JURNAL EMACS e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.7 No.3 September 2025: 321-326 DOI: 10.21512/emacsjournal.v7i3.14317

Comparative Study of CNN-based Deep Learning Models for Animal, Digit, and Flower Image Classification

Puti Andam Suri[1*], Michael Alvin Setiono[2], Andrew[3], Muhammad Fajar[4]

[1-4] Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta Indonesia, 11480
puti.suri@binus.ac.id, michael.setiono001@binus.ac.id, andrew038@binus.ac.id,
muhamad.fajar@binus.edu

*Correspondence: puti.suri@binus.ac.id

Abstract – *This study explores how four convolutional neural network (CNN) models MobileNetV2, DenseNet121, EfficientNetB0, and InceptionV3 perform in classifying images from three different datasets: animals, handwritten digits (MNIST), and flowers. The main goal is to understand which model offers the best balance between accuracy and efficiency when applied to datasets with varying complexity. Each model was trained and tested using identical preprocessing steps, and its performance was evaluated based on accuracy, precision, recall, and F1-score through a confusion matrix. Training and testing times were also measured to assess computational efficiency. The results show that DenseNet121 consistently achieved the highest accuracy: 98% on animal images and 88% on flower images, while MobileNetV2 provided a close performance (97% and 82%) but with much faster processing times, between 11 and 55 minutes. EfficientNetB0, on the other hand, performed poorly on the more complex flower dataset, achieving only 5% accuracy. These findings suggest that DenseNet121 is ideal for projects where accuracy is the main concern, whereas MobileNetV2 is more suitable for real-time applications that require quick responses without a major drop in accuracy. Overall, this research highlights the importance of aligning model selection with both dataset characteristics and computational limitations in practical image classification tasks.*

Keywords: *Image*; *Classification; Deep Learning*

I. INTRODUCTION

As time progresses, AI technology has become more advanced and apparent in this technological driven world (Al-Saffar et al., 2017); (Alshazly et al., 2019); (Alzubaidi et al., 2021). This change affects us in many ways, from how we study, work, and even live. Many applications that are our daily driver have begun to utilize AI to enhance their performance, suit each individual user better, and do stuff that would likely be impossible without artificial intelligence. In computer vision, specifically image classification. Deep learning is a popular choice as it is very flexible, can handle large amounts of data, and understand complex features which frequently occur in image datasets (Basak et al., 2021). Studies show that image recognition using deep learning is more accurate than traditional handcrafted computer vision (Comber et al., 2012). Due to its high interest and popularity, many people have started to make different kinds of deep learning algorithms and architectures to produce better classification results, which in turn causes the existence of many kinds of deep learning models. While each model has its own strengths and weaknesses, we are interested in testing and comparing these models' performance, specifically in how they perform in image classification (Dutta et al., 2017); (Eli-Chukwu, 2019).

Image classification is an important field in computer vision, it acts as a way for us to process visual data and allow technologies such as medical diagnosis or autonomous media censorship. A study was conducted on using deep learning to generate steering instructions

for self-driving cars based on event-based image vision (Hatcher & Yu, 2018). Another research conducted by Obaid et al., stated that deep learning models for image classification has made many remarkable achievements in many large-scale identification tasks in the field of computer vision (Obaid et al., 2020). Considering its high importance and various benefits, aiding the development of image classification technology would help humanity and may allow other more beneficial technologies to be possible in the future. We hope that from our paper, readers who are developing or working with image classification technologies can gain insights and choose the right model which best fits their own use case.

There have been previous studies that also compare deep learning models' performance to one another. One paper analyzed the performance of different deep learning models for medical image classification. In it, the study compares advanced CNN (Convolutional Neural Networks) with DNN (Deep Neural Networks) to diagnose two diseases, Diabetic retinopathy which is a sight related disease caused by diabetes, and Emphysema, which is a disease characterized by the loss of tissues in the lung. The dataset used in the paper includes CT images of the human lung and images of the eye captured with a FUNDUS camera (Kamel, 2024).

Different from the paper mentioned above, our paper's aim is to test each models' performance more generally. Each of the deep learning models will be trained and tested using different kinds of datasets. The test results of each model will then be compared to one another to find the most efficient and accurate model. In this study, the performance of each model will determine which of deep learning architecture performs better, timewise and accuracy wise. To measure each model's accuracy, the metric that will be used is its accuracy, namely by confusion matrix. As for its efficiency or speed, the metric that will be measured is the model's training and testing time.

Despite the rapid progress in deep learning research, there remains a lack of comparative studies that examine CNN architectures across datasets with different complexity levels and data characteristics. Most existing studies focus on domain-specific datasets such as medical or industrial imagery, which limits the generalizability of their findings. This research aims to fill that gap by conducting a broader comparison using three diverse datasets: animals, handwritten digits, and flowers to represent varying levels of visual complexity. The main objective is to determine which CNN architecture offers the best trade-off between accuracy and computational efficiency, providing a practical reference for selecting appropriate models in real-world image classification applications.

According to Al-Saffar et al. (2017) deep learning is a very popular research direction where tasks like image classification and object detection have very big results and progress but on the other hand, there are still many potentials that can be accomplished by deep learning. During the process of training, the module itself can autonomously learn the parameters required for spatial transformation and does not need to add any additional supervision during the training, therefore many people have used deep learning in the field of image classification with great results (Khamparia & Singh, 2019).

To compare the accuracy between the four deep learning architectures that have been chosen, each model's performance will be mapped using the confusion matrix metric. In a confusion matrix there are 4 terms used: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These 4 terms each measure how many times the model correctly and incorrectly guesses if a data belongs to certain category. These 4 terms are also used to calculate a few metrics that can be used to determine a model's performance (Loureiro et al., 2018); (Maqueda et al., 2020); (Obaid et al., 2020).

Research conducted by Hemdan et al., used some deep learning classifiers such as VGG19, DenseNet21, InceptionV3, ResNetV2, Inception-ResNet-V2, Xception, and MobileNetV2 (Hemdan et al., 2020). The workflow starts by preprocessing all images that have been collected, and then followed by training the model and validation. Lastly the model is tested, and the overall performance is evaluated. The performance can be analyzed by using metrics inside the confusion matrix to compute the accuracy, precision, recall, and f1-score of each model. In conclusion, the best

performance scores of deep learning classifiers are VGG19 and DenseNet21 (Pommé et al., 2020).

Recent studies have explored more efficient or hybrid architectures in image classification. For example, Tan & Le (2021) proposed EfficientNetV2, which achieves faster training and better parameter efficiency (Zhao et al., 2021) conducted an empirical comparison among CNN, Transformer, and MLP architectures, revealing their respective strengths under different scales. Other works, such as MLP-Mixer (Tolstikhin et al., 2021) and AS-MLP (Lian et al., 2021), push the boundaries of pure MLP models in vision tasks.

In total, this research paper contains 6 sections. Section 1 explains about the research problems of some deep learning models for image classification. Section 2 describes the deep learning models that will be used for the accuracy and speed test. Section 3 explains the dataset that will be used for the models. Section 4 explains the evaluation method that is used for getting the results after testing the models. Section 5 will show the results based on the test with the dataset. Section 6 will summarize all the aforementioned results.

II. METHODS

2.1 Dataset

There are 3 different kinds of data se being used to test the models with various situations. The first dataset includes 4 different kinds of animals with each class containing 1000 images. The second dataset is the MNIST-digit handwriting image dataset containing 10 classes sampled 2000 images for each digit. And the third dataset contains 16 different kinds of flowers sampled 500 images for each flower.

2.2 Research Approach

In this research, the approach used is quantitative where an experiment will be carried out to train the deep learning models for each dataset based on the architecture that has been chosen. After training the models, their accuracy and training time elapsed will be measured and compared with other models.

2.3 Architecture

There are many deep learning architectures that are readily available to be used in open-source libraries such as Pytorch and Tensorflow. In this research the architecture used will include DenseNet121, EfficientNetB0, InceptionV3, and MobileNetV2 taken from the Tensorflow 2.11.0 library.

2.4 Preprocessing

To make sure each model is treated fairly, the dataset used for each model's training will be the same. To prepare the dataset each image's features will be normalized and resized to 224x224. For images that are grayscale by default, an additional step is required to ensure compatibility with models that expect color images with explicit color channels. This is done by adding an extra dimension to represent the color channel. Then the images are split for the models' training and testing dataset. After that, the training data will be used to train the models, and the testing data is used to validate whether the models are overfitted or not by testing it with images outside of its training data.

2.5 Evaluation Method

After training the models using the preprocessed dataset, the confusion matrix will be used to determine the accuracy, precision, recall, specifity, and F1-score of each model.

III. RESULTS AND DISCUSSION

After conducting experiments, the four different deep learning models performance that have been trained on three different image datasets are compared, observed, and analyzed.

3.1 Model Performance Metrics

The accuracy of each model is evaluated using metrics in confusion matrix which are accuracy, precision, recall, and F1-score. Meanwhile the efficiency of each model is measured by how long it took to train and test the model's performance.

Table 1. Performance comparison of CNN models on the Animal dataset.

Model	Mobile NetV2	Dense Net121	Efficient NetB0	Inception V3
Accuracy	0.97	0.98	0.26	0.98
Precision	0.97	0.98	0.06	0.98
Recall	0.97	0.98	0.25	0.98
F1-Score	0.97	0.98	0.10	0.98
Training Time	11m 40.8s	37m 7.8s	19m 57.1s	14m 32.1s

Testing Time	20.6s	1m 15.5s	37.3s	30s

Table 2. Performance comparison of CNN models on the MNIST-digit dataset.

Model	Mobile NetV2	Dense Net121	Efficient NetB0	Inception V3
Accuracy	0.92	0.89	0.91	0.84
Precision	0.93	0.90	0.92	0.86
Recall	0.92	0.89	0.91	0.84
F1-Score	0.92	0.89	0.91	0.84
Training Time	55m 57.2s	174m 30.2s	98m 47.5s	68m 17.4s
Testing Time	1m 29.9s	5m 40.1s	2m 41.2s	2m 15.3s

Table 3. Performance comparison of CNN models on the Flower dataset.

Model	Mobile NetV2	Dense Net121	Efficient NetB0	Inception V3
Accuracy	0.82	0.88	0.05	0.79
Precision	0.85	0.89	0.00	0.80
Recall	0.82	0.88	0.06	0.78
F1-Score	0.82	0.88	0.01	0.78
Training Time	23m 21.1s	76m 0.5s	39m 10.4s	32m 54.5s
Testing Time	39.8s	2m 22.1s	1m 10.9s	1m 0.9s

3.2 Observation and Analysis

3.2.1 Model Performance

Based on the Model Performance Metrics table above, the performance of each model varies across different datasets.

In Table 1, DenseNet121 and InceptionV3 have the highest accuracy of 98% then closely followed by MobileNetV2 at 97% and EfficientNetB0 has the lowest accuracy with 26%.

In Table 2, MobileNetV2 has the best results with an accuracy of 92%, closely followed by EfficientNetB0 at 91%, DenseNet121 at 89%, and lastly InceptionV3 84%.

Lastly for Table 3, DenseNet121 has the highest accuracy of 88% and then followed by MobileNetV2 at 82% and InceptionV3 at 79%. EfficientNetB0 has the lowest accuracy across 3 experiments at only 5% for the flower dataset.

3.2.2 Training and Testing Time

Based on the experiments conducted, an underlying pattern can be found within each model's efficiency across all datasets. MobileNetV2 is observed to be the most efficient among all the models with the fastest training and testing time, and then closely followed by InceptionV3. This can be attributed due to these two models being lightweight hence having similar and relatively fast times.

EfficientNetB0 places third based on its time taken for training and testing, taking an average of twice as long when compared with MobileNetV2. DenseNet121 on the other hand while consistently showing high accuracy across all datasets is also the heaviest, with its training time consistently taking almost 3 times as long as MobileNetV2's.

3.2.3 Dataset Characteristics

The three datasets used for this experiment test the models to various possible characteristics of data. The animal dataset is a simple dataset having only 4 classes and each class having 1000 images while also having the highest resolution among the other datasets.

The MNIST-digit dataset is more complex than the animal dataset, having 10 classes and each class having 2000 images. This dataset is the largest among the others, even with the lower resolution it still takes the largest amount of computational resource shown by the longer training and testing times.

Lastly the flower dataset is the most complex, having 16 classes and each class containing 500 images. This dataset is the most challenging for the models to classify which can be seen from the lowest average accuracy across all datasets.

IV. CONCLUSION

In conclusion, DenseNet121 and MobileNetV2 produced the highest average accuracy across all the datasets, making them a good choice for image classification tasks with different dataset characteristics. Although DenseNet121 achieved the highest average accuracy, it has a downside of requiring many computational resources where the training time is long compared to other models used in this research. Therefore, MobileNetV2 is recommended for situations where the need for high efficiency is present while keeping a high accuracy and DenseNet121 is recommended where computation resource is not a problem and achieving a high accuracy is the top priority.

EfficientNetB0 has the lowest average accuracy amongst 4 of the models where the

lowest accuracy of this model is 5%. This is likely because EfficientB0 is the first version of the EfficientNet architecture with limited pattern learning capability hence it couldn't adapt well with complex datasets which is shown when tested with the flower dataset which contains 16 different classes. EfficientNetB4 was intended for use at the start of the research, but due to the conflicting input shape, EfficientNetB0 was chosen instead.

This research still has its flaws and weaknesses, one of which is that each model was trained using datasets with the same data preprocessing. While this ensures that each model is treated the same, another perspective can be taken which is that each model has different architectures and hence needs different preprocessing methods to perform well. In a real case scenario, different data preprocessing techniques should be tested and the best perfoming ones should be chosen to ensure the model producing the most accurate results.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request and can be accessed through this link:

1. Animals: https://www.kaggle.com/datasets/ayushv322/animal-classification
2. MNIST: https://www.kaggle.com/datasets/hojjatk/mnist-dataset
3. Flowers: https://www.kaggle.com/datasets/l3llff/flowers

Author Contribution Statement

Puti Andam Suri: Research Lead and Paper Enhancement, Michael Alvin Setiono: Model Creation and writing, Andrew: Model Creation and writing, Muhamad Fajar: Advisor and paper quality control.

REFERENCES

Al-Saffar, A. A. M., Tao, H., & Talab, M. A. (2017). Review of deep convolution neural network in image classification. Proceedings of the 2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 26–31. https://doi.org/10.1109/ICRAMET.2017.8253139

Alshazly, H., Linse, C., Barth, E., & Martinetz, T. (2019). Handcrafted versus CNN features for ear recognition. Symmetry, 11(12), 1493. Retrieved April 21, 2024, from https://www.mdpi.com/2073-8994/11/12/1493

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 1–74. https://doi.org/10.1186/s40537-021-00444-8

Basak, H., Kundu, R., Chakraborty, S., & Das, N. (2021). Cervical cytology classification using PCA and GWO enhanced deep features selection. SN Computer Science, 2(5). https://doi.org/10.1007/s42979-021-00741-2

Comber, A., Fisher, P., Brunsdon, C., & Khmag, A. (2012). Spatial analysis of remote sensing image classification accuracy. Remote Sensing of Environment, 127, 80–89. https://doi.org/10.1016/j.rse.2012.09.005

Dutta, S., Manideep, B., Rai, S., & Vijayarajan, V. (2017). A comparative study of deep learning models for medical image classification. IOP Conference Series: Materials Science and Engineering, 263(4), 042097. https://doi.org/10.1088/1757-899X/263/4/042097

Eli-Chukwu, N. C. (2019). Applications of artificial intelligence in agriculture: A review. Technology & Applied Science Research, 9(4), 4377–4383. Retrieved May 30, 2024, from www.etasr.com

Hatcher, W. G., & Yu, W. (2018). A survey of deep learning: Platforms, applications and emerging research trends. IEEE Access, 6, 24411–24432. https://doi.org/10.1109/ACCESS.2018.2830661

Hemdan, E. E.-D., Shouman, M. A., & Karar, M. E. (2020). COVIDX-Net: A framework

of deep learning classifiers to diagnose COVID-19 in X-ray images. arXiv Preprint. Retrieved April 4, 2024, from http://arxiv.org/abs/2003.11055

Kamel, I. (2024). Artificial intelligence in medicine. Journal of Medical Artificial Intelligence, 7(0), 4. https://doi.org/10.21037/jmai-24-12

Khamparia, A., & Singh, K. M. (2019). A systematic review on deep learning architectures and applications. Expert Systems, 36(3), e12400. https://doi.org/10.1111/exsy.12400

Lian, D., Yu, Z., Sun, X., & Gao, S. (2021). AS-MLP: An Axial Shifted MLP Architecture for Vision. arXiv.

Loureiro, S. M. C., Guerreiro, J., & Tussyadiah, I. (2021). Artificial intelligence in business: State of the art and future research agenda. Journal of Business Research, 129, 911–926. https://doi.org/10.1016/j.jbusres.2020.11.001

Maqueda, A. I., Loquercio, A., Gallego, G., García, N., & Scaramuzza, D. (2018). Event-based vision meets deep learning on steering prediction for self-driving cars. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Retrieved April 22, 2024, from http://openaccess.thecvf.com/content_cvpr_2018/html/Maqueda_Event-Based_Vision_Meets_CVPR_2018_paper.html

Obaid, K., Zeebaree, S., & Ahmed, O. M. (2020). Deep learning models based on image classification: A review. International Journal of Science and Business. Retrieved April 22, 2024, from https://www.academia.edu/download/64726148/612.pdf

Pommé, L., Bourqui, R., Giot, R., & Auber, D. (2022). Relative confusion matrix: Efficient comparison of decision models. Proceedings of the 2022 26th International Conference on Information Visualisation (IV). https://doi.org/10.1109/IV56949.2022.0002

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller Models and Faster Training. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021). arXiv

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., … Uszkoreit, J. (2021). MLP-Mixer: An all-MLP Architecture for Vision. arXiv.

Zhao, Y., Wang, G., Tang, C., Luo, C., Zeng, W., & Zha, Z.-J. (2021). A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP. arXiv.