Feline Feelings Unleashed: Harnessing Deep Learning Through Photos For Cat Pain Detection

Chutimon Namboonlue [1*], Vera Sa-ing [2]

Abstract

Effective pain detection in cats, who cannot verbally communicate, is challenging and crucial. This study analyzes the performance of EfficientNetB7, a pre-trained convolutional neural network, for classifying feline pain using a dataset of 57 images per category, labeled ‘pain' or ‘no pain' by Thai veterinarians. The images were preprocessed and run through various configurations of EfficientNetB7, differing in batch sizes and learning rates, with ImageNet weights as the initial training parameters. The models were evaluated based on accuracy, precision, and recall. The most effective model, using the SGD optimizer with a learning rate of 0.001 and a batch size of 100, achieved 79% accuracy, 74% precision, and 90% recall. These findings demonstrate the potential of deep learning for non-verbal pain detection in veterinary settings, especially with the high recall rate essential for identifying animals in distress. This research opens avenues for integrating such AI models into veterinary practice, enhancing animal welfare.

Keywords : convolutional neural network, deep learning, cat pain detection, feline feelings

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author: vera@g.swu.ac.th

INTRODUCTION

Under typical circumstances, humans are capable of manifesting a range of emotions via facial expressions, one of which pertains to the experience of pain. This specific expression is the result of coordinated movements of designated muscle groups in the facial region, referred to as 'Action Units'[1].

In 1978, Ekman et al. undertook a study exploring the link between different groups of Action Units and human emotions. The framework that illustrates this connection is known as the Facial Action Coding System (FACS) [1]. This system has since become a widely referenced tool in numerous studies focusing on the analysis of human emotions [2]. Therefore, this idea was applied to create a system that shows the relationship between various Action Units and the facial expressions of animals in order to study behavior related to various facial expressions. Each animal has its own system based on its anatomy. For example, the system in cats is called CatFACS [3].

Researchers are using Animal FACS systems to enhance the understanding of the correlation between facial expressions and the manifestation of pain. This has led to the development of the 'Grimace Scale', a pain assessment tool predicated on the observation of prominent facial regions rather than individual muscle movements. This approach is designed for ease of use, enabling both professionals and non-specialists to conduct assessments [4]. However, it's important to note that the reliance on observational techniques may introduce subjectivity and potential bias into the evaluation process [5].

In 2019, Finka et al. advanced the field of animal behavior by quantitatively analyzing cat facial expressions in response to pain. Using the Cat Facial Action Coding System (CatFACS) as a framework, the team annotated 48 facial landmarks in a series of cat photographs. They assessed the variations in these landmarks under conditions of pain using principal component analysis (PCA). This study represents a pioneering application of statistical analysis to quantify the facial expressions of cats in pain [6]. In subsequent research conducted in 2022, Feighelstein et al. developed a predictive model for assessing pain in cats, utilizing photographic analysis of their facial expressions. The team compared the efficacy of a multi-layer perceptron model, which employed the same 48 reference points as previously established to create regional feature vectors, against that of a fine-tuned ResNet50 model. Their findings suggested that the multi-layer perceptron model provided more accurate predictions in the assessment of feline pain, potentially due to the limited amount of training data [7]. Following this, in 2023, the same researchers expanded their study to investigate the impact of individual facial regions (ears, eyes, mouth) on the accuracy of pain classification. They compared the effectiveness of a random forest model, a multi-layer perceptron model, and a fine-tuned ResNet50 model. Each model was trained using the full image of the cat, then with each facial region occluded, and finally by revealing only individual facial regions to ascertain the contribution of each region to the overall pain assessment. The findings revealed that, when analyzing images with a fully visible face, the random forest model achieved the highest accuracy at 77.2%, followed by the multi-layer perceptron (MLP) at 69.3%, and the convolutional neural network (CNN) at 63.6%.

Notably, the mouth region exhibited the most significant influence on classification accuracy, while the ears had the minimal impact on the overall accuracy score [8]. EfficientNetB7 [9], as part of the EfficientNet family, embodies a unique design philosophy that seeks to optimize the balance between model size, computational efficiency, and performance. This architecture introduces a compound scaling method that uniformly scales network width, depth, and resolution, thereby ensuring that the model remains efficient while delivering state-of-the-art results. The integration of mobile inverted bottleneck convolution blocks with squeeze-and-excitation (SE) blocks enhances feature extraction and utilization, contributing to its superior efficiency. These deep learning models were measured by using the most widely measurement method [10]- [12] that is the confusion matrix. This matrix measures the performance of the classification between the predicted correction and the predicted incorrection. Moreover, this technique was used to compare the testing performance of each compared model. So, this research uses this measurement to evaluate compared models.

The classification of cat pain from images represents a task that necessitates not only exceptional accuracy but also the judicious allocation of computational resources, especially in real-world veterinary settings. Traditional CNN architectures have achieved commendable results, but their often substantial computational demands can hinder their practical applicability. EfficientNetB7, with its ability to maintain high accuracy while being computationally efficient, emerges as an ideal candidate for addressing this challenge. Previous research has primarily focused on a single type of pre-trained convolutional neural network (CNN) model. In response, this study seeks to assess the performance of an alternative CNN architecture to determine its effectiveness in the given context. Moreover, the dataset utilized herein originates from Thailand, which presents a variable feline demographic when compared to that of the United Kingdom. This demographic distinction may have implications for the generalizability of the results, an aspect that will be thoroughly examined in the results discussion.

METHODOLOGY

A. Data Collection

The cat images, which included both genders, various breeds, a range of ages, and different reasons for hospital visits, were collected from August 2023 to January 2024. Sourced from multiple veterinary clinics, this approach ensured a diverse sample for the study. Photos were excluded from the study if they lacked complete facial features, such as both eyes, ears, or the muzzle area. Additionally, any images where facial features were partially or fully obscured, as seen in cases like black cats or chimeric cats with a half-black face, were also omitted. Initially, the images were annotated by the contributing veterinarians. The ‘no pain’ category comprised images corresponding to a score of 0 on the Colorado State University Feline Acute Pain Scale (CSU-FAPS) [13], whereas the ‘pain’ category included those with a score of 2 or higher. Subsequently, the photos underwent a secondary validation process by an expert P.V., as illustrated in Fig. 2.

Fig. 1 Cat facial expressions represent (a) pain and (b) no pain

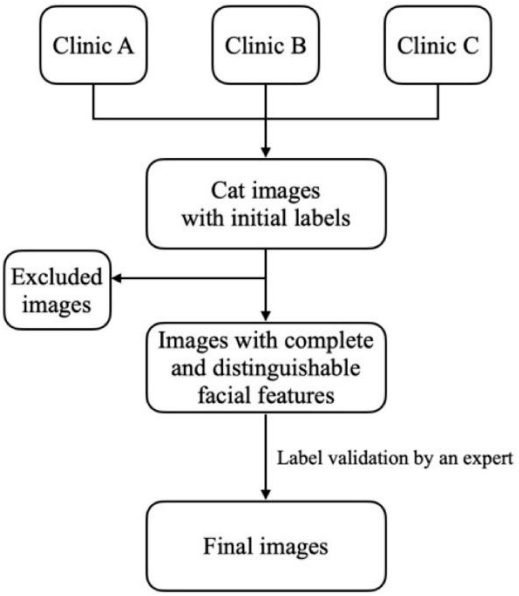

Fig. 2 Flowchart of data collection and selection

B. Data Preprocessing

Nine facial landmarks of the cats were identified using a specialized cat facial landmark detection algorithm [14]- [16]. The photographs were then rotated to align the x-coordinates of the centers of the left and right eyes on the same horizontal plane. Subsequently, the images were cropped into a square format to concentrate on the facial features, and resized to ensure consistent dimensions. The overall processes were depicted in Fig. 3. These images were split into training, validation, and test sets in the ratio of 47:5:5.

In the preparation of the datasets for the deep learning pipeline, the training, validation, and test datasets were independently passed through an image augmentation process that consisted of flipping, rotation, and contrast enhancement. The examples of augmented images were shown in Fig. 4. Upon completion of this augmentation process, the resultant datasets consisted of 470 images in the training set, and 50 images each in the validation and test sets, for every class.

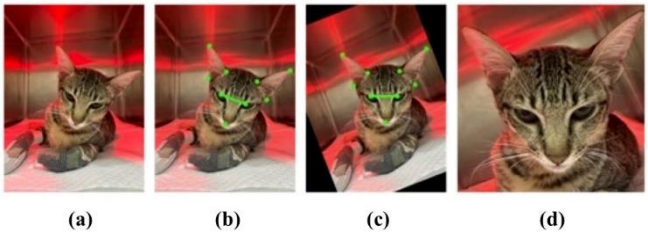

Fig. 3 Overall data preprocessing of alignment correction and resizing process (a) Original image prior to modification. (b) Image annotated with detected facial landmarks. It is observed that the coordinates for the left and right eyes are not aligned on a horizontal axis. (c) Adjusted image post-rotation to align the eye levels horizontally. (d) Final cropped image emphasizing the facial profile, optimized for subsequent analytical procedures.

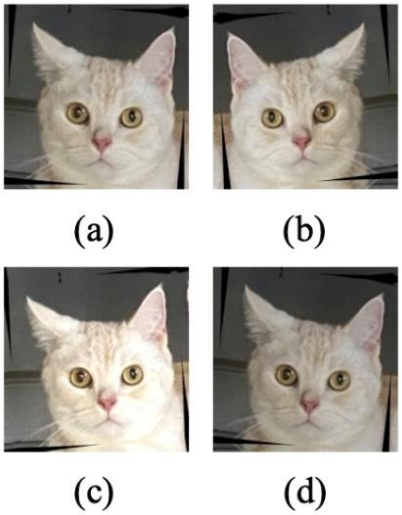

Fig. 4 Examples of augmented images consist of (b) flipped, (c) contrast enhanced, and (d) slightly rotated from the original image (a).

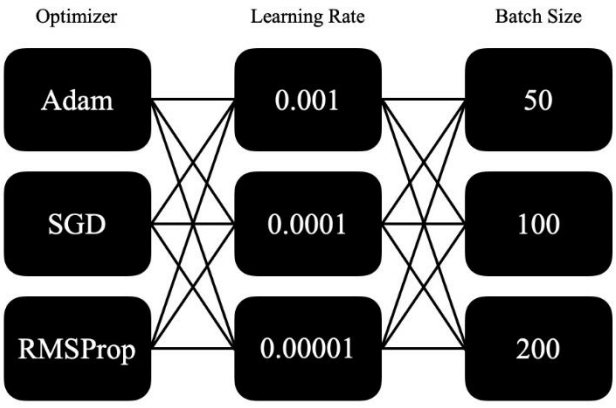

Fig. 5 Models were trained with different combinations of optimizers, learning rates, and batch sizes.

C. Modeling

In this study, we employed the EfficientNetB7 convolutional neural network architectures, initializing them with pre-trained weights from the ImageNet database. We modified the fully connected layers to include two dense layers comprising 64 and 32 neurons, respectively, utilizing the Rectified Linear Unit (ReLU) activation function. To reduce overfitting, a dropout layer was subsequently integrated into the model for regularization purposes. The output layer consists of a single neuron with a sigmoid activation function to classify the binary outcomes. All convolutional layers were frozen during the initial training phase with our augmented dataset. Subsequently, we unfroze the last convolutional block to fine-tune the parameters further.

The experimental setup involved the use of three optimizers—Adam, Stochastic gradient descent (SGD), and RMSprop—with learning rates of 0.001, 0.0001, and 0.00001, and the models were trained with batch sizes of 50, 100, and 200, as illustrated in Fig. 5, resulting in a total of 27 distinct models.

D. Evaluation Metrics

In this research, a confusion matrix was utilized to visualize the classification performance, measuring the proportion of correct and incorrect predictions against the actual outcomes. This methodology delineates four result categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The values derived from these categories facilitate the computation of key performance metrics, namely accuracy, precision, recall, and the F1-score, which are essential for assessing the efficacy of the deep learning model.

a) Accuracy : Accuracy (ACC) is a metric that illustrates the overall performance of a model in making correct predictions. It is expressed as a percentage reflecting the proportion of predictions that the model gets right. This percentage is calculated by dividing the sum of correct predictions, which includes both true positives and true negatives, by the total number of cases in the confusion matrix. The equation for accuracy is as follows:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

b) Precision : Precision (PCS), or positive predictive value, measures the ratio of true positive predictions (TP) to the total number of predicted positives, which includes both true positives (TP) and false positives (FP). The equation for precision is as follows:

$$PCS = \frac{TP}{TP+FP}$$

c) Recall : Recall (REC), also known as sensitivity, is a metric widely utilized in the medical field and numerous other research domains to assess a model's ability to correctly identify positive instances. A high recall value indicates a high likelihood of the model correctly predicting true positive cases, such as accurately identifying cats experiencing pain. Recall is calculated by dividing the number of true positives (TP) by the total number of actual positives, which includes both true positives and false negatives (FN). The equation for recall is as follows:

$$REC = \frac{TP}{TP+FN}$$

d) F1-score : Optimal deep learning evaluation aims for high recall and precision, yet in practice, achieving 100% for both is rarely possible due to their trade-off. The F1-score is introduced as the harmonic mean of precision and recall to balance this trade-off. A high F1-score indicates that the model has a robust balance between precision and recall. The F1-score is calculated using the following equation:

$$F1S = \frac{2*PCS*REC}{PCS+REC}$$

EXPERIMENTAL RESULTS

The research utilized the Python programming language and Keras library with a TensorFlow backend, executed on an A100 GPU within the Google Colab environment. Various parameters were systematically adjusted to fine-tune the model. The results were as shown in Fig. 6.

The model that achieved the highest accuracy within the Adam optimizer group was trained with a learning rate of 0.001 and a batch size of 100. It attained an accuracy of 73%, a precision of 65%, a recall of 100%, and an F_1-score of 78%. Meanwhile, in the RMSProp group, the model with a learning rate of 0.0001 and a batch size of 200 achieved the highest accuracy within the group. This model's accuracy, precision, recall, and F_1 score were 74%, 68%, 90%, and 78%, respectively. The model that achieved the highest overall accuracy among the three optimizer groups was trained using the SGD optimizer, with a learning rate of 0.001 and a batch size of 100, attaining an accuracy of 79%. Additionally, it achieved a precision of 74%, the highest among all tested models. Regarding the confusion matrix in Fig. 7, the model demonstrated 90% sensitivity. This means that, given 50 images of cats in pain, the model was able to correctly detect 45 of them. However, the model's specificity, which measures the proportion of true negatives in the predicted negatives, showed a less promising result of 68%. This suggests that, given 50 images of cats without pain, the model was able to accurately detect only 34 of them.

CONCLUSION

In this study, we explored the practicality of using EfficientNetB7, empowered by transfer learning, to identify pain in cats through image analysis. Our work was particularly challenging due to the dataset's uncontrolled

Optimizer	LR	Batch size	Accuracy	Precision	Recall	F1 Score
Adam	0.001	50	0.54	0.53	0.64	0.58
Adam	0.001	**100**	**0.73**	**0.65**	**1**	**0.78**
Adam	0.001	200	0.66	0.61	0.92	0.73
Adam	0.0001	50	0.66	0.68	0.6	0.63
Adam	0.0001	100	0.61	0.58	0.78	0.66
Adam	0.0001	200	0.62	0.61	0.66	0.63
Adam	0.00001	50	0.63	0.6	0.78	0.67
Adam	0.00001	100	0.59	0.59	0.6	0.59
Adam	0.00001	200	0.71	0.65	0.9	0.75
SGD	0.001	50	0.55	0.54	0.76	0.63
SGD	0.001	**100**	**0.79**	**0.74**	**0.9**	**0.81**
SGD	0.001	200	0.62	0.6	0.72	0.65
SGD	0.0001	50	0.5	0.51	0.66	0.57
SGD	0.0001	100	0.58	0.59	0.54	0.56
SGD	0.0001	200	0.37	0.4	0.52	0.45
SGD	0.00001	50	0.37	0.39	0.44	0.41
SGD	0.00001	100	0.37	0.41	0.6	0.48
SGD	0.00001	200	0.22	0.11	0.08	0.09
RMSProp	0.001	50	0.63	0.61	0.74	0.67
RMSProp	0.001	100	0.66	0.63	0.78	0.7
RMSProp	0.001	200	0.57	0.58	0.52	0.55
RMSProp	0.0001	50	0.7	0.67	0.8	0.73
RMSProp	0.0001	100	0.71	0.63	1	0.78
RMSProp	0.0001	**200**	**0.74**	**0.68**	**0.9**	**0.78**
RMSProp	0.00001	50	0.64	0.59	0.9	0.71
RMSProp	0.00001	100	0.5	0.5	0.54	0.52
RMSProp	0.00001	200	0.61	0.6	0.66	0.63

Fig. 6 Training outcomes for the 27 evaluated models. Note that bolded characters indicate the maximum accuracy achieved by each optimizer.

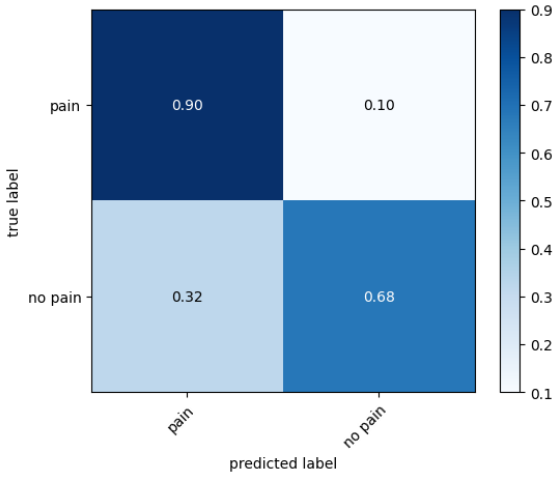

Fig. 7 Confusion matrix representing the results of the best model's predictions.

environmental conditions, such as varying angles and lighting in the photographs. The limited size of our dataset, comprising only 47 training images, 5 validation images, and 5 test images per category, added to the complexity of our task.

Despite the aforementioned limitations, the findings are promising. The top-performing model, optimized with an SGD optimizer, a learning rate of 0.001, and a batch size of 100, achieved a noteworthy accuracy of 79%. Its sensitivity rate of 90% is especially encouraging, suggesting a strong ability to identify cats experiencing pain. However, the model's specificity, at 68%, indicates room for improvement in reducing false positives.

This research contributes to the evolving field of applying AI in veterinary medicine, demonstrating the potential of deep learning as an assistance in detecting pain in cats. The potential of such technology in enhancing animal welfare is clear, particularly in providing more accurate pain assessment. Future research should focus on expanding the dataset and refining the model to better serve the needs of veterinary professionals. This study serves as a foundation for further exploration into the integration of AI in veterinary practices, offering a promising avenue for the advancement of animal care.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system," Environmental Psychology & Nonverbal Behavior, 1978.

[2] E. A. Clark, J. Kessinger, S. E. Duncan, M. A. Bell, J. Lahne, D. L. Gallagher, and S. F. O'Keefe, "The facial action coding system for characterization of human affective response to consumer productbased stimuli: A systematic review," Frontiers in Psychology, vol. 11, pp. 1–21, 2020.

[3] C. Caeiro, A. Burrows, and B. Waller, "Development and application of catfacs: Are human cat adopters influenced by cat facial expressions?" Applied Animal Behaviour Science, vol. 189, pp. 66–78, 2017.

[4] M. C. Evangelista, R. Watanabe, V. S. Y. Leung, B. P. Monteiro, E. O'Toole, D. S. J. Pang, and P. V. Steagall, "Facial expressions of pain in cats: The development and validation of a feline grimace scale," Nature News, pp. 1–11, Dec 2019.

[5] M. C. Evangelista and P. V. Steagall, "Agreement and reliability of the feline grimace scale among cat owners, veterinarians, veterinary students and nurses," Scientifc Reports, vol. 11, no. 5262, pp. 1–9, Mar 2021.

[6] L. R. Finka, S. P. Luna, J. T. Brondani, Y. Tzimiropoulos, J. McDonagh, M. J. Farnworth, M. Ruta, and D. S. Mills, “Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar,” Scientifc Reports, vol. 9, no. 9883, pp. 1–12, Jul 2019.

[7] M. Feighelstein, I. Shimshoni, L. R. Finka, S. P. L. Luna, D. S. Mills, and A. Zamansky, “Automated recognition of pain in cats,” Scientifc Reports, vol. 12, no. 9575, pp. 1–10, Jun 2022.

[8] M. Feighelstein, L. Henze, S. Meller, I. Shimshoni, B. Hermoni, M. Berko, F. Twele, A. Schutter, N. Dorn, S. K astner, L. Finka, S. P. L. Luna, D. S. Mills, H. A. Volk, and A. Zamansky, “Explainable automated pain recognition in cats,” Scientifc Reports, vol. 13, no. 8973, pp. 1–16, Jun 2023.

[9] M. Tan and Q. V. Le, “Effcientnet: Rethinking model scaling for convolutional neural networks,” pp. 1–11, 2020.

[10] L.-E. Pomme, R. Bourqui, R. Giot, and D. Auber, “Relative confusion matrix: Effcient comparison of decision models,” in 2022 26th International Conference Information Visualisation (IV), 2022, pp. 98–103.

[11] W. Choomueang, C. Withoonchatri, P. Janwong, and V. Sa-Ing, “Ccvnet: Covid-19 detection in chest x-ray imaging based on convolutional neural network,” in Proceedings of the 2023 7th International Conference on Medical and Health Informatics, 2023, p. 1–5.

[12] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data– recommendations for the use of performance metrics,” in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 245–251.

[13] H. Shipley, A. Guedes, L. Graham, E. Goudie-DeAngelis, and E. Wendt-Hornickle, “Preliminary appraisal of the reliability and validity of the colorado state university feline acute pain scale,” Journal of feline medicine and surgery, vol. 21, no. 4, pp. 335–339, Apr 2019.

[14] B. Lee, “Cat hipsterizer,” Available online: https://www.kaggle.com/code/kairess/cat-hipsterizerl, 2018.

[15] W. Zhang, J. Sun, and X. Tang, “Cat head detection - how to effectively exploit shape and texture features,” in Computer Vision – ECCV 2008, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 802–816.

[16] G. Martvel, I. Shimshoni, and A. Zamansky, “Automated detection of cat facial landmarks,” pp. 1–20, 2023.