CNN and Vision Transformer for Classifying Basic Cat Facial Expressions

Izhar Octafirlian Susilo
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
izhar.susilo@binus.ac.id

Nicholas Andrew Marvell
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nicholas.marvell@binus.ac.id

Henry Lucky
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
henry.lucky@binus.ac.id

Derwin Suhartono
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
dsuhartono@binus.edu

***Abstract*—This study assesses how well two deep learning models—ResNet18 and MobileViT—perform in identifying the four cat facial emotions of calm, angry, alarmed, and pleased. Two partitioning strategies—the original dataset split and a manually curated split—were used to train and evaluate the models on a small dataset. The results show that, especially when using the manually split data, ResNet18 continuously outperformed MobileViT in terms of accuracy. These results imply that CNN-based models perform better in situations involving tiny amounts of data. The findings have potential uses in domains like behavioral analysis, pet care, and veterinary diagnostics.**

Keywords—CNN, Vision Transformer, ResNet18, MobileViT, cat expression

I. INTRODUCTION

The recognition of facial expressions in animals, particularly cats, has emerged as a growing area of interest in the field of Artificial Intelligence (AI). Accurately interpreting feline emotions has potential benefits in various domains, including pet care, veterinary diagnostics, and animal welfare **[1]** Despite this potential, identifying cat facial expressions remains a challenge due to the presence of unique fur patterns and subtle, often indistinct, emotional cues that can be difficult to detect.

To identify their expressions, image recognition is applied; the most common method used is Convolutional Neural Networks (CNNs), which are the most widely implemented. Previous studies **[2] [3]** have shown that CNNs can successfully classify a cat's facial expressions, and their accuracy can be improved with more training. The studies utilized CNN models trained on the ImageNet and Stanford Dogs datasets, achieving high accuracy in pet emotion recognition; however, performance was inconsistent across different breeds. Another study **[4]** examined 1000 cat facial images and found that CNN models become more accurate with repeated training sessions. Additionally, studies **[5]** showed that CNN-based landmark models analyzing cat facial landmarks achieved higher accuracy compared to texture-based methods. However, CNNs mainly focus on small details, such as edges and textures, which may not be sufficient to fully understand the cat's emotions.

However, a new method called Vision Transformers (VITs) has recently gained popularity. VITs look at an image differently; they do this by breaking the image into smaller parts and understanding how these parts relate to each other. Research studies in **[6][7]** showed that ViTs are good at recognizing emotions, especially in complex images. The studies compared the results of CNN and ViT. They found that, when comparing CNN and ViT-based models on facial expression datasets (JAFFE, CK+), ViTs achieved higher accuracy in recognizing nuanced facial expressions. Moreover, studies **[8]** have shown that CNNs tend to perform better on small datasets, whereas ViTs excel with larger datasets. Despite this, limited research has been conducted on comparing CNN and ViT specifically for recognizing animal expressions, particularly in cats. Additionally, the impact of dataset size on performance differences between CNN and ViT remains unclear.

This research aims to compare CNN models (ResNet18) and ViT models (MobileVIT) for classifying basic cat facial expressions. Between those two methods, a comparison of their performance on a small-scale dataset will be made, particularly considering the impact of different dataset-splitting methodologies. By making this comparison, the approach that is more effective in recognizing cat facial expressions and how dataset size influences their accuracy will be determined. The results of this research can contribute to AI to improve enhancing the understanding of cat emotions, which can benefit both pet owners and veterinarians.

II. RELATED WORKS

Emotion recognition using artificial intelligence has been studied widely including on human facial expression and animal expressions. One of the methods often used is Convolutional Neural Networks (CNN). CNN is considered the traditional method, however there is a new and upcoming method called Vision Transformers (ViTs). CNN works by processing an image through several layers, while ViTs process them by splitting them into patches. Researchers are shifting to ViTs as Vision Transformers offer greater ability to capture global relationships within an image.

A study using CNN **[2]** utilizes Faster-RCNN. Faster-RCNN differs from CNN as it detects regions of interest before classifying emotions which optimizes its ability to analyze both facial expression and body language in pets. The study trained its model on various datasets which enabled the model to differentiate happiness, fear, and neutrality. The study specifically underlined the difficulty in emotion classification as there were a lot of differences in facial anatomy and non-uniform emotional cues across species. The study showed that Faster-RCNN successfully classified both pet species and emotion in a high accuracy..

When using CNN, a study **[3]**, utilizes CNN features to unravel the intricate language embedded in the facial expressions of cats. The study used a dataset with a total of 1000 photos for training and 300 for validation. The study shows that the model's accuracy improves every iteration (Epoch) done. Epoch represents one complete pass of the training dataset through the model which allows it to refine its learning over time. It shows that for the first iteration, the accuracy is only at 55.56%, but over time, it improved and eventually reached 95.71% by the 10th iteration. Similary **[4]** demonstrated that CNNs maintained reliable performance in facial emotion recognition (FER) tasks, further supporting their effectiveness.

Another study **[5]** also uses CNN in recognizing domestic cats' facial expression, with a result of 98% overall. While the RAF-DB, FER2013, a recently balanced dataset were used in the study **[6]** to assess thirteen Vision Transformer (ViT) models on facial emotion recognition tasks. According to the results, the most successful models were Mobile ViT and Tokens-to-Token ViT, which were followed by PiT and Cross Former. The study demonstrates ViT's remarkable ability to identify emotions in complicated and chaotic images.

For ViT (Vision Transformers) that has been gaining attention to handle complex datasets, there has beena study **[7]**, that explains how ViT can recognize facial expressions, unlike the CNN which they process an image by scanning the local patterns, but with ViT they break images into patches and use self-attention to figure out relationships between different parts of the image. This allows them to pick up on subtle emotional details, even in noisy or complicated pictures. The study found that ViTs did particularly well at recognizing emotions that aren't super obvious or require understanding the full context of the image

But how to compare CNN and ViT when it comes to human expression or animal expression? There's this study **[8]**, where they look at the strengths and weaknesses between those 2 methods, now CNN is great at picking up local features like the edges, textures, or small details which makes them more effective for smaller datasets. ViT is better at capturing the bigger picture and handling large datasets which makes them a strong option for something like this research paper which is about Classifying Basic Cat Facial Expressions. But still CNN works better when there is a limited dataset and easier to train.

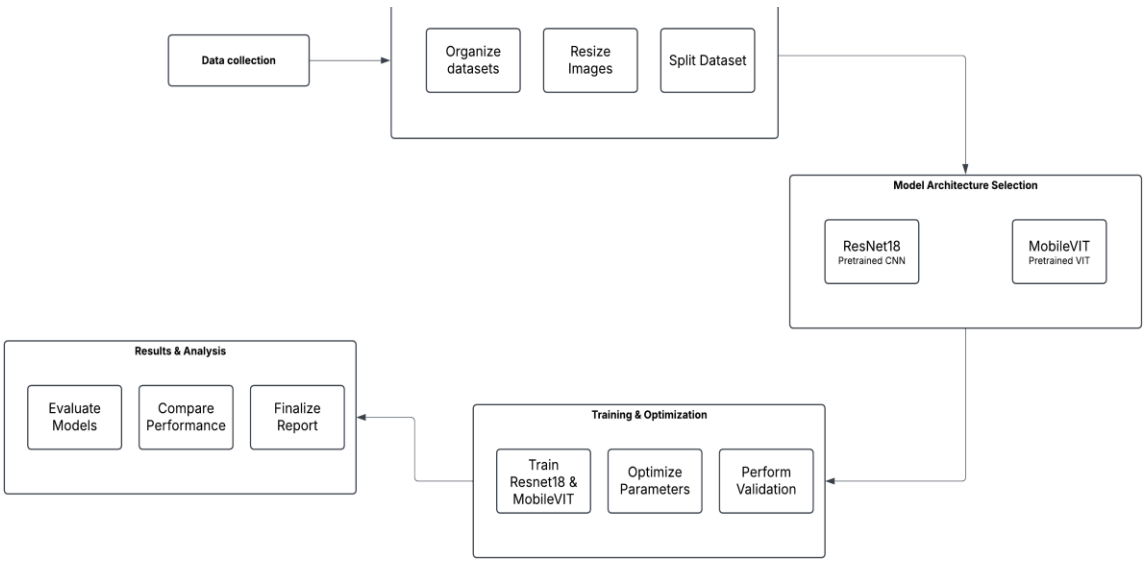
Besides, **[9]** created a hybrid approach that combines CNNs with the Improved Whale Optimization Algorithm (IWOA) to detect emotions in dogs. The results achieved noteworthy enhancements over baseline CNNs, revealing that optimization techniques can offer enhanced deep learning performance for similar classification tasks.

III. METHODOLOGY

A. Research Design

The performance of two deep learning models, ResNet18 (CNN) and MobileViT (Vision Transformer), in categorizing cat facial expressions is assessed using a quantitative, comparative design. Previous studies on emotion recognition in cats and dogs have utilized classification and object detection models [10], which helped inform the selection of model architectures. The main objective is to determine which model provides the highest accuracy when classifying cat images into four emotional states: Pleased, Angry, Alarmed, and Calm. The experimental workflow is illustrated in Figure 1, based on the model architecture selection referenced in [11].

Figure 1. Experimental workflow for classifying cat facial expressions using ResNet18 and MobileViT models.



B. Dataset

The dataset will be sourced from the "Domestic Cats Facial Expressions Computer Vision Project" on Roboflow Universe, containing approximately 1,348 pet images. Furthermore, all images in the dataset will be used. In addition, every single image will be manually labelled with one of the four target emotions: "Pleased," "Angry," "Alarmed," or "Calm." For dataset partitioning, two types of splitting will be applied. The first is the original split that has already been implemented in the dataset, and the second will be performed manually to suit specific experimental needs.

The original labelled dataset is split into three subsets: the Training Set, the Validation Set, and the remaining 3% for the Test Set. However, the manually labelled dataset will be partitioned into three subsets: a Training Set (70%) and both Validation and Test Sets (15%).

Justification for Using Two Splits:

The original dataset split (89% training, 8% validation, 3% test) provided by the dataset creator was used to evaluate baseline model performance using the default configuration. Creating a manually split version (70/15/15) will improve class distribution and gain more control over evaluation size.

This dual-split approach enables comparison between the provided split and a custom-designed split, providing insights into how data partitioning affects model performance, particularly on small-scale datasets.

C. Model Architecture

A pre-trained ResNet18 model will be used, and the final fully connected layer will be modified to output four classes. The MobileViT architecture will be employed, utilizing pre-trained weights if available, with the classification head adjusted for four classes.

Model Selection Justification:
ResNet18 was chosen as the CNN baseline due to its proven performance on small-scale datasets, its relatively low computational cost, and its ability to generalize well, even with limited data. Its residual connections help mitigate vanishing gradients, which improves training stability.

MobileViT was chosen to evaluate the potential of Vision Transformers in low-resource settings. Its design aligns with prior research on small-scale datasets [12] and has demonstrated promising performance in constrained environments [15]. MobileViT combines global attention mechanisms from Transformers with local feature modelling similar to CNNs, as introduced in lightweight architectures such as Token-to-Token ViT [17]. This hybrid approach enables effective extraction of emotional context even with limited data.

B. Training Procedure

The models will be implemented using PyTorch, and pre-trained weights will be loaded to leverage transfer learning. For optimization, we will use the Adam optimizer for both models, and cross-entropy loss will also be loaded to leverage transfer learning. Both models will then be trained on the training set for a fixed number of epochs, which is 10. Monitor validation loss is used to prevent overfitting by using early stopping. The results that we will use to evaluate the models' performance are Accuracy, Precision, Recall, and F1-score.

C. Analysis

1. Comparative Performance: Compare evaluation metrics for ResNet18 and MobileViT.
2. Statistical testing was not conducted due to the limited dataset size.
3. Qualitative Analysis: Analyze confusion matrices to identify which emotion categories are most challenging for each model.

IV. RESULTS AND DISCUSSIONS

A. Overview

Deep learning models, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have demonstrated effectiveness in various image classification and segmentation tasks [8]. These architectures are increasingly applied in specialized domains, including animal facial expression recognition.

This experiment compared two such models—ResNet18 (CNN) and MobileViT (ViT)—in classifying four basic cat facial expressions: Pleased, Angry, Alarmed, and Calm. The dataset was processed using two different splitting strategies: the original (89/8/3) and a manually curated version (70/15/15), to observe the impact of partitioning on model performance. All training was conducted using Google Colaboratory on a T4 GPU [12].

Figure 2. Accuracy comparison of ResNet18 and MobileViT across four emotional categories using both original and manual dataset splits.

Model	**Original Splitting**				
	Angry	**Alarmed**	**Calm**	**Pleased**	**Overall Accuracy**
ResNet18	0.900	0.630	0.440	0.500	0.690
MobileViT	0.870	0.500	0.600	0.670	0.690
	Manual Splitting				
	Angry	**Alarmed**	**Calm**	**Pleased**	**Overall Accuracy**
ResNet18	0.880	0.760	0.730	0.900	0.830
MobileViT	0.670	0.460	0.610	0.850	0.700

B. Model Performance Comparison

As shown in Figure 2, ResNet18 outperformed MobileViT in most categories using the manually split dataset, achieving an overall accuracy of 83%, compared to 69% with the original split. Specific improvements were observed in Alarmed (63% → 76%), Calm (44% → 73%), and Pleased (50% → 90%), with only a minor drop in Angry (−2%).

MobileViT, while more sensitive to dataset configuration, showed promising results in some cases. In the manual split, its accuracy for Pleased increased from 67% to 85%, though Angry dropped from 87% to 67% [13]. This inconsistency reflects the model's reliance on both dataset quality and volume.

Overall, ResNet18 displayed stronger generalization across all emotional categories under limited data conditions.

C. Analysis and Implications

Understanding cat facial expressions is inherently difficult due to subtle fur pattern variations and non-obvious emotional cues [14]. From the findings, ResNet18's architecture made it more stable and effective with a small dataset, while MobileViT required more data to converge effectively. Looking ahead, one promising direction is to explore hybrid architectures, combining CNNs' ability to extract local features with ViTs' capacity for global attention [16].

Observations also support prior research indicating that ViTs, while powerful, need more data and training to reach their potential [18]. In contrast, CNNs remain more efficient and easier to train in data-constrained environments [19].

V. CHALLENGES & FUTURE WORKS

A major challenge encountered was data imbalance, particularly in underrepresented classes such as Pleased. This imbalance skewed learning outcomes and reduced accuracy in those categories. Another constraint was the limited overall dataset size, which negatively impacted the performance of Vision Transformer models, confirming their dependency on large-scale datasets for optimal training. In future work, we plan to expand and rebalance the dataset, as well as include more emotional classes for deeper analysis. Additionally, further exploration of Transformer-based models and hybrid architecture may lead to more robust performance under real-world constraints.

VI. CONCLUSION

This study evaluated the performance of ResNet18 and MobileViT for classifying basic cat facial expressions using two different dataset partitioning strategies. The results show that ResNet18 consistently outperformed MobileViT, especially when using the manually split dataset. These findings reinforce that CNNs are more effective in small-data scenarios, while ViTs may be better suited for longer training and large-scale datasets [20].

In conclusion, ResNet18 remains the more practical model for classifying cat facial expressions when data is limited.

REFERENCES

[1] E. M. C. Bouma, M. L. Reijgwart, P. Martens, and A. Dijkstra, "Cat owners' anthropomorphic perceptions of feline emotions and interpretation of photographs," Applied Animal Behaviour Science, vol. 270, p. 106150, Dec. 2023, doi: 10.1016/j.applanim.2023.106150.

[2] R. O. Sinnott, U. Aickelin, Y. Jia, E. R. J. Sinnott, P.-Y. Sun, and R. Susanto, "Run or pat: using deep learning to classify the species type and emotion of pets," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), vol. 2, pp. 1–6, Dec. 2021, doi: 10.1109/csde53843.2021.9718465.

[3] A. K. Pandey, P. Jain, B. Yadav, and V. Pandey, "Facial emotion recognition of cat breeds by using convolution neural network," IRE Journals, Dec. 01, 2023.

[4] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," Procedia Computer Science, vol. 235, pp. 2079–2089, Jan. 2024, doi: 10.1016/j.procs.2024.04.197.

[5] N. A. Ali, C. L. N. O. Oyana, and O. S. Salum, "Domestic cats facial expression recognition based on convolutional neural networks," International Journal of Engineering and Advanced Technology, vol. 13, no. 5, pp. 45–52, Jun. 2024, doi: 10.35940/ijeat.e4484.13050624.

[6] S. Bobojanov, B. M. Kim, M. Arabboev, and S. Begmatov, "Comparative analysis of vision transformer models for facial emotion recognition using augmented balanced datasets," Applied Sciences, vol. 13, no. 22, p. 12271, Nov. 2023, doi: 10.3390/app132212271.

[7] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial Emotion Recognition with Vision Transformers," Applied System Innovation, vol. 5, no. 4, p. 80, Aug. 2022, doi: 10.3390/asi5040080.

[8] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision Transformers and convolutional neural Networks for image classification: A literature review," Applied Sciences, vol. 13, no. 9, p. 5521, Apr. 2023, doi: 10.3390/app13095521.

[9] Y. Mao and Y. Liu, "Pet dog facial expression recognition based on convolutional neural network and improved whale optimization algorithm," Scientific Reports, vol. 13, no. 1, Feb. 2023, doi: 10.1038/s41598-023-30442-0.

[10] Y. Wu, Emotion Detection of Dogs and Cats Using Classification Models and Object Detection Model. ResearchGate, May 2023.

[11] V. A. Saputra, M. S. Devi, N. Diana, and A. Kurniawan, "Comparative analysis of convolutional neural networks and vision transformers for dermatological image classification," Procedia Computer Science, vol. 245, pp. 879–888, Jan. 2024, doi: 10.1016/j.procs.2024.10.315.

[12] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using Deep Learning: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1, Jan. 2021, doi: 10.1109/tpami.2021.3059968.

[13] M. Yan, "Advancements in Image Recognition: Comparing CNNs and vision Transformers," Applied and Computational Engineering, vol. 104, no. 1, pp. 143–149, Nov. 2024, doi: 10.54254/2755-2721/104/20241205.

[14] M. Feighelstein et al., "Explainable automated pain recognition in cats," Scientific Reports, vol. 13, no. 1, Jun. 2023, doi: 10.1038/s41598-023-35846-6.

[15] A. Vatcharaphrueksadee, M. Maliyaem, and P. Sawakchart, "Hybrid models for facial emotion recognition and intensity detection: generalization across human and cartoon faces using CNN and Vision Transformer," Journal of Advances in Information Technology, vol. 16, no. 4, pp. 478–490, Jan. 2025, doi: 10.12720/jait.16.4.478-490.

[16] S. Mehta and M. Rastegari, "MobileVIT: light-weight, general-purpose, and mobile-friendly vision

transformer," arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2110.02178.

[17] S. Shaees, H. Naeem, M. Arslan, M. R. Naeem, S. H. Ali, and H. Aldabbas, "Facial emotion recognition using transfer learning," 2020 International Conference on Computing and Information Technology (ICCIT-1441), pp. 1–5, Sep. 2020, doi: 10.1109/iccit-144147971.2020.9213757.

[18] L. Yuan et al., "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 538–547, Oct. 2021, doi: 10.1109/iccv48922.2021.00060.

[19] R. H. Frye and D. C. Wilson, "Comparative analysis of transformers to support Fine-Grained Emotion Detection in Short-Text data," Proceedings of the ... International Florida Artificial Intelligence Research Society Conference, vol. 35, May 2022, doi: 10.32473/flairs.v35i.130612.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," arXiv (Cornell University), Jan. 2020, doi: 10.48550/arxiv.2012.12877.