTs- perform in detecting multiple household objects under controlled conditions. The findings show that ViTs consistently outperformed plain CNN and transfer learning in key performance areas. While the plain CNN achieved a peak IoU of 97.10%, VGG16 reached 98.01%, and the ViTs model attained the highest IoU of 98. 42%, with a smoother and more stable learning curve. Additionally, the mean square error (MSE) was lower for ViTs at 5. 5.197%, compared to the plain CNN and the fine-tune VGG 16, which settled at 11. 40% and 10. 21%, indicating better prediction precision. Loss metrics for ViTs were also consistently lower, decreasing to 5. 88% compared to plain CNN and VGG16, which settled at 11.23% and 10. 01%, demonstrating more efficient learning with less fluctuation during training. However, this came at the expense of increased training time. The VGG16 required approximately 1, 326 seconds per epoch, compared to about 147 and 150 seconds per epoch for plain CNN and the ViTs. By comparing the three models in terms of IoU, MSE, loss behaviour, and execution time, this research highlights the growing strength of transformer-based models in object detection tasks. These results not only reinforce the potential of ViTs but also offer valuable insights for researchers and practitioners aiming to balance performance with computational costs in real-world detection tasks.*

Key words: *Computer Vision (CV); Deep Learning (DL), CNNs (Convolutional Neural Networks); You Only Look Once (YOLO); Vision Transformers (ViTs); VGG (Visual Geometry Group); MSE (Mean Squared Error), IoU (Intersection over Union)*

1 Introduction

The primary purpose of an object detection model is to identify the presence of an object within an image and assign a class label to it by using a bounding box (Hassan & Tukur, 2025). Computers see and process images in different ways (Balarabe & Jordanov, 2022; Hassan & Tukur, 2025). In image classification, an entire image is assigned a label to determine its category (Balarabe & Jordanov, 2021, 2024). In semantic segmentation, however, different objects are given different colour pixels, while similar objects are labelled using the same colour. The emergence of convolutional neural networks has paved the way for the exponential growth of object detection. From robotics, autonomous vehicles, medical imaging and surveillance, object detection continues to be applied to solve problems in many domains. With the growing traction towards this field, several techniques have been proposed by researchers with the view to

addressing some shortcomings of the existing methodologies. CNNs used to be the most reliable techniques for object detection (Wang et al., 2022). However, as more application areas emerge, researchers have continued to develop cutting-edge techniques for detecting objects in images (Hassan & Tukur, 2025). The most recent among such architectures is (ViTs) (Reedha et al., 2022). Vision transformers offer several advantages compared to CNNs, although they rely heavily on voluminous data (Hassan & Tukur, 2025). CNNs models, on the other hand, have been the architecture of choice for object detection tasks, including in-housewares, where there is a limited dataset and constrained computational resources (Hassan & Tukur, 2025). The transfer learning models, which are pre-trained CNNs, have proven to be reliable for practical object tasks, including housewares detection (Jakhar and Kaur, 2020).

Despite the need for efficient methods to segment, classify, detect and identify objects to secure sensitive information, there are still challenges that should be addressed (Hassan & Tukur, 2025; Maurício et al., 2023). Each model, from the earlier ones to the most recent ones, has some strengths and weaknesses. The earlier object detection methods used the sliding window approach, where classifiers are applied to the contents of a window at each spatial location (Hassan & Tukur, 2025). As an improvement to the earlier detectors, the later techniques abandoned the sliding windows, reduced the search space by first determining the region proposals, and applied advanced classifiers (Hassan & Tukur, 2025). Deep learning-based detectors, such as R-CNN and Fast R-CNN, were introduced to address the issues related to the speed at which objects are detected (Hassan & Tukur, 2025). In addition to these proposal-based detectors, Faster R-CNN was developed. Unlike the R-CNN and Fast R-CNN, the Faster R-CNN brought an assurance of increased detection speed by introducing the region proposal network (RPN) (Hassan & Tukur, 2025). Mask R-CNN was later introduced for the task of semantic segmentation. While insufficient data affects a model's performance during and after its training, large amounts of data, on the other hand, require advanced hardware to maintain a desired level of speed and accuracy (Hassan & Tukur, 2025). Hence, the introduction of ViT, which has the benefits of long-range relationship, parallelised processing, less bias, and the ability to process humungous amounts of data (Hassan & Tukur, 2025). The goal of this research is to explore the performance of these architectures, CNNs, fine-tuned models and ViTs, in a low computational power environment under the constraint of a small sample dataset. In addition, the study aims to compare at least three object detection techniques, the plain CNNs, fine-tuned VGG16 and ViTs, to identify the best of them in terms of performance using small sample image datasets. These goals will be achieved through exploring innovative training methods that can enhance model performance without requiring extensive datasets, focusing more on hyperparameter optimisation, developing and evaluating new methodologies that leverage architectures like CNNs and ViTs to improve generalisation in limited data scenarios and comparing the performance of CNNs, fine-tuned models, and the ViTs architecture on the same small sample datasets. The outcomes of this research may lay the grounds to explore new gaps for researchers in the future and bring to the fore the strengths and the weaknesses of the three architectures.

In the area of home applications, object detection has been applied for identification and recognition of commodities and events, description and pattern analysis of images, rain and shadow detection, classification of species, etc. With the advent of social media and its growing acceptability and penetration into different societies, the data has become more diverse, and the need for surveillance has become even more imperative (Hassan & Tukur, 2025). Consequently, Goldman et al. (2019) introduced a new technique for object detection, which uses SKU110K to detect real-time events, such as festivals, social gatherings and talks, online. Cuenat and Couturier (2022) came up with the idea to compare CNNs models with transformer models for digital holography. this research aimed to recreate the amplitude for the models. Additionally, the phase of the experiment involved assessing the distance of the object in an image from the main hologram. The authors compared the performance of EfficientNet and a transformer-based architecture using a dataset of 3,400 images. The samples in the dataset were separated into four categories: negative images without filters, negative images with filters, positive images without filters, and positive images with filters. The result obtained showed that ViTs have better performance than CNNs.

In a study conducted by Nafisah et al. (2023), the authors proposed a technique that uses a combination of CNN-based and ViT architectures to detect COVID-19 in X-ray images. According to the authors, the CNN-based EfficientNetBy and ViT-based SegformerB5 produced comparable results. However, they highlighted the disadvantage of using higher computational power systems in experimentation, which is the main drawback of using ViT-based architectures. In a work presented by Springenberg et al. (2023), The authors introduced a technique called

the effectiveness of self-supervised learning to help improve the performance of CNNs and ViTs techniques in situations where labelled data is limited. The model proposed by the authors recorded an additional 5% improvement in the classification of dermatological images compared to the state-of-the-art supervised methods. The finding shows that transformer-based models may be more useful than CNNs in environments with limited resources.

Guo et al. (2023) conducted research where they attempted to solve the multi-modal medical image problem by designing a new architecture that employs a transformer to fuse information from MR and PET images. By incorporating the proposed approach, they established that there is a surface rate that achieved 2% improvement compared to single-modality conventional techniques. The development of tasks on advanced medical image identification and classification has been stepped up in recent years, where techniques like DL have been used to compare the performance of CNN-based and ViT architectures. Sharma et al. (2025) made a global survey on the use of transformer models in image segmentation and classification for medical images to show the efficiency of the architecture in capturing distant context, which is significant for accurate demarcation. The finding shows that there is an improvement in performance by 0.5% on the Dice score of brain tumour segmentation as compared to traditional deep learning techniques. The research (Xin et al., 2024), integrate AI to identify dermoscopic images for early diagnosis of skin cancer. The author incorporates CNN and ViT to assess the effectiveness of skin transplantation. Two datasets are used, namely HAM10000 and a clinical dataset collected through dermoscopy. This approach achieves an accuracy of 94.3% and 94.01%, respectively.

Despite the remarkable contribution made by the authors of the reviewed work so far, it is evident enough attempt has not has been made to compare the performance of plain CNN, transfer learning model and a ViT-based model in each object detection task. For the plain CNN, there is an advantage of developing the model from scratch, selecting the right layers and choosing the right number of blocks. The downside, however, is the amount of time and data needed to train the model. Notwithstanding, it is impressive that researchers understand how these important architectures compare in terms of their performance. This is the gap that this research intends to fill.

2 Materials and Methods

This research follows a step-by-step experimental approach to achieve its objectives, which are structured around five (5) different key modules. Each builds upon the output of the previous one, ensuring a coherent and systematic approach. The methodology has the following phases:

i. Data Selection: a dataset of 2100 thousand images was gathered and manually annotated, and used for the experiment.
ii. Dataset Pre-processing: The preparation and cleaning of this data to make it suitable for the analysis and model, which includes image annotations, normalisation, resizing, and scaling, was carried.
iii. Proposed CNNs and ViTs Models: An overview of the proposed model architecture and the approach of its unique features and advantages.
iv. Training, Testing, and Evaluation: The model was trained on a training dataset, tested, and evaluated using metrics like IoU and Mean Squared Error.
v. Performance Analysis: the performance of each of the selected models was evaluated using some key evaluation metrics to identify its strengths and weaknesses.

The research design focuses on the comparative analysis of two deep learning algorithms (CNNs and ViTs). These models are selected for their popularity and applicability in object detection tasks.

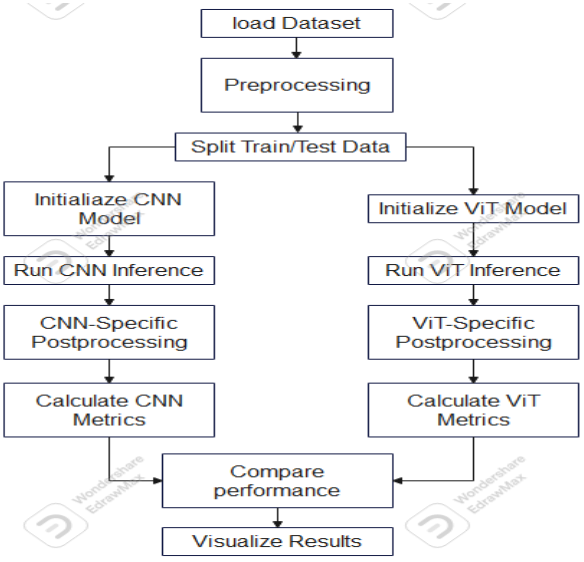

Figure 1: System Architecture

The overall goal of this research is to evaluate and compare their performance when applied to the image

dataset, using various deep-learning metrics (IoU and mean square error). Therefore, this research follows an experimental approach, where different deep learning algorithms (CNN and ViTs) are implemented and evaluated based on their performance.

2.1 Technologies Used

Several tools and technologies were used in conducting the experiment and evaluating the performance of these algorithms. For the hardware, a laptop with Intel Core i3 (5th Gen), 4GB RAM, 256GB HDD storage, Windows 10 operating system, and Python 3.8+ as programming language was used. Anaconda/Jupyter notebook was utilised as a programming environment. Libraries such as NumPy, Pandas, Matplotlib, scikit-learn, seaborn, TensorFlow, Pytorch, Beautiful Soup (bs4), OpenCV (cv2) and LabelMe for image annotations, were also used as highlighted in the next section.

2.2 Dataset and Annotation

The dataset used in this research contains 2100 images of household items. It was built using a mobile phone specifically for object detection. The images were saved to a folder in JPEG format. Subsequently, each image was then annotated with the .xml extension and imported into a programming environment for further analysis. The dataset annotations contain multiple rows, each representing an individual image, its corresponding object, and the bounding box coordinates (xmin, ymin, xmax, and ymax), capturing a specific combination of these values. The annotation was carried out using the LabelMe annotator.

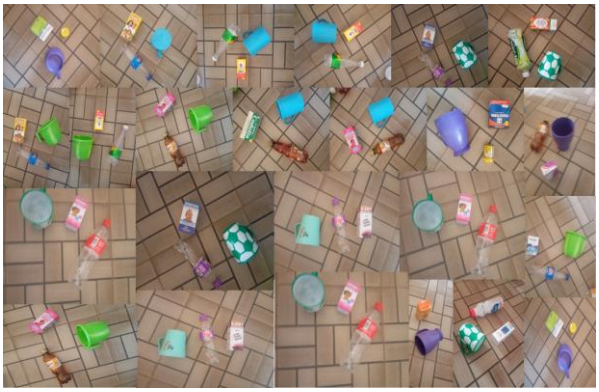

Figure 2: Sample Images before Annotation

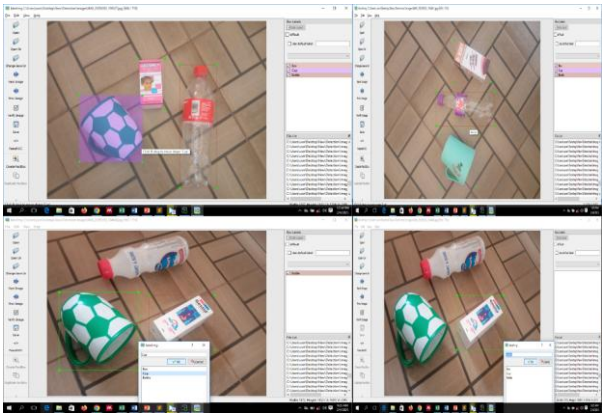

Figure 3: Sample Images After Annotation

2.3 Evaluation Metrics

Mean Squared Error is one of the most used metrics for regression tasks in object detection. It is used to calculate the mean squared difference between the actual values and the predicted values. A lower average square error indicates better performance, as it means the model's predictions are closer to the true target values (Raza & Victor, 2021). Equation 1 gives the mathematical expression of the formula for calculating the MSE. Where n represents the number of samples, y_i represents the actual target, and y_j represents the predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_j)^2 \qquad (1)$$

Intersection over union (IoU) is a common metric used to evaluate the performance of object detection models. As shown in equation 3.3. IoU is a number that quantifies the degree of overlap between two boxes. In object detection and segmentation tasks it used to evaluate how well the predicted region overlaps with the corresponding ground truth region (Zhang & Zhang, 2024).

$$IoU = \frac{TP}{TP+FP+FN} \qquad (2)$$

Table 1: Experimental Parameters

Parameter	Experimental Setup
Sample Size	2,100 images
Image Resolution	224 × 224
Epochs	50
Batch Size	32
Patch Size	32
Optimizer	Adam
Learning Rate	1e-4
Train/validation/test splits	80%, 20%
Training Time	7,332 to 66,300 sec

3 Results

This section discusses the result of the entire experiment. The three models compared in this work, the plain CNN, fine-tuned VGG16 architecture and vision transformers, and were all trained using 50 epochs to generate the results discussed in this section.

3.1 False Negative

From the experiments, when training the plain CNN, fine-tuned, and ViTs individually, each showed that the false negatives dropped as epochs increased. This is evident are irregularities in the curve, as normally, a perfectly trained result should present a perfect downward curve, indicating that the model is learning to decrease false negative results during the training as epochs increase. In this case, a high rate of false negatives, with misses of 54.98% at the first 20 epochs, was initially recorded for ViTs, while 66.26% and 55.49% for the plain CNN and fine-tuned models were recorded.

The irregularities indicate the and fine-tuned VGG16 models missed nearly half of the objects early in the training process for plain, while the ViTs-based architecture missed nearly a quarter of the objects in the first 20 epochs. The variation of above 50% for plain CNN and fine-tuned, and over 20% for ViTs suggests a drop in the model's ability to detect object. Between 21 and 35 epochs, the models and the detection ability improved, resulting in a drastic reduction false negative to 10%, 11% and 11% for the ViTs, VGG16 and the plain CNN, respectively. However, false positives slightly rose toward the final epochs from 10% to 11% for the plain CNN while fine-tuned, and ViTs models remain unaffected. Although the models showed variation when training a plain CNN, it generally adapted well toward the end, with a significant drop to 10.01% and 5.88% at the 50th epoch for the fine-tuned VGG16 and ViTs, implying that the detection accuracy improved with the increase in the training time.

3.2 Intersection over Union (IoU)

As shown in Figures 6, 7, 8, and 10, in the first 20 epochs, the plain CNN, fine-tuned VGG, and ViTs had already achieved IoU of 53.02%, 60.24%, and 65.42%, continuing up to 80.35%, 82.53%, and 85.29% at 40 epochs. The models' prediction ability continued to improve, exceeding 92.14%, 94.04%, and 95.04% at 45 epochs, while achieving the highest IoU of 97.10% for the plain CNN, 98.01% for the VGG16 and 98.42 for the ViTs at 50 epochs.

3.3 Mean Squared Error (MSE)

An increase in the detection accuracy implies that the models are better at predicting the correct outputs. As the number of samples increases, the MSE typically decreases and vice versa. Since there is an improvement in accuracy, it is expected that the model makes fewer mistakes, which means there are fewer errors to average and square, and the errors that occur are likely to be smaller in magnitude, since the models are better at predicting the correct outputs. As a result, the MSE decreases in magnitude, indicating that the model's predictions are very close to the true values (TP).

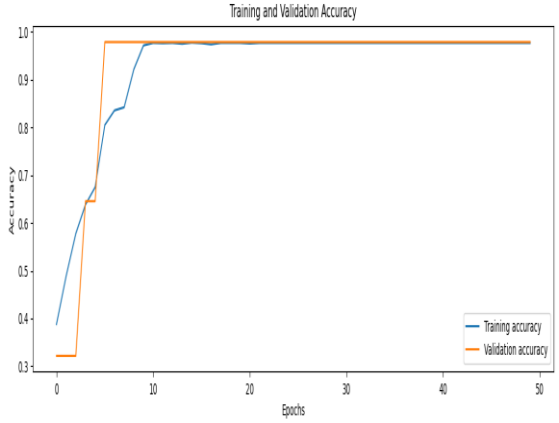

Figure 4: Plain CNN Training Curve for 50 Epochs

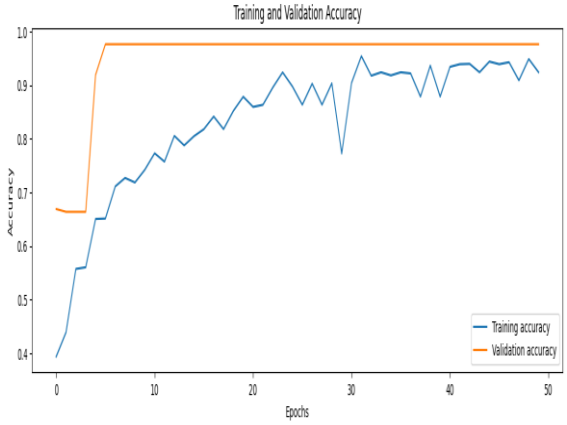

Figure 5: Transfer Learning Training Curve for 50 Epochs

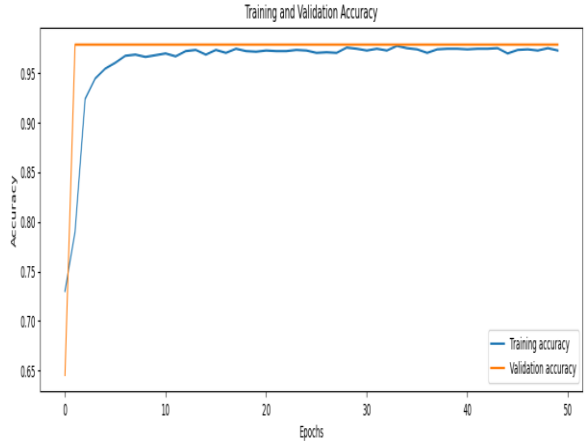

Figure 6: ViTs Training Curve for 50 Epochs

As shown in the Figures below 6, 7, and 8 above, in the first 20 epochs, the models achieved accuracies of 53.02%, 60.24%, and 65.42%, and the MSE decreased to 35.77%, 32.40%, and 30.77% for plain CNN, fine-tuned models, and ViTs, respectively, continuing to decline to 15.20%, 13.02%, and 12% by 40 epochs. At 50 epochs, the models reached accuracies of 97.10%, 98.01%, and 98.42%, with the MSE dropping to 11.40%, 10.21%, and 5.97% for plain CNN, fine-tuned models, and ViTs. The models learn to improve IoU as the number of epochs increases, which reduces errors and leads to better detection performance without irregularities.

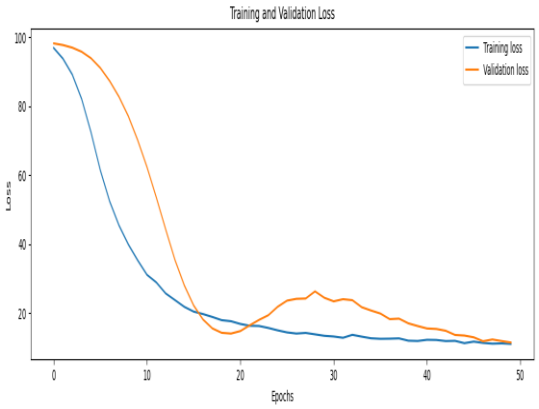

Figure 7: Plain CNN Loss Curve for 50 Epochs

3.4 Execution Time

Training ViTs, as shown in Figure 12, took significantly longer than the plain CNNs, but less time than the fine-tuned VGG16 model. This may be because the plain CNNs model automatically trains and predicts all at once. Conversely, the fine-tuned model trains first before validating. However, training the plain CNNs on a small sample dataset tends to be faster than fine-tuning due to millions of pre-trained parameters and the effect of a low computational power environment.

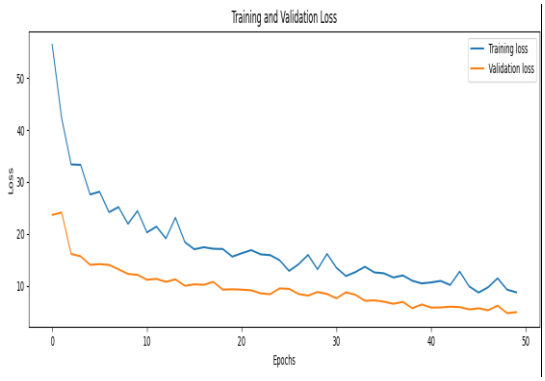

Figure 8: Transfer Learning Loss Curve for 50 Epochs

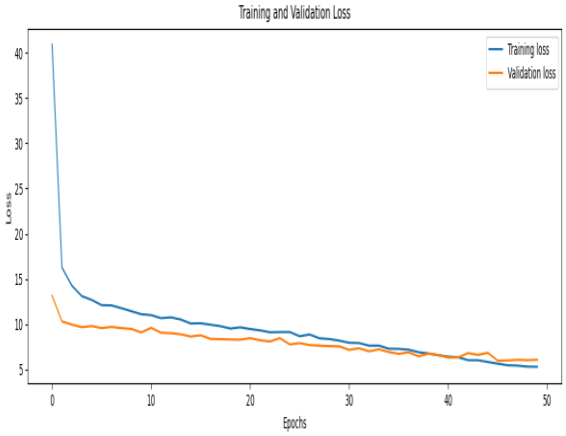

Figure 9: ViTs Loss Curve for 50 Epochs

Similarly, ViTs' longer training time is primarily due to the complexity of the transformer architecture and the computational cost of multi-head self-attention layers, which evaluate the global context across image patches. Limited computational resources when working with this model may also affect its overall performance. ViTs took approximately 2,989.2 seconds, compared to 2,932.8 seconds for plain CNNs, and much less than fine-tuned models at 66,300 seconds. Despite the higher computational cost, the quality of detection and learning consistency improved, justifying the additional resources. Compared to the plain CNNs, ViTs were slower to train but proved more effective in feature learning and generalisation, especially with small datasets containing complex objects. The extended training time is expected, as ViTs require more memory and processing power due to their parallel processing and patch-wise feature extraction.

Table 2: Overall Result Summary Comparison

Metric	Plain CNN	FinetunedVGG16	ViTs	Best Result
Accuracy	Peaked at 97.10%	Peaked at 98.01%	Peaked at 98.42%	ViTs
MSE	Dropped to 11.40%	Dropped to 10.21%	Dropped to 5.97%	ViTs
Loss	Fluctuated, moderate drop to 11.23%	Consistently low and stable drop to 10.21%	Consistently very low and stable drop to 5.88%	ViTs
Execution Time	7,332 seconds, Fast (w/o VGG16)	66,300 seconds, Slow with VGG16	7,473 seconds, Moderate, but efficient	CNN (faster), ViTs (better)

3.5 Loss Metrics

Figure 11 gives and insight into the model training loss. At first 20 epochs, the loss metrics for the plain CNNs dropped to 29.24%, 25.40% for the VGG16 and 24.24% for the ViTs. The same trend was maintained at the 40th epoch, where they dropped to 12.20%, 11.56% and 11%, and nearly halved for the VGG16 and the ViTs at the 50th epoch, standing at 11.23%, 09.01% and 5.88% for the CNNs, the fine-tune architecture and the ViTs, respectively. The box_loss across training and validation on images shows that the box_loss drops during training and prediction as the number of epochs increases. The consistency suggests robust training and prediction irrespective of the epochs used; they all learned to reduce error as the epochs increased.

4 Summary of Results

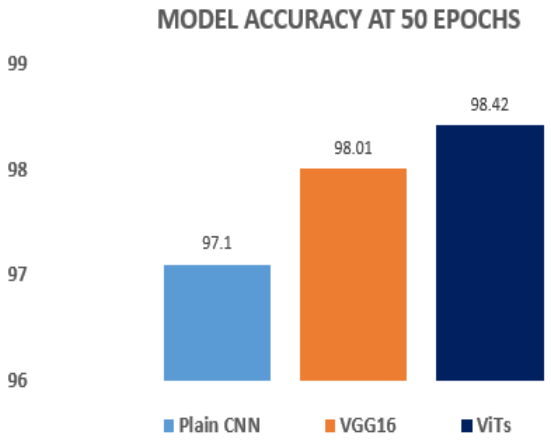

Figure 10: Models IoU (%)

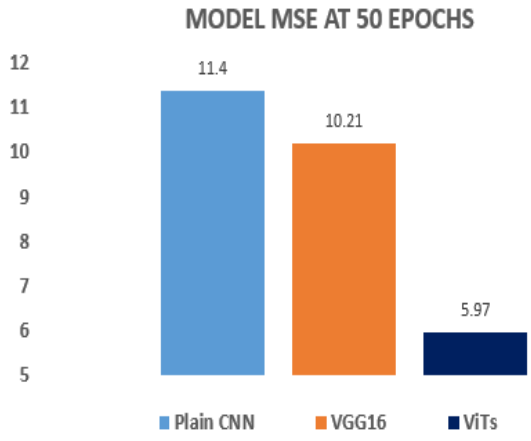

Figure 11: Models Loss (%)

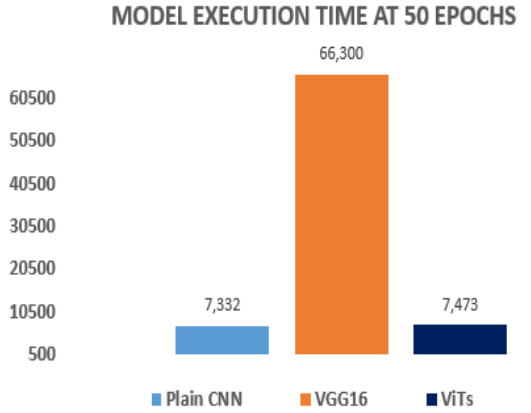

Figure 12: Models Execution Time (Sec)

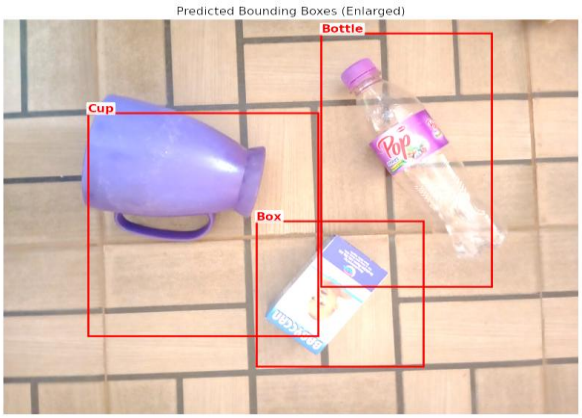

Figure 13: Sample A

5 Comparison between Benchmark study and Proposed Research

This section compares an existing benchmark system with the proposed research experiment, focusing on dataset characteristics, domain specific, training

configurations, computational environment, and model performance results.

The benchmark system used a public dataset with 11,000 annotated images, resized to 224 × 224 for processing. It employed CNN and ViT architectures, trained on hardware that included an Intel i7 processor and an NVIDIA RTX 3060 GPU with 12 GB of RAM. Although the number of training epochs, OS version, and execution time were not specified, the model achieved peak accuracy of 94.30% (ViT) and 94.01% (CNN), with ViT delivering the best results among the evaluated models. In comparison, our work used a custom, self-collected dataset comprising 2,100 annotated images captured under controlled conditions. The images were also resized to 224 × 224 to match the benchmark setup. Three models were evaluated: Plain CNN, fine-tuned VGG16, and ViT, all trained in a low-computational environment. All models demonstrated consistent performance and were trained for 50 epochs, recording distinct execution times: 7,332 seconds for the plain CNN, 66,300 seconds for the fine-tuned model, and 7,473 seconds for ViT. Despite limited computational resources and a smaller dataset, the proposed approach outperformed the benchmark. Specifically, the fine-tuned CNN achieved an IoU of 98.01%, while the ViT had the highest IoU at 98.42%. The plain CNN attained an IoU of 97.10%. However, the fine-tuned model required significantly more training time, highlighting a trade-off between IoU and computational demand. The fastest model was the plain CNN, while the ViTs provided the best balance between IoU and acceptable execution time, considering the dataset and hardware constraints.

Table 3: Comparison between Benchmark study and Proposed Research

Parameter	**Benchmark Study (Xin et al., 2024)**	**Proposed Research**
Dataset Type	Public dataset with annotations	Custom dataset with annotations
Application Domain	Healthcare object detection	Household object detection (bottles, boxes, cups)
Dataset Size	11,000 images	2,100 images
Image Resolution	224 × 224	224 × 224
Annotation Format	Not specified	XML (Pascal VOC format)
Model Architectures	CNN, ViT	Plain CNN, Fine-tuned VGG16, ViT
Training Strategy	From scratch	From scratch (CNN), Transfer learning (VGG16), ViT
Hardware Configuration	Intel i7, 2× NVIDIA RTX 3060, 12 GB RAM	Intel Core i3 (5th Gen), 4 GB RAM
Operating System	Not reported	Windows 10
Number of Epochs	Not reported	50
Training Time	Not reported	7,332 s (CNN), 66,300 s (Fine-tuned), 7,473 s (ViT)
Evaluation Metric	Accuracy	IoU, MSE
Best Result Achieved	94.30% (ViT)	98.42% (ViT)

6 Contribution

While many related works rely on publicly available datasets, we generated and pre-processed tailored datasets for the specific requirements of this research. This process included careful data collection, preprocessing, and manual annotation to ensure accuracy and relevance to the problem domain. This contribution ensures the model is trained on high-quality, domain-specific data, leading to more reliable and application-specific results. In contrast to previous researchers in the same field, who often omit critical experimental details such as loss metrics, execution time and the number of training epochs used, we have explicitly documented and reported these metrics, thereby enhancing the reproducibility and transparency of the research and allowing other researchers to replicate the study and verify the results. It also supports a better understanding of the computational efficiency and resource requirements of the proposed approach. Thus, the efficacy of each model in object detection is hereby established.

7 Conclusion

From these experiments, which were conducted using a newly created dataset, the ViTs model has a faster convergence time than fine-tuning the VGG16 architecture but slower than plain CNN. It is equally faster but it depends on the patch size, the complexity, and the size of the dataset. We have shown that ViTs took longer than the plain CNN but still faster than transfer learning, with the advantage of better IoU and loss performance. The increased execution time for the transfer learning model and ViTs is justified by their superior generalisation and predictive capabilities over the plain CNN architecture. In the future, we plan to assess the possibility of building an ensemble model to leverage the advantage of these models. We also plan to go beyond the experimental stage and deploy the proposed model to real.

References

Balarabe, A. T., & Jordanov, I. (2021). *LULC IMAGE CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORK Anas Tukur Balarabe and Ivan Jordanov School of Computing , University of Portsmouth , UK*. 5985–5988.

Balarabe, A. T., & Jordanov, I. (2022). *INTERPOLATION AND CONTEXT MAGNIFICATION FRAMEWORK FOR CLASSIFICATION OF SCENE IMAGES*. 93–100.

Balarabe, A. T., & Jordanov, I. (2024). A Deeper Look Into Remote Sensing Scene Image Misclassification by CNNs. *IEEE Access*, *12*(December 2023), 123078–123098. https://doi.org/10.1109/ACCESS.2024.3354976

Cuenat, S., & Couturier, R. (2022). Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography. *2022 2nd International Conference on Computer, Control and Robotics, ICCCR 2022*, 235–240. https://doi.org/10.1109/ICCCR54399.2022.9790134

Guo, Z., Zhao, W., Wang, S., & Yu, L. (2023). HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14225 LNCS*, 755–764. https://doi.org/10.1007/978-3-031-43987-2_73

Hassan, N. A., & Tukur, A. B. (2025). *Comparing the Performance of Convolutional Neural Networks and Vision Transformers in Object Detection : A Review*. *18*(12), 184–201. https://doi.org/DOI: https://doi.org/10.9734/ajrcos/2025/v18i12798

Jakhar, D., & Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: definitions and differences. *Clinical and Experimental Dermatology*, *45*(1), 131–132. https://doi.org/10.1111/ced.14029

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences (Switzerland)*, *13*(9). https://doi.org/10.3390/app13095521

Nafisah, S. I., Muhammad, G., Hossain, M. S., & AlQahtani, S. A. (2023). A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection. *Mathematics*, *11*(6). https://doi.org/10.3390/math11061489

Raza, D. M., & Victor, D. B. (2021). Crime Using Random Forest. *Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)*, *7*, 980–987.

Reedha, R., Dericquebourg, E., Canals, R., & Hafiane, A. (2022). Transformer Neural Network for Weed and Crop Classification of High Resolution UAV Images. *Remote Sensing*, *14*(3), 1–20. https://doi.org/10.3390/rs14030592

Sharma, R. R., Sungheetha, A., Tiwari, M., Pindoo, I. A., Ellappan, V., & Pradeep, G. G. S. (2025). *Comparative Analysis of Vision Transformer and CNN Architectures in Medical Image Classification*. *Icsice 24*, 1343–1355. https://doi.org/10.2991/978-94-6463-718-2_112

Springenberg, M., Frommholz, A., Wenzel, M., Weicken, E., Ma, J., & Strodthoff, N. (2023). From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. *Medical Image Analysis*, *87*. https://doi.org/10.1016/j.media.2023.102809

Wang, C., Xu, H., Zhang, X., Wang, L., Zheng, Z., & Liu, H. (2022). Convolutional Embedding Makes Hierarchical Vision Transformer Stronger. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13680 LNCS*, 739–756. https://doi.org/10.1007/978-3-031-20044-1_42

Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., Zhou, Q., Wang, S., Li, L., Yang, F., Xu, S., & Chen, H. (2024). An improved transformer network for skin cancer classification. *Computers in Biology and Medicine*, *149*(June), 105939. https://doi.org/10.1016/j.compbiomed.2022.105939

Zhang, H., & Zhang, S. (2024). *Focaler-IoU: More Focused Intersection over Union Loss*. 1–4. http://arxiv.org/abs/2401.10525