A Comparative Analysis of Dog Emotion Classification Based on Four CNN Architectures

Yutong Liu [1], Bowei Shi [2, *], Heming Sun [3]

[1] Vermont Academy, Vermont, United States

[2] Notre Dame High School, Calgary, Canada

[3] Beijing Etown Academy, Beijing, China

* Corresponding Author Email: *boweis1@learn.cssd.ab.ca*

Abstract. This study was research on the validation accuracy of different convolutional neural network (CNN) architectures in identifying the dog's emotions based on their facial images. The goal is to recognize which model can best verify an animal's emotion. The four CNN models are: a custom-designed Original CNN (OCNN), MobileNetV2, EfficientNetB3, and EfficientNetB7. The experiment used a labeled dataset from Kaggle, which contains 4,000 dog facial images with four emotional states: happy, angry, sad, and relaxed, to train and preprocess the model. The research is using accuracy, precision, recall, and F1-score to evaluate the performance of the model. It was shown that EfficientNetB3 had the highest F1-score of 0.72, showcasing better model complexity and performance. In contrast, OCNN was underfitting, while EfficientNetB7 showed overfitting. This experiment highlights the importance of model choice in emotion classification tasks and provides some new views for the development of emotional computing systems in animal welfare.

Keywords: Emotion Classification in Dogs, Transfer Learning, Efficientnet, Mobilenet, Comparative Experiments.

1. Introduction

With the increasing population of society and its advances, nowadays, pets have gradually become an important part of human life. According to the newest American Pet Products Association 2025 stats, 94 million U.S. households (71%) have a pet, up from 82 million in 2023, and the pet industry continues to demonstrate growth and resilience, with total expenditures in the U.S. pet industry projected to reach $152 billion in 2024 [1]. Showcasing pets is becoming more and more important in human life. Among different kinds of pets, dogs are one of the most popular animals which have a huge number. For instance, relevant data from the 2018 U.S. General Social Survey indicate that 46.1% of the population owned a dog, far exceeding the proportions owning cats (25.0%), birds (3.7%), or other pets (11.3%), evincing that dogs are the most common pet type in American households [2].

Based on the Journal of Veterinary Behavior, dogs not only provide companionship but also give people emotional support and psychological health [3]. Meanwhile, dogs have a rich capacity for emotional expression, such as happiness, anger, sadness, and fear, so knowing their emotions began to play a key role as a pet owner. It not only improves animal welfare but also fosters deeper communication between people and their pets. However, because of the language barrier between species, it is difficult for people to identify what their pets really need. According to "Owners' beliefs" reckoning the emotional capabilities of dogs and cats [4], people often rely on subjective experience when interpreting pet emotions, which carries the risk of misunderstanding. This misinterpretation can be seen as heterospecific emotion recognition in pets, causing individuals to mediate affiliative behaviors and avoid potentially harmful interactions [5]. Therefore, exploring more scientific and accurate methods is momentous.

In recent years, many researchers have focused on automatic emotion recognition in animals, like dogs. For example, Correia-Caeiro's research (2021) demonstrated that dogs primarily rely on bodily emotional expressions as a source of information, highlighting the potential for automated visual analysis [6]. Similarly, Xie's research (2022) developed deep learning models to classify canine emotions from facial cues, reflecting a growing trend toward computer-vision-based emotion

recognition systems in animal behavior research [7]. Researchers have begun using machines learning to classify different animals' emotions based on their facial features. Compared with SVM and KNN, Convolutional Neural Networks (CNNs) have shown huge advantages in extracting tricky cues from raw images automatically. The literature review indicates that among the three algorithms, CNN utilizes less time while providing greater accuracy, making it advantageous to adopt CNN for future applications [8].

However, most of the previous studies have relied on single model architecture. In contrast, this study systematically compares multiple CNN configurations to determine which structural variations best capture emotional cues in dog facial images.

The research uses features restored in an input image to go through processes like convolution and pooling. After that, the features will be restored in basic vectors, which simplifies the classification by only comparing the discernible points of the pictures. Such work not only increases the performance of CNN technical understanding but also contributes to the development of smarter pet emotion identification systems.

2. Methodology

2.1. Comparative Experiment

2.1.1. Experiment Overview:

The significant purpose of the research **is** to analyze and compare the performance of four custom fully-connected-layer architectures through **a** systematic comparative experiment**.**

The research designs an appropriate environment for a certain CNN architecture to train, validate, and test, by setting up the required and correctly corresponding versions of the packages of TensorFlow and Keras, and the CUDA toolkit. The research controls the variables, including the programming environment, the training-validation-split ratio, the data preprocessing, etc.

The research experiments on four CNN architectures: a custom-designed sequential CNN network and three pretrained models, as MobileNetV2, EfficientNetB3, and EfficientNetB7. The paper uses abbreviations for these architectures, such as OCNN, MNV2, ENB3, and ENB7, respectively, to enhance the concision.

2.1.2. Control Group:

The control group of the research is OCNN, and its architecture roadmap is shown in Figure 1.

OCNN is constructed by three convolution blocks, and each one includes a Conv2D and a MaxPooling2D for facial features extraction with attenuating calculating dimensions. After convolution, the architecture uses Flatten to rescale the features into vectors. After vectors are passed through a 256-channel Dense layer, they are sent through a Dropout layer. The research chooses to use the Dropout with a drop rate of 0.4 to prevent overfitting, because this process can deactivate random neurons during training. The same strategy is used in another lightweight CNN research, which aims to reduce neuron interdependence and achieve a more compact architecture for one- and two-dimensional networks [9]. Then, in order to output multitype confidence, the last processing layer of hidden layers is a dense layer using SoftMax activation.

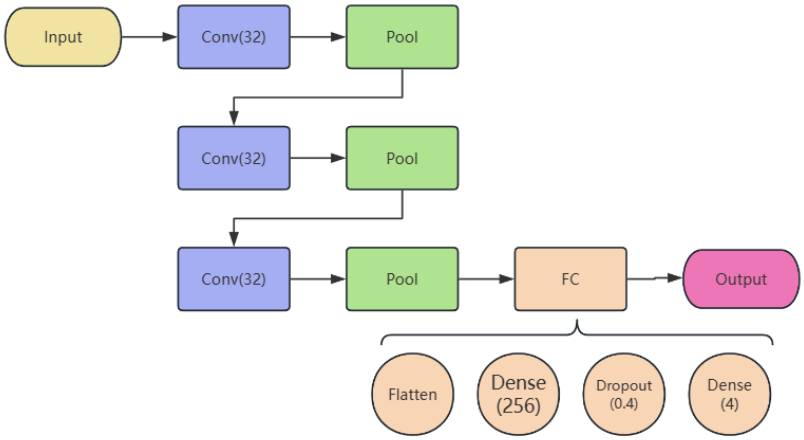
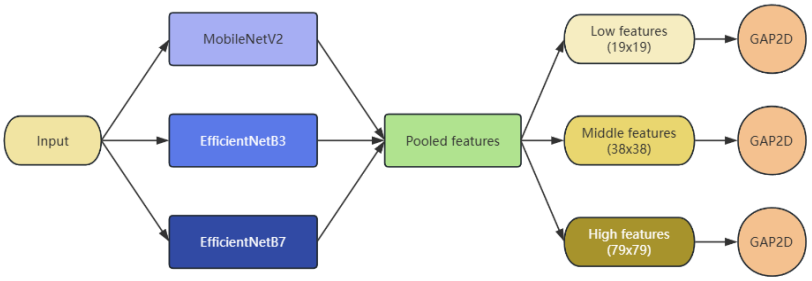


Fig 1. The Architecture of OCNN (Picture credit: Original).

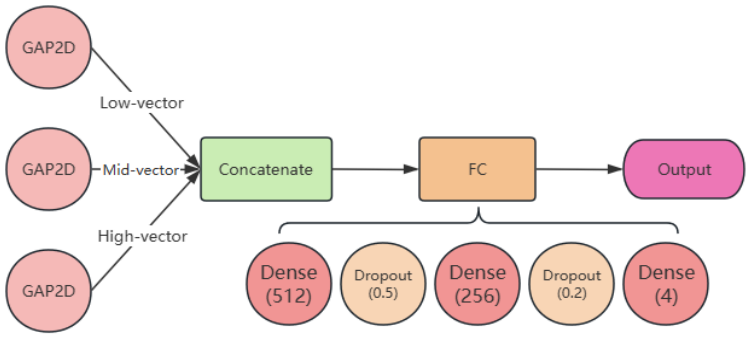
2.1.3. Experiment Group:

The experiment group of the research is constructed by three similar architectures that only differ in the base model. These are MNV2, ENB3, and ENB7, as shown in Figure 2. This group is used to evaluate the effects of architectural complexity and feature representation level on the classification performance.

Each architecture extracts three layers of features for concatenating, and they are: low-level features, middle-level features, and high-level features with corresponding convolutional output sizes as 19 * 19, 38 * 38, and 79 * 79, respectively. These pooled features are flattened in the same Global Average Pooling 2D, generating a feature vector for the corresponding level. This process is shown in Figure 2 (a).



(a)



(b)

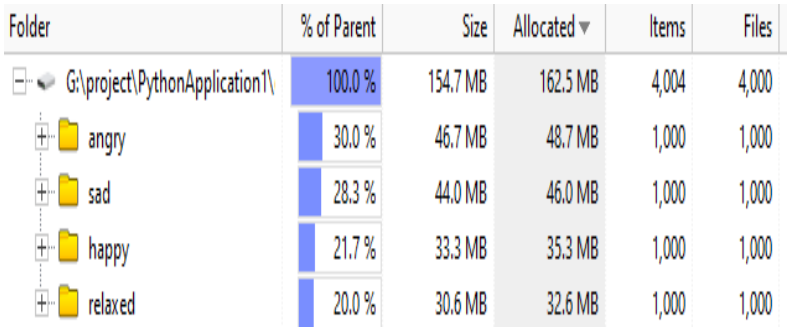
Fig 2. (a) Multi-level features extraction of each architecture,
(b) Processing roadmap after concatenation (Picture credit: Original).

As shown in Figure 2 (b), the product vector after concatenating is sent to Fully Connected where two steps of dense with 512 and 256 channels and a step of dense using SoftMax activation are used. There are two dropout layers in between dense blocks that have a drop rate of 0.5 and 0.2.

2.2. Dataset Description and Preprocessing

2.2.1. Dataset Description:

The research utilizes a dataset comprising dog facial images from Kaggle [10]. The four major categories of data images are angry, sad, happy, and relaxed. There are 1000 RGB-channel images in each emotion category, and in total, there are 4000 images for later processing, as shown in Figure 3. The dataset will be split into two sets: training and validation. After this, the researchers are using the other clip of dog facial images from a different dataset of Kaggle with a number of 500 to set up a testing set, serving to evaluate the model's effectiveness after training [11].



Folder	% of Parent	Size	Allocated	Items	Files
G:\project\PythonApplication1\	100.0 %	154.7 MB	162.5 MB	4,004	4,000
angry	30.0 %	46.7 MB	48.7 MB	1,000	1,000
sad	28.3 %	44.0 MB	46.0 MB	1,000	1,000
happy	21.7 %	33.3 MB	35.3 MB	1,000	1,000
relaxed	20.0 %	30.6 MB	32.6 MB	1,000	1,000

Fig 3. The overview of using categorical images (Picture credit: Original).

2.2.2. Data Preprocessing:

The preprocessing section in all groups of the experiment uses identical parameters when randomly adjusted. At first, the sizes of input images are adjusted to 256 x 256 pixels, and rescaled to a range of [0, 1] to assuage the learning interruption of brightness and contrast. The research uses data augmentation during training through the Image Data Generator. For example, the images are horizontally flipped, rotated, zoomed, or brightness shifted in a certain range, in order to simulate the probable situation in real-time classification, where a slight change in vision angle, illumination, or position can occur. Figure 4 shows some possible adjusted images after the preprocessing. Furthermore, the images are split by a ratio of 4:1 into the training and validation sets.

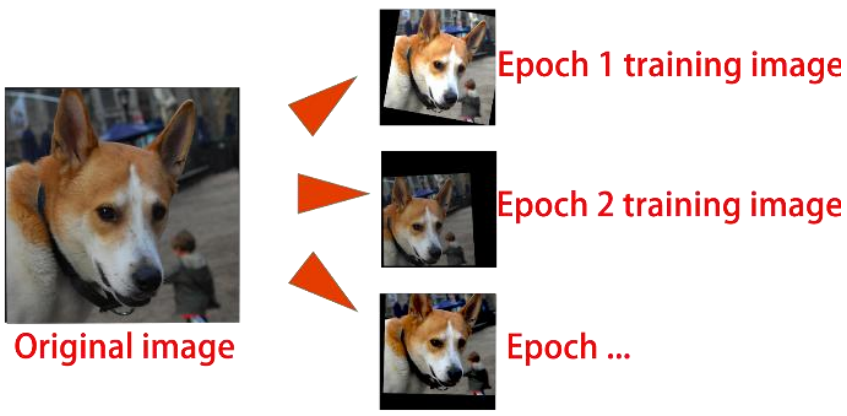


Fig 4. Examples of Preprocessed Images (Picture credit: Original).

2.3. Training Configuration

2.3.1. Hyperparameters:

Each architecture uses categorical cross-entropy as the loss function; The optimizer is Adam. The evaluation metrics include accuracy, loss, precision, and recall. The hyperparameters, such as learning

rate, batch size**,** and epoch amount, are 0.001, 32, and 20, respectively, for both control and experiment groups, maintaining control variables.

2.3.2. L2 Regularization:

To appease the overfitting during the training process, the research introduces L2 Regularization in Dense layers of Fully Connected for the experiment group. L2 Regularization adds the terms of square sums of weights into the loss function, putting a penalty on over-weighted situations during the training process to control the model complexity and increase generalization ability. The research sets the regularization coefficient, lambda, as 0.001, which means that every weigh**t** parameter will be multiplied by one-thousandth and be summed in the loss function.

This method's effectiveness in enhancing robustness and generalization ability has been proven in many studies. For example, in Ng's research (2004), L2 regularization significantly ameliorates the model stability and smooths the overfitting rate, especially when the weights scale tends to be huge [12].

2.3.3. Callbacks:

The research introduces Early Stopping with a patience of 3 epochs in the callbacks, monitoring validation loss, and automatically stopping the training process to avoid overfitting, which could realize the maximum balance between training efficiency and generalization ability.

At the same time, Early Stopping helps the researchers quickly find out the epoch when the model starts overfitting. In terms of increasing clarity for multiple architectures training tasks, it provides a standardized and consistent termination point for different architectures, thus making the performance comparison fairer.

Besides, Early Stopping reduces unnecessary computational costs and training time, especially when working with large-scale architectures such as ENB7. The research ensures minimal resource waste by stopping the training when there is no significant improvement in validation accuracy.

3. Results and Discussion

3.1. Training and Validation

3.1.1. Accuracy and Loss:

In accordance with evaluation metrics, ENB3 reaches the highest validation accuracy of 73.1% at epoch 10, and it triggers Early Stopping later than all three other models, which demonstrates it has the most stability and the best performance throughout the experiment group. The ENB7, MNV2, and OCNN have the highest validation accuracy of 72.3%, 67.7% and 46.4% at epoch 7, 6, and 9, respectively, from the best performance to worst in the rest of the architectures. ENB7 has the highest increasing rate of both training and validation at first, but experiences fluctuation after five episodes. Figure 5 shows the line chart of training and validation accuracy over epochs.

In Figure 5, ENB7 has the highest decreasing rate of loss at the beginning of the epoch. ENB3 and MNV2 also have an obvious decreasing inclination, but they are slightly smoother than ENB7. Although OCNN has the lowest loss over epochs, its improvement is still less significant than that of other architectures because it almost maintains its epoch-1 loss after 12 epochs.

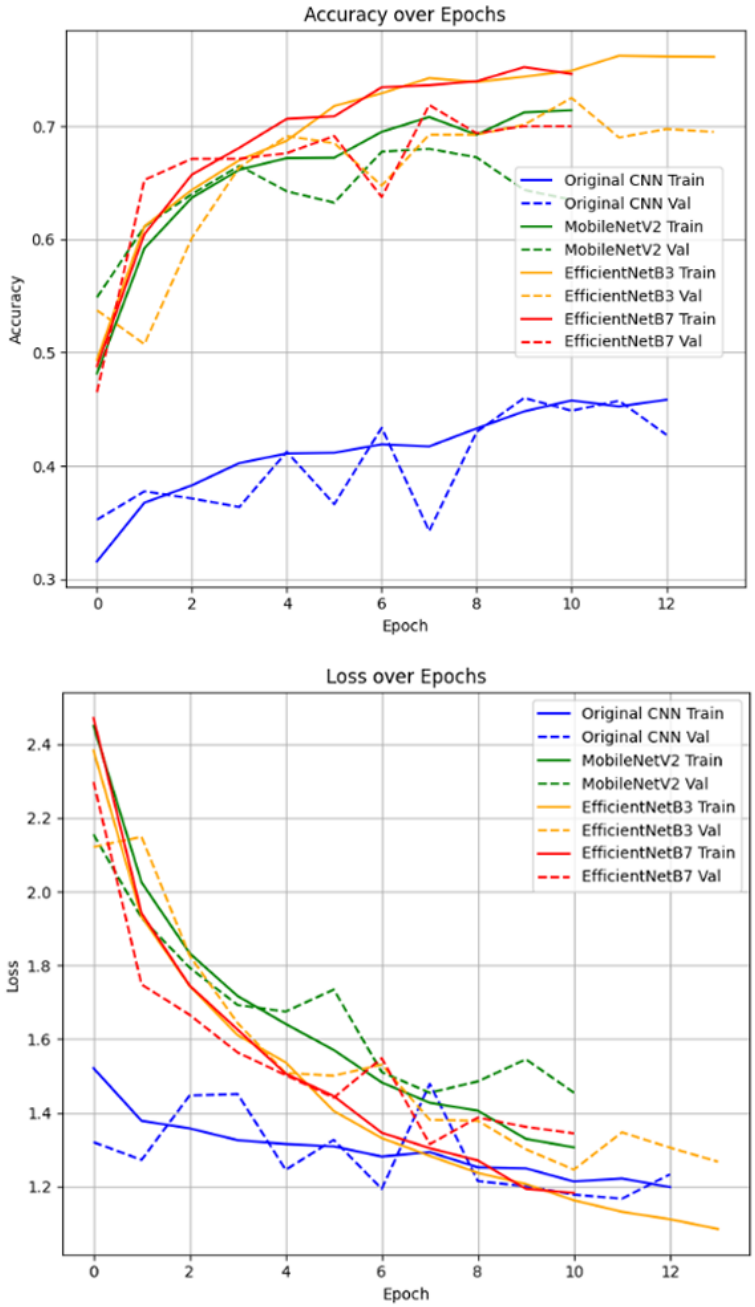


Fig 5. The Accuracy and Loss over Epochs (Picture credit: Original).

3.1.2. Precision and Recall:

Recording the evaluation metrics of precision and recall can help the researchers to observe whether the architecture is precise at a certain point. More importantly, it reveals the model's bias and capability boundaries when dealing with imbalanced class distributions.

In Figure 6, the precisions of ENB3, ENB7, and MNV2 are steadily increasing while OCNN's is experiencing fluctuation. It displays the advantages of deep-level feature extraction and better adaptability in dog emotion classification. Specifically, ENB7 has the biggest increasing rate of both training and validation precision and finally reaches the highest validation precision of 75.8%. ENB3, MNV2, and OCNN reach architecturally the highest validation precision of 75.4%, 74.1% and 57.6%.

In Figure 6, ENB7 still has the largest recall increasing rate at the beginning, while ENB3 reaches the highest validation recall of 68.0%. ENB7, MNV2, and OCNN reach their architecturally highest validation recall of 67.5%, 58.7% and 29.8%. The considerable point of OCNN's recall is its slight change in the process.

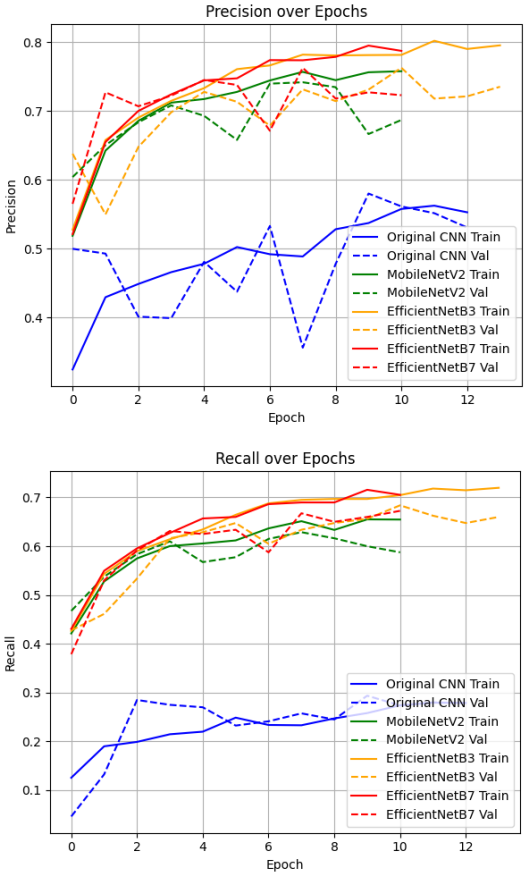


Fig 6. The Precision and Recall over Epochs (Picture credit: Original).

3.1.3. F1-score:

F1-score is the adjusted average of precision and recall. It is useful when the research needs to compare a balance between precision and recall, especially when comparing the performance of four different CNN architectures. It has a strong sensitivity of extremely low precision or recall, making the evaluation more conservative.

$$F_1{}^{weighted} = \sum_{i=1}^{n} w_i \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \tag{1}$$

Expression (1) displays a mathematical method to calculate the weighted F1 Score for architectures. In the equation, $F_1{}^{weighted}$ is the weighted F1-score which covers all the classification classes, w_i is the weight for class i, P_i *and* R_i are precision and recall of class i. Through this, research can evaluate the compressive ability of architecture.

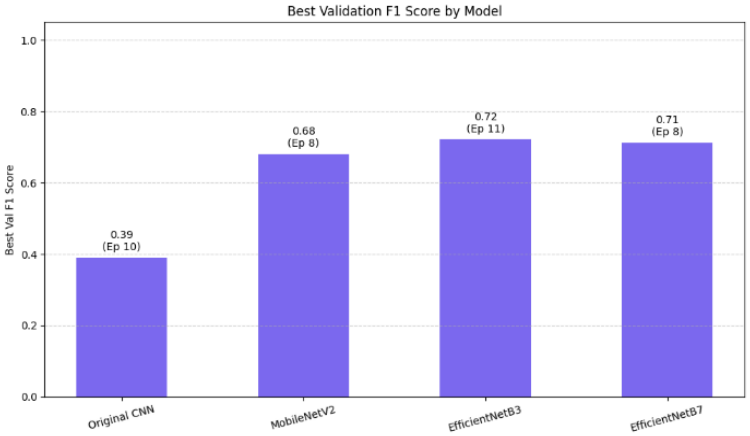


Fig 7. The Best F1-score (Picture credit: Original).

Figure 7 shows the calculated F1-score for four architectures. Specifically, ENB3 has the best performance of F1-score evaluation, reaching 72% at epoch 11. ENB7 ensues, then there are MNV2 and OCNN. According to OCNN's low and hardly improving recall, the extremely low F1-score (39%) is reasonable.

3.2. Testing

Objectively, Figures 8, 9, 10 and 11 exhibit the testing performance of four CNN architectures based on the testing set with 125 images of each emotion clip and in total 500 images.

At first, ENB3's classification ability is the most equilibrant and accurate. Architecture shows a powerful capability in classifying the four categories of emotion. It also has the largest number of correctly classified cases for emotion "angry", "happy", and "sad", indicating an overall better performance of feature memorizing and generalizing. Secondly, ENB7 and MNV2 also display effective classification based on the testing set, but they present confusion when classifying emotions such as "happy" and "relaxed". Compared to the experiment group, the control group, OCNN, has a critical confusion when classifying the testing set. There are only 36 out of 125 images that are correctly categorized into the emotion "angry", and the rest of the emotions also occur with many more mistakes, especially the class of "sad", which also has an obvious confusion similar to its classification as "angry".

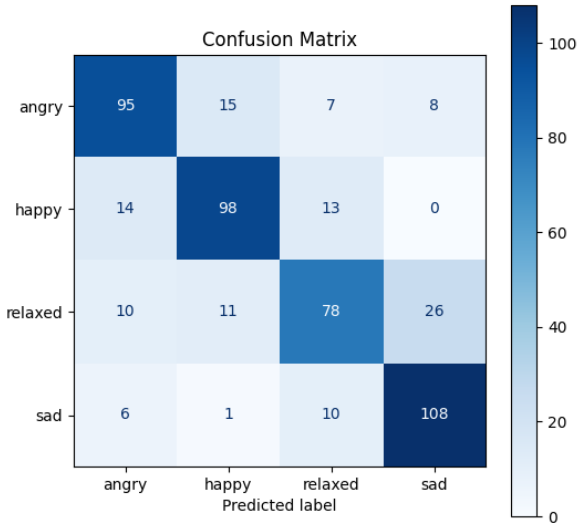


Fig 8. The Testing Confusion Matrix of ENB7 (Picture credit: Original).

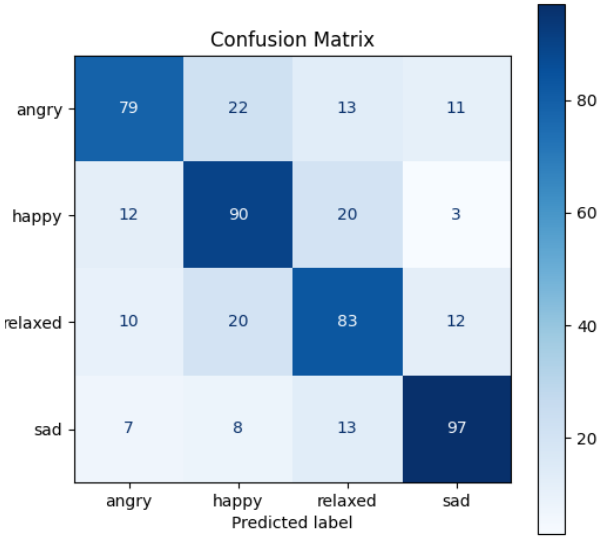


Fig 9. The Testing Confusion Matrix of MNV2 (Picture credit: Original).

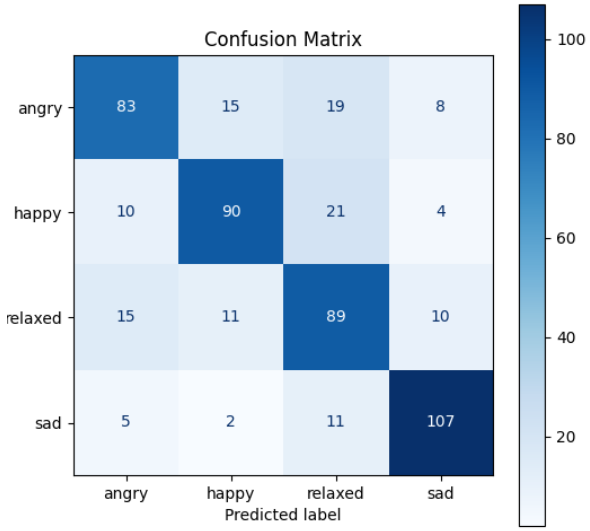


Fig 10. The Testing Confusion Matrix of ENB3 (Picture credit: Original).

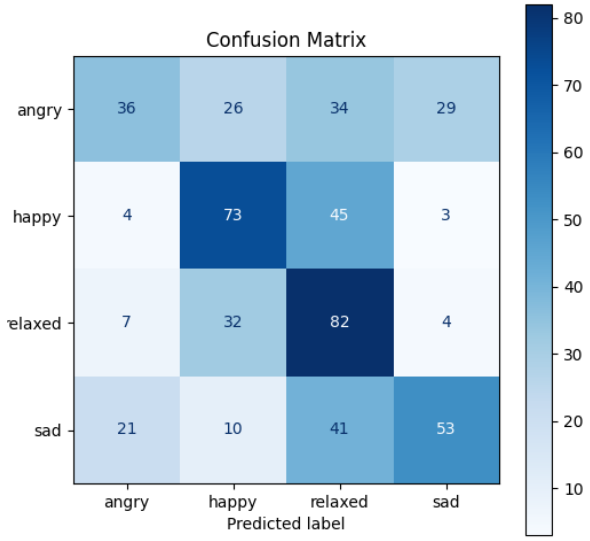


Fig 11. The Testing Confusion Matrix of OCNN (Picture credit: Original).

3.3. Implications of Architectures

Figure 12 shows the relationship between the validation precisions of four architectures at the last epoch and the architecture sizes, revealing the connection between the architecture complexity and the corresponding performance. In this research, ENB3 presents the best validation precision of 74% and remains middle-amount parameters of 12.2 million, showing a desirable cost-performance ratio.

The control group, OCNN, has only 0.1 million parameters, and it has the worst overall performance among other architectures, showing that its three convolution blocks have limitations when dealing with the dog emotion classification task, and feature extraction tends to be superficial and insignificant. Although it shows the lowest categorical cross-entropy, its loss almost fails to decrease over the epochs, which means that it cannot generalize well based on augmented images. It is possibly attributed to its difficulty in capturing slight changes in a dog's facial expression, leading to confusion when classifying.

The experiment group, including ENB3, ENB7, and MNV2, has overall better performance. Specifically, ENB3 has the best training, validation, and testing accuracy. This can be ascribed to ENB3's depth and width and efficient Compound Scaling strategy. Compared to this, ENB7, though, reaches a large scale of 66.6 million params; its highest validation precision is lower than ENB3. This indicates that the complexity is over the need of the dog emotion classification task, which can be

attributed to the small dataset and redundant feature extraction. ENB7 also has the longest training, validation, and testing time, which is not convenient for small-scale research and classification. As a lightweight architecture similar to OCNN, MNV2 has 2.3 million parameters and a significant improvement in cost-performance ratio due to its overall results. It is also convenient to choose when the research has a particular need for a short time and low computational cost.

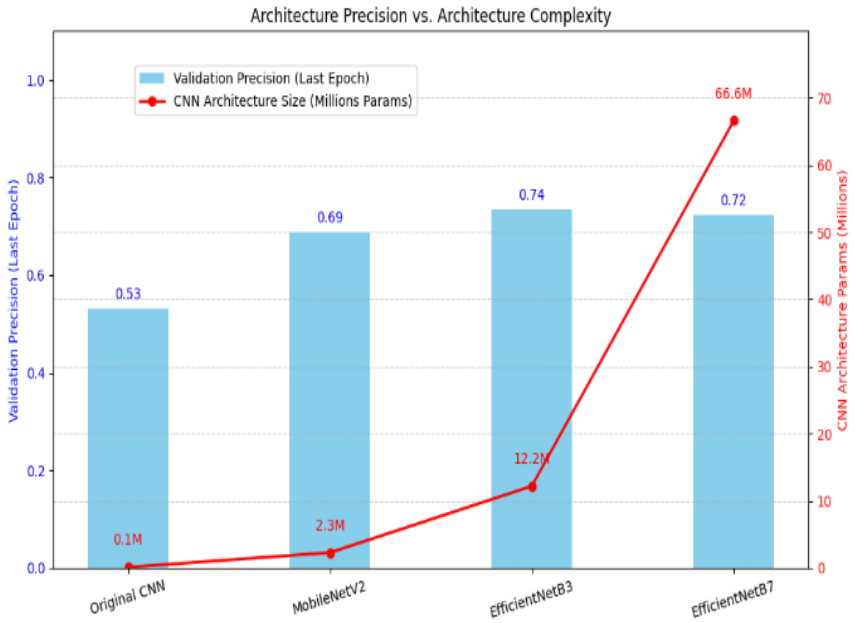


Fig 12. Architecture Precision vs. Architecture Complexity (Picture credit: Original).

4. Conclusion

This study did a comprehensive comparison between four convolutional neural network (original CNN) architectures-CNN, MobileNetV2, EfficientNetB3, and EfficientNetB7- to classify dog emotions based on facial images. By including consistent datasets, preprocessing techniques, and evaluation metrics (accuracy, precision, recall, and F1-score), the paper observed that the type of architectures significantly affects classification performance.

Among the four architectures, EfficientNetB3 constantly performed better than the other three architectures, achieving the highest F1-score, about 72%, and illustrating the best balance between complexity and generalization. Its relative complexity and appropriate scaling strategy led to an effective feature extraction from the original data without extra computation. In contrast, the original CNN illustrates a noticeable underfitting feature, which means that it is unable to capture and predict fine-grained emotional features because of its shallow structure. On the other hand, although EfficientNetB7 is a powerful and superior complex model, it suffered from overfitting and inefficiency due to its redundant structure. For MobileNetV2, a lightweight model, it showed a good trade-off between performance and resource demand, considering its limited computational resources.

Hence, these discoveries underscore the importance of selecting the appropriate model to identify the animals' emotions, especially in the fields of animal welfare, where the number of datasets is limited, and the facial details are unclear and subtle, which means that it will be hard for a simple model to recognize. This research can both deepen the technological understanding of the classification of emotions and push the development of more intelligent, empathetic interaction systems between humans and animals.

In the future, enlarging the dataset to include more aspects, like facial angles or light conditions, will improve the ability of generalization. In this way, the model could better recognize more difficult and complex images. In addition, ethical considerations are critical to ensure these technologies can be utilized responsibly and help to keep and promote animal welfare.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Wani, A., Pawar, U., Gatagat, Y., & Thalor, M.: ' Handwritten character recognition using CNN, KNN and SVM', International Journal of Technology Engineering Arts Mathematics Science, 2021, 1(2), 19–26

[2] Applebaum, J. W., Peek, C. W., & Zsembik, B. A.: ' Examining U.S. pet ownership using the General Social Survey', The Social Science Journal, 2020, 60(1), 110–119

[3] Wells, D. L.: 'Horse-training techniques that may defy the principles of learning theory and compromise welfare', Journal of Veterinary Behavior, 2010, 5(5), 226–234

[4] Pickersgill, O., Mills, D. S., & Guo, K.: 'Owners' beliefs regarding the emotional capabilities of their dogs and cats', Animals,2023, 13(5), 820

[5] Albuquerque, N., Mills, D. S., Guo, K., et al.: 'Dogs can infer implicit information from human emotional expressions', Animal Cognition, 202225, 231–240

[6] Correia-Caeiro, C., Guo, K., & Mills, D. S.: 'Bodily emotional expressions are a primary source of information for dogs, but not for humans', Animal Cognition, 2021, 24, 267–279.

[7] Xie, Y.: 'Dogs emotion recognition and parameter analysis based on EfficientNet with transfer learning', Science and technology publictions. 2024.

[8] American Pet Products Association. State of the Industry Report, 2025.

[9] Begum, M., et al.: 'LCNN: Lightweight CNN architecture for software defect feature identification using explainable AI', IEEE Access, 2024, 12, 55748.

[10] Dataset Training. https://www.kaggle.com/datase ts/danielshanbalico/dog-emotion. 2025.

[11] Dataset Testing. https://www.kaggle.com/datasets /devzohaib/dog-emotions-prediction. 2025.

[12] Ng, A. Y.: 'Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine Learning (ICML '04), 78', 2004, Association for Computing Machinery, New York, NY