Harnessing Computer Vision to Identify Pests and Predators: A Comparative Model Analysis

Mahsa Mohaghegh
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
0000-0003-2228-8300

Lara Jaber
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
0009-0002-8475-1612

James Stratford
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
0009-0001-0971-8313

William Maxwell
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
0009-0008-3412-3283

Connor Tamatea
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
0009-0004-5056-5875

Lachlan Crawford
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
0009-0003-0740-3582

***Abstract*— This study focuses on developing an intelligent system using machine learning algorithms and motion-sensor camera traps to achieve real-time classification of pest predators in the setting of the bushlands of New Zealand. The primary goal is to provide an optimized tool for ecological monitoring, aiding in the preservation of native Kiwi bird habitats and aligning with the eco-city initiatives. By training and comparing various deep learning models, including Convolutional Neural Networks (CNNs), Multi-Layer Perceptron (MLP), and Vision Transformers (ViT), the system aims to assist in accurately identifying and managing pest populations. We found that our best-performing model on our data was ResNet-50 with an overall accuracy of 98.15% and an average f1-score of 0.982 across the five classes. This was closely followed by DenseNet-121 with an overall accuracy of 97.95% and an average F1 score of 0.978 and our CNN with a Vision Transformer model with an overall accuracy of 97.58% and an average F1 score of 0.976.**

Keywords— image classification, transfer learning, vision transformers, CNN, MLP

I. INTRODUCTION

For many years, conservation efforts have existed around New Zealand to protect native bird populations from introduced pest predators that threaten their presence. As a flightless nocturnal bird, the Kiwi is particularly vulnerable to predation by introduced species, making its protection a matter of utmost importance. Analog methods such as ink pads and chew cards have been the standard method of predator detection; however, they have their limitations and drawbacks. In this context, the application of image classification technology emerges as a powerful tool in the ongoing efforts to monitor and conserve these unique species. Recent advancements in computer vision and deep learning techniques have demonstrated significant potential for improving the accuracy and efficiency of predator detection in natural environments [1]. This study builds upon the promising results achieved by previous research in wildlife monitoring and image classification, particularly in identifying small animal species such as cats, rats, possums, mice, and stoats [1]. Furthermore, other studies have compared different machine learning models relevant to our paper in the past, with ResNet being a particularly popular model [2]. However, we wanted to take a novel approach to this study and analyze the performance of certain models that have not been extensively compared in other articles.

Our research aims to harness the capabilities of cutting-edge technologies by comparing various deep learning models including Convolutional Neural networks (CNNs) such as ResNet-50 and DenseNet-121, and Vision Transformers (ViT). Each of these models has its strengths and weaknesses. Through comprehensive evaluation and in-depth analysis, we seek to identify the most effective and realistic approach for real-time identification and management of pest predators within New Zealand's clean natural landscape. The images used in the dataset to train the models were directly sourced from the local New Zealand bush, containing pest predators such as cats, mice, possums, rats, and stoats. These pests were primarily chosen because of their widespread proliferation and the problems they create for native wildlife. We predominantly focused on examining the overall accuracy achieved by each model. With the overall accuracy acting as a baseline guide, we used class-specific performance metrics such as precision, recall, and F1 scores to give deeper insights into the models' effectiveness in distinguishing between different predator species. The top-performing models were two transfer learning architectures (ResNet-50 and DenseNet-121), as well as the Vision Transformer (ViT) model. ResNet-50 achieved the highest overall accuracy, outperforming DenseNet-121 and the ViT model by diminutive margins of between 0.2% and 0.57%.

This study will investigate and answer what is the most effective machine learning model for image classification of pest predators in the native New Zealand bush.

II. RELATED WORK

Within the domain of computer vision and image classification, several studies have already investigated the utilization of deep learning techniques and transfer learning to enhance the accuracy of image classification. Recent research has showcased the effectiveness of various models in addressing various challenges in animal image classification. This section provides an overview of major contributions made by prior studies, with a focus on the significant insights that have emerged recently.

Firstly, as proposed by Pillai et al. [3], CNNs have been demonstrating great performance by achieving an 81% accuracy rate in categorizing ten distinct monkey species with a dataset of 1370 training images and 272 testing images.

This breakthrough showcases the capability of CNNs to accurately identify species, which is directly relevant to our objective of real-time identification and management of pest predators in New Zealand's pristine natural landscapes. Furthermore, by combining CNNs with transfer learning techniques such as ResNet-34, it is sought that it can achieve an even greater accuracy rate of 96.43% [4] where the models solved the challenge of image content retrieval in large datasets. This result indicates the substantial improvements that can be realized by leveraging deep learning for image analysis, aligning with our goal of enhancing the efficiency and accuracy of pest predator identification. Furthermore, an extensive investigation [5] delved into image classification using pre-trained CNN models, particularly focusing on scenarios with limited data availability and computational constraints. Their findings highlighted the effectiveness of pre-trained models, with one model, Densenet, achieving an impressive average accuracy of 99.65%. This underscores the significance of transfer learning, a concept directly relevant to our work as we seek to make the most of limited data resources for pest predator identification in the New Zealand bush. Given the popularity of ResNet models, we have chosen to experiment with alternative architectures such as Densenet121, which, as revealed in another study [5] achieved the utmost performance with an outstanding accuracy in image classification.

Additionally, with the deployment of a CNN model [6] it was theorized that it could be beneficial in the assistance of animal husbandry. With a provided dataset comprising 37,322 images across 50 diverse animal categories, a CNN model was trained to be able to identify livestock which resulted in a robust CNN model with an accuracy of 90.85%. This achievement further exemplifies the potential of CNNs for animal image classification, suggesting optimistic prospects for our approach of employing CNN models for pest image classification.

A distinct methodology was employed [7] where features from RGB and NIR images were extracted through a CNN model, incorporating both an LMT and Dynamic Transformer Encoder. In parallel, an MLP was utilized to extract pertinent information from the location data associated with the images. The dataset underpinning this research is available from Plantnet (n.d.). Of note, the model demonstrated Top 30 Error Rates of 0.7278, 0.7297, 0.6594, and 0.6567. After reading this study we wanted to incorporate a Vision Transformer model of our own on our unique dataset and compare it with other models using the same dataset, classes, and performance metrics.

A 2014 study [8] introduced VGG models from the University of Oxford's Visual Geometry Group. Evaluating the ImageNet dataset, the VGG-16 and VGG-19 architectures stood out. Their single model achieved a top-5 test error rate of 7.3% in the classification task, with further improvements to 6.8% in localization. By employing an ensemble of models, they pushed the error rate down to an impressive 6.67%, solidifying VGG's position among the top-performing image classification models of its era. Due to this model's success, we also wanted to incorporate it in our model comparisons.

A similar study was undertaken [9] that explored this approach in more depth. Four object detection CNN models were evaluated on a dataset containing a wide collection of different animals. The overall performance of their models produced a mean average precision of 80.02 and a mean average recall of 73.6. These metrics are very acceptable and provide promising evidence for further study into the implementation of object detection in the future.

For a concise summary of the references, datasets, metrics, challenges, and strengths related to the studies mentioned above, refer to Table 1 below.

TABLE I. COMPARATIVE ANALYSIS OF VARIOUS MACHINE LEARNING MODELS ON DIVERSE DATASETS

Reference	Dataset	Model	Performance	Challenge	Strengths
[3]	10 Monkey Species Images	CNN	0.81	-Assessing how well the CNN model generalizes to new and unseen monkey species. -Data imbalance -Morphological and behavioral features of different monkey species.	-Automated identification which provides a valuable tool for wildlife conservation efforts. - Effective species classification.
[4]	CATS/DOGS	Resnet-34 and transfer learning	0.9643	-small sample data recognition. -limitations of traditional methods.	-Use of transfer learning -Improved accuracy and efficiency in image recognition.
[5]	Wang dataset	CNN (VGG, ResNet50, DenseNet)	0.7894 0.8905 0.9965	-Data availability and constraints.	-Outstanding image classification performance with densenet considering limited data availability.
[8]	ILSVRC	VGG	92.7%	-Dependent on public dataset. -Specific focus	-Utilization of small features. -Focus on depth. -State of the art performance.

III. Proposed Method

In this study, we investigated the adaptation of seven various deep learning architectures for image classification tasks. These models were: ResNet-50, DenseNet-121, GoogLeNet, EfficientNet-B0, a Vision Transformer model (ViT-base-patch16-224-in21k), VGG-16 and an MLP model. Our objective was to deploy the best-suited model by comparing all the different models developed. Each of these deep learning architectures has its unique characteristics and advantages, therefore having such a wide range of models is beneficial for comparison. The MLP model was included as a baseline to provide a simple and well-understood benchmark for evaluating the performance of more complex deep learning architectures in our image classification study. This was to ensure that any improvements observed in the more complex deep learning architectures could be attributed to their inherent capabilities rather than simply to the dataset. ResNet-50, known for its depth and skip connections, excels in capturing intricate image features. DenseNet-121's dense connectivity fosters parameter efficiency and robustness [10]. Additionally, these transfer learning architectures are pre-trained in a large-scale dataset, ImageNet, which consists of millions of images spread across 1000 categories [4]. This is done to recognize common patterns which speed up the learning process for our theory. Vision Transformers take a different approach and break images up into sequential patches and use the attention mechanism to learn these sequences [11]. By fine-tuning these models and tailoring them to our specific dataset, we aim to leverage their strengths while addressing potential limitations.

A. Dataset

The dataset used in training and testing is a comprehensive collection of images, captured via strategically positioned camera traps within the bushlands of New Zealand. The setting of these locations is varied as the New Zealand bush is diverse and different angles were experimented with. Upon the motion sensor detecting movement, the camera trap is triggered and multiple snapshots are taken. Depending on whether it is taken in the day or night, a camera flash may be used. The camera traps are set up across many sites and are often set up with bait to attract predators into the camera's viewport. The cameras were also often angled downwards to avoid the camera's view being blocked completely by the animal and to lower the chance of the camera being set off by tree branches moving in the wind.

a. Composition

Our dataset contains a unique blend of native New Zealand fauna, offering a realistic snapshot of the local conditions and challenges associated with camera traps, such as variations in lighting, lens flare, and image blurriness, attributable to environmental variables and equipment or logistic constraints. These challenges provide a realistic layer of complexity and are vital in fine-tuning our models to adapt to real-world constraints. Fig 1, Fig 2, and Fig 3 show examples of the images that are in our dataset.

Fig 1. Cat image taken in the night.

Fig 2. Possum image taken in the night.

b. Focus and Categorization

Given our emphasis on pest predators of Kiwi, we extracted five classes from our extensive dataset: cats, mice, possums, rats, and stoats, totaling 268,445 images. This collection of images provided a robust foundation for our training and testing sets. Fig 4 shows the distribution of classes in the whole set.

c. Pre-processing and Data Augmentation

We subjected our image dataset to a series of preprocessing steps to facilitate accurate and efficient model training. The purpose of these steps is to standardize the data and to expedite convergence during the training phase.

i. Image Resizing

Images were resized to a resolution of 224x224 pixels. This resolution was chosen because it is the input size during the pre-training stages for many pre-trained models. This ensures

Fig 3. Stoat image.

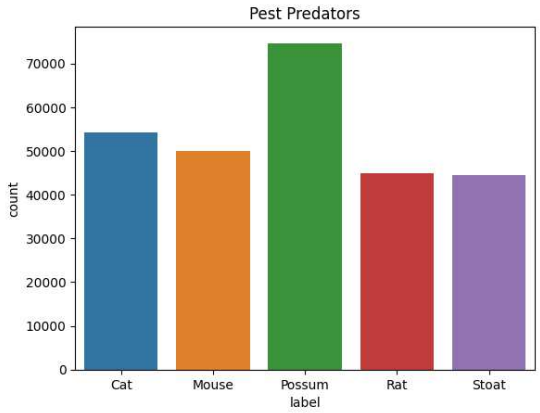

Fig 4. Class distribution

compatibility across established networks such as ResNet, DenseNet, and Vision Transformers. The chosen resolution is a balanced compromise, it retains sufficient details within the image crucial for accurate feature extraction and interpretation, while also optimizing for computational efficiency, limiting the strain on resources during the training phase [12].

ii. Normalization

After resizing, images were normalized across the RGB channels. Each color channel was normalized with a mean of 0.5 and a standard deviation of 0.5. Normalization transforms each pixel's channels into a range of –1 to 1. This is an important step as raw pixel values will range between 0 and 255. By normalizing the images, model convergence is sped up and training is stabilized [13].

iii. Rotation

In addition to the above pre-processing steps, the dataset also was randomly rotated between 0 and 30 degrees. Image rotation is a common data augmentation technique, proven to enhance model accuracy and generalization to unseen data [14]. This augmentation technique is designed to introduce variety and enable the models to learn and adapt to different orientations of the objects within the images. This enhances the ability of models to generalize data in a production-ready environment.

d. Dataset significance

This dataset is much more than just a bunch of pictures; it's a detailed look into New Zealand's bushlands and the animals that live there. It gives us a lot of information about where these animals live, how they behave, and how they interact with each other. It connects our models to the main goal of our study, capturing and accurately identifying pest predators in the unique environment of the New Zealand bush. The challenges inherent to the dataset, such as lighting variances, obstructions, and movement blurs, replicate the real-world complications that models would encounter, providing an authentic testing ground for model robustness and adaptability.

B. Model Training

For the training process, we set the training parameters to five epochs to allow the models sufficient learning while avoiding overfitting, enhancing the models' ability to generalize to unseen data. The cross-entropy loss function was utilized given its effectiveness in classification tasks, providing a clear measure of the disparity between the predicted probability distribution and the actual distribution.

- Learning Rate: A learning rate of 0.01 was applied to balance the speed and stability of the convergence during training, preventing oscillation and ensuring steady progress towards the local minimum.
- Batch Size: A batch size of 32 was chosen to optimize the computational resources, allowing for efficient updates to the model weights and improved model convergence time [15].

Three deep learning architectures were the focus of this study, namely, ResNet-50, DenseNet-121, and Google's Vision Transformer. Other popular pre-trained models were also trained and compared against a baseline, as well as an MLP model. ResNet-50 and DenseNet-121 were both sourced from the Pytorch and Torchvision packages, while the Vision Transformer model (vit-base-patch16-224-in21k) was sourced through the HuggingFace library, transformers. Adjustments were made to the classification layer of these models so that they were fitted to our number of classes.

IV. Results and Discussions

A. Training Environment

To facilitate our data analysis and model training, we utilized Google Colab, a cloud-based platform that provides free access to computing resources including GPUs and TPUs [5]. Google Colab allows for collaborative coding and data analysis using Jupyter notebooks, making it a convenient choice for our research.

B. Model Overview

The seven network models that were developed and experimented on were ResNet-50, DenseNet-121, EfficientNet-B0, VGG16, GoogLeNet, a Vision Transformer, and an MLP, where their performances were evaluated.

By analyzing their training and testing accuracies. Additionally, precision, recall, F1-scores, and the confusion matrix were used to further observe the performance of the models.

C. Performance Evaluation

Table 2 illustrates the overall accuracy performance of our selected models. The ResNet-50 model displayed the highest accuracy at 98.15%, surpassing other models with slight margins.

The models were tested on the test set and results were analyzed through a confusion matrix to give a deeper insight into its class classification performance. One such confusion matrix is shown below in Fig 5, where the predictions of labeled images from our best-performing model (ResNet-50) can be seen, and in Fig 6 the comparative accuracies across all trained models can be seen.

TABLE II. MODEL METRICS AGAINST CLASSES

	F1-Score				
Model	***Cat***	***Mouse***	***Possum***	***Rat***	***Stoat***
ResNet-50	0.99	0.97	0.99	0.98	0.98
DenseNet-121	0.99	0.97	0.98	0.97	0.98
ViT-base-patch16-224-in21k	0.98	0.97	0.98	0.97	0.98
GoogleNet	0.99	0.97	0.99	0.98	0.98
EfficientNet-B0	0.89	0.90	0.79	0.78	0.81
VGG16	0.89	0.79	0.86	0.85	0.76
MLP	0.76	0.76	0.73	0.77	0.73

D. Challenges and Improvements

Several challenges were encountered in the model's ability to accurately classify specific classes, particularly those involving 'Cat' and 'Possum' species. The confusion matrix unveiled a pronounced number of instances where 'Cat' was misclassified as 'Possum' and vice versa, more so than in other classes.

This higher difficulty suggests that there is room for enhancement in the model's feature recognition capabilities, specifically regarding the distinct attributes of the 'Cat' and 'Possum' classes. A similar pattern of misclassification was observed between the 'Rat' and 'Mouse' classes, signifying a difficulty in distinguishing between the two. In these instances, the model frequently confused 'Mouse' with 'Rat' and the reverse, pointing to potential areas for improvement in differentiating more effectively between these species.

Regardless, through comparative analysis of existing baselines, it becomes evident that ResNet50 and other transfer learning models including DenseNet-121, EfficientNetB0, and GoogleNet, exhibit great performance in our study particularly when applied to identifying pest predators within New Zealand's natural landscapes. A study [4] conducted a similar approach to enhance image classification by leveraging transfer learning, specifically ResNet-34, as a fundamental strategy. Our findings reveal a small, improved performance boost when using ResNet50, achieving an accuracy rate of 98.15%, which outperformed the baseline accuracy of 96.43% with ResNet-34. Similarly, DenseNet-12, EfficientNetB0, and GoogleNet achieved an accuracy rate above 90%, further substantiating the effectiveness of our chosen models.

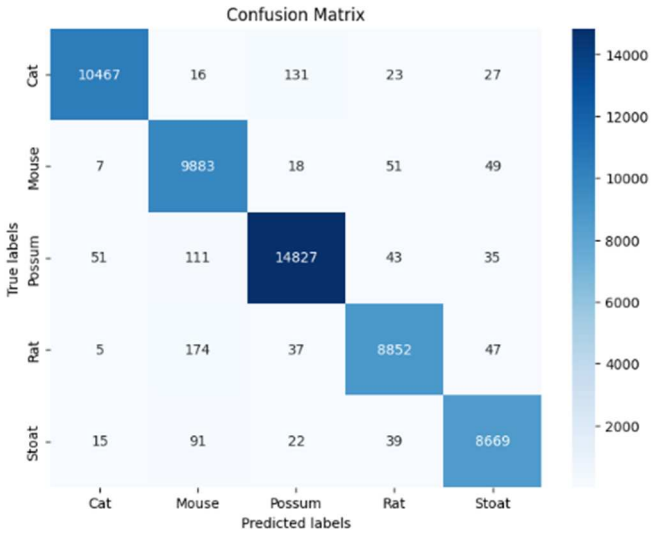

Fig 5. Confusion matrix of the highest performed model, ResNet-50.

V. CONCLUSION

In this study, we embarked on a novel approach to classify images of pest predators inhabiting the thriving landscapes of the New Zealand bush. Our primary objective was to investigate the comparative effectiveness of various cutting-edge models to combat the weaknesses of traditional conservation methods. These models include ResNet-50, DenseNet-121, and a Vision Transformer, all of which provide precise identification of target species within New Zealand's native ecosystems. To ensure a fair and consistent environment, we meticulously controlled parameters throughout our experiment with a fixed learning rate, batch size, number of epochs, and loss function. Our research hopefully contributes to the significance of harnessing deep learning models for the classification of pest predators in the New Zealand bush to increase conservation effectiveness. While most models achieved perfectly adequate performance metrics, the remarkable accuracies achieved by ResNet-50, DenseNet-121, and Vision Transformers underscore the transformative potential of advanced technologies in ecological monitoring and conservation. These models show significant promise in complementing or even replacing traditional conservation methods currently in use. As with all things, this isn't a perfect method with no drawbacks. Notably, we identified complexities in classifying the 'Cat' and 'Possum' species, along with the 'Rat' and 'Mouse' classes. These pests, sharing strikingly similar features, challenge both human and machine learning capabilities. Our future endeavours will involve an exploration of further image augmentation techniques and alternative data collection methods, including the promising avenue of thermal imaging. Additionally, we anticipate delving into object detection tasks within our current dataset and expanding our

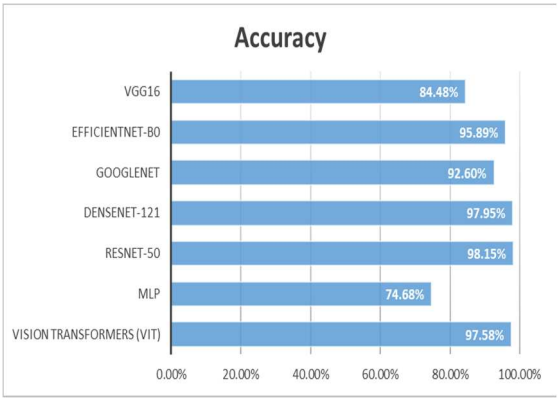

Fig 6. Comparison of model accuracies.

understanding by incorporating more diverse datasets. These extended datasets would introduce a broader array of species and environmental conditions, enriching our research. In essence, our study underlines the immense potential of integrating deep learning models as a formidable tool for ecological monitoring and conservation. These models, with their capacity for continuous improvement and fine-tuning, facilitate quick, data-driven decision-making in a field critical to preserving New Zealand's unique biodiversity and ecosystems. Looking ahead, our research journey holds the promise of reshaping how we approach conservation efforts on a global scale, ushering in an era of heightened effectiveness and sustainability.

VI. Acknowledgments

We would like to thank the team at the Palmerston North City Council (PNCC) for their support during our study.

References

[1] S. B. Islam and D. Valles, "Identification of Wild Species in Texas from Camera-trap Images using Deep Neural Network for Conservation Monitoring," 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0537-0542, doi: 10.1109/CCWC47524.2020.9031190.

[2] H. Nguyen et al., "Animal Recognition and Identification with Deep Convolutional Neural Networks for Automated Wildlife Monitoring," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 2017, pp. 40-49, doi: 10.1109/DSAA.2017.31.

[3] R. Pillai, R. Gupta, N. Sharma and R. K. Bansal, "A Deep Learning Approach for Detection and Classification of Ten Species of Monkeys," 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), Tumakuru, India, 2023, pp. 1-6, doi: 10.1109/ICSSES58299.2023.10199762.

[4] X. Han and R. Jin, "A Small Sample Image Recognition Method Based on ResNet and Transfer Learning," 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 2020, pp. 76-81, doi: 10.1109/ICCIA49625.2020.00022.

[5] J. H. Dewan et al., "Image Classification by Transfer Learning using Pre-Trained CNN Models," 2023 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI), Chennai, India, 2023, pp. 1-6, doi: 10.1109/RAEEUCCI57140.2023.10134069.

[6] Y. Qi, C. Baiyang and L. Lan, "Deep Learning Based Image Recognition In Animal Husbandry," 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2021, pp. 318-321, doi: 10.1109/ICCWAMTIP53232.2021.9674177.

[7] Pan, H., Xie, L., & Wang, Z. (2022b). Plant and animal species recognition based on dynamic Vision transformer architecture. Remote Sensing, 14(20), 5242. https://doi.org/10.3390/rs14205242

[8] Simonyan, K. (2014b, September 4). Very deep convolutional networks for Large-Scale image recognition. arXiv.org. https://arxiv.org/abs/1409.1556

[9] R. Gandhi, A. Gupta, A. K. Yadav and S. Rathee, "A Novel Approach of Object Detection using Deep Learning for Animal Safety," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 573-577, doi: 10.1109/Confluence52989.2022.9734225.

[10] N. Darapaneni, B. Krishnamurthy and A. R. Paduri, "Convolution Neural Networks: A Comparative Study for Image Classification," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 327-332, doi: 10.1109/ICIIS51140.2020.9342667.

[11] A. Dosovitskiy, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv.org, Oct. 22, 2020. https://arxiv.org/abs/2010.11929

[12] H. Talebi and P. Milanfar, "Learning to Resize Images for Computer Vision Tasks." arXiv, Aug. 17, 2021. doi: 10.48550/arXiv.2103.09950.

[13] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization Techniques in Training DNNs: Methodology, Analysis and Application." arXiv, Sep. 27, 2020. doi: 10.48550/arXiv.2009.12836.

[14] C. Khosla and B. S. Saini, "Enhancing Performance of Deep Learning Models with different Data Augmentation Techniques: A Survey," in 2020 International Conference on Intelligent Engineering and Management (ICIEM), Jun. 2020, pp. 79–85. doi: 10.1109/ICIEM48762.2020.9160048.

[15] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," ICT Express, vol. 6, no. 4, pp. 312–315, Dec. 2020, doi: 10.1016/j.icte.2020.04.010.