ENHANCING INFANT PAIN DETECTION WITH HYBRID ATTENTION MECHANISMS IN LIGHTWEIGHT MOBILENETV3 ARCHITECTURES

Anindya Apriliyanti Pravitasari[1*], Triyani Hendrawati[1], Anna Chadidjah[1], and Tutut Herawan[2]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jl. Ir. Sukarno Km. 21, Bandung 45363, Indonesia

[2]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

Emails: anindya.apriliyanti@unpad.ac.id*, triyani.hendrawati@unpad.ac.id, anna.chadidjah@unpad.ac.id, tutut@um.edu.my

ABSTRACT

Creating an automated pain detection system for infants less than a year old is essential because they are unable to communicate their discomfort verbally. Conventional assessment techniques like FLACC (Face, Legs, Activity, Cry, Consolability) require considerable time and may not be effective for infants with vocal cord impairments. Utilizing infants' facial expressions for real-time, automated pain detection presents a promising approach that facilitates rapid medical response. This study adopts a machine learning approach using infant facial expressions as input and explores the efficacy of various MobileNetV3 architectures, both Small and Large, enhanced with attention mechanisms. We introduced modifications involving 12 model variants, including the integration of CBAM (Convolutional Block Attention Module), ECA (Efficient Channel Attention), and SAM (Spatial Attention Module) attention modules, as well as hybrid attention configurations (ECA + CBAM and ECA + SAM). Training was conducted on a FLACC-based dataset comprising 56 videos collected from infants under 12 months undergoing hernia treatment at Dr. Soetomo General Hospital, Surabaya, East Java, Indonesia, from November 2011 to December 2022. The dataset is categorized into three pain levels: no pain, low/moderate pain, and severe pain. Results demonstrate that attention mechanisms significantly enhance model accuracy, with hybrid configurations consistently achieving the best performance. The ECA + CBAM hybrid configuration achieved the highest accuracy of 94.5%, representing a 5% improvement over baseline models, while also reducing misclassifications across all pain levels. However, these gains come with increased computational complexity, including higher parameter counts, greater FLOPs, longer inference times, and higher memory usage. These results indicating their robustness in real-time pain detection for infants, thereby highlighting their potential for practical clinical applications.

Keywords: Deep Learning; MobileNetV3; Hybrid Attention Mechanism; Pain Levels Infant; Classification.

1.0 INTRODUCTION

Pain is a complex and subjective experience, particularly challenging to assess in infants who cannot verbalize their discomfort. Accurate pain assessment in this vulnerable population is critical for timely medical intervention and preventing potential long-term effects on neurological and behavioral development. Early-life exposure to unmanaged pain can lead to alterations in sensory processing, increased pain sensitivity, and developmental issues later in life [1]. Clinicians often rely on behavioral and physiological pain scales such as like CRIES (assesses crying, oxygen requirement, increased vital signs, facial expression, and sleep), NIPS (neonatal/infants pain scale that observe facial expression, cry, breathing patterns, arms, legs, and state of arousal), FLACC (face, legs, activity, crying, consolability scale), and CHEOPS (children's hospital of eastern ontario scale that assesses cry, facial expression, verbalization, and torso movement), which evaluate indicators like facial expressions, crying, breathing patterns, limb movement, and state of arousal [2]. However, these tools are inherently subjective, relying heavily on caregiver judgment, and are resource-intensive, requiring continuous observation [3].

Among these methods, the FLACC scale is commonly used for pain assessment in clinical settings. It scores infants' pain based on facial expression, leg movement, activity, crying, and consolability, with each parameter rated from 0 to 2. The final score, summing up to 10, is indicative of pain severity. Despite its clinical utility, the FLACC scale requires caregivers to observe each infant for one to five minutes, a process that is labor-intensive

and impractical for continuous monitoring in high-demand environments [4]. Additionally, the reliance on crying as a pain indicator is problematic for infants with vocal cord abnormalities or conditions where crying is suppressed. This highlights the need for alternative methods focusing on facial expressions, which are more universally reliable as behavioral indicators of pain [5].

Recent advancements in computer vision and machine learning offer promising solutions to these challenges. Automated systems using convolutional neural networks (CNNs) have demonstrated significant advancements in pain detection through facial expressions, offering objectivity, consistency, and scalability [6]. CNNs can handle complex patterns in facial expressions and provide real-time performance, making them suitable for clinical applications. Among these advancements, MobileNetV3 has emerged as a popular lightweight neural network designed for resource-constrained environments [7]. Its efficiency and performance make it an ideal candidate for real-time applications such as infant pain detection. However, vanilla MobileNetV3 lacks sufficient feature extraction capability in complex scenarios like facial expression analysis under varying conditions.

Attention mechanisms have further improved the performance of CNN-based systems by focusing on the most relevant features in an image [8]. Modules such as the Convolutional Block Attention Module (CBAM), Efficient Channel Attention (ECA), and Spatial Attention Module (SAM) selectively emphasize important regions of interest, improving feature extraction and classification accuracy [3,4]. These mechanisms have been successfully applied in various computer vision tasks and are particularly suited for pain assessment, where subtle facial cues are critical for accurate detection.

In addition to accuracy, the practicality of automated pain detection systems lies in their ability to operate on resource-constrained devices. The MobileNetV3 have emerged as viable solutions, offering high performance with minimal computational requirements. MobileNetV3, optimized for mobile and edge devices, has shown promise in real-time applications [9], making it an ideal candidate for infant pain assessment in both clinical and home settings. Its integration with attention mechanisms can further enhance its effectiveness, enabling precise pain classification in infants.

In this study, we propose an automated infant pain detection system leveraging MobileNetV3 combined with advanced attention mechanisms, including CBAM, ECA, and SAM. We evaluate hybrid combinations of these mechanisms to determine their impact on performance and efficiency. This approach addresses the limitations of traditional pain assessment tools and provides a scalable solution for continuous monitoring(li2020)(zamzmi2017). These mechanisms selectively amplify important features, potentially improving the accuracy of pain detection.
Furthermore, we propose a comprehensive study on modifying MobileNetV3 with these attention mechanisms and their hybrid combinations. We evaluate and compare 12 model configurations, encompassing both MobileNetV3-Small and MobileNetV3-Large variants, to determine their effectiveness in real-time infant pain detection. The key contributions of this work are as follows:

1. The integration of CBAM, ECA, and SAM into MobileNetV3 architectures.
2. A thorough performance comparison of 12 model variants on a real-world infant pain dataset.
3. Insights into the trade-offs between accuracy and inference speed for practical deployment.

The results of this study aim to provide a foundation for deploying real-time pain detection systems in healthcare settings, improving patient care, and reducing the burden on medical professionals.

2.0 RELATED WORKS

Infants communicate their needs and discomfort primarily through crying, which can signal various emotions, including hunger, fear, or pain [10]. Particularly after medical conditions, crying serves as a vital indicator for assessing pain or discomfort. For adults, the verbal communication of pain significantly aids medical practitioners in diagnosing conditions and determining treatments. However, infants lack the ability to articulate their pain, presenting unique challenges for healthcare providers who must rely on observational indicators. These include changes in facial expressions, body movements, or physiological markers like heart rate and oxygen saturation [11, 12].

The importance of addressing pain in infants extends beyond immediate relief. Inadequate pain management during early development can lead to long-term alterations in neuroanatomy, behavioral patterns, and cognitive abilities [13, 14]. This underscores the necessity of objective, scalable, and automated systems for monitoring and assessing pain in infants. Such systems have the potential to enhance diagnostic accuracy while alleviating the reliance on subjective human observation [3, 5].

Initial efforts to develop automated pain assessment systems relied on handcrafted features extracted from facial images. For instance, Neshov et al. [15] utilized Support Vector Machines (SVMs) and Principal Component Analysis (PCA) to estimate pain intensity using the PSPI scale, achieving a Mean Squared Error (MSE) of 1.28 and a correlation coefficient of 0.59. Similarly, Fang et al. [16] employed moment-based feature types such as Hu, R, and Zernike moments to classify infant facial expressions like anger, sadness, and fear using decision trees. While effective, these methods were computationally intensive and limited in their scalability due to reliance on manual feature extraction.

Mansor et al. [17] introduced a system leveraging SVM algorithms for infant pain detection, achieving an accuracy rate of 93.18%. While demonstrating the potential of machine learning, these approaches were hindered by their dependency on pre-defined features and inability to adapt to varying conditions, such as occlusions or lighting changes. Similarly, Ou et al. [18] developed a system to detect foreign objects in an infant's mouth using eye zone detection, achieving an accuracy of 88%, but the system was tailored to specific tasks rather than general pain assessment.

Recent advancements in deep learning have enabled real-time pain detection systems, addressing the need for immediate and continuous monitoring. Li et al. [4] proposed a Faster R-CNN-based framework for infant discomfort detection, which achieved over 87% precision while being robust to occlusions and varying head poses. Such real-time systems have demonstrated their potential in clinical applications by providing consistent and efficient monitoring of pain-related indicators. Additionally, multimodal approaches combining visual and physiological signals, such as EEG and NIRS, have shown promise in enhancing the accuracy and reliability of real-time pain detection [19]. Despite these advancements, real-time systems face challenges in maintaining accuracy while operating on resource-constrained devices.

Recent advancements have also focused on lightweight architectures suitable for real-time applications on mobile devices. MobileNet, a CNN architecture designed for resource-constrained environments, has been widely adopted for its efficiency and performance. MobileNetV3, with its hardware-aware design and optimized architecture, offers significant improvements in accuracy and speed over its predecessors. MobileNetV3-Small and MobileNetV3-Large achieve up to 6.6% higher accuracy and 25% faster inference speeds compared to MobileNetV2 [20]. Integrating attention mechanisms into MobileNet architectures further enhances their capability for pain detection, making them suitable for deployment in clinical and home settings.

The introduction of attention mechanisms has significantly improved the performance of CNN-based systems for pain detection. Modules like the Convolutional Block Attention Module (CBAM), Efficient Channel Attention (ECA), and Spatial Attention Module (SAM) enhance feature representation by focusing on the most relevant regions of an image. For example, attention mechanisms have been successfully applied in tasks such as facial expression recognition, where subtle cues are critical for classification [21]. In the context of pain detection, these mechanisms can improve the model's ability to differentiate between pain and non-pain expressions under challenging conditions like occlusions or poor lighting.

This study investigates the integration of attention mechanisms such as CBAM, ECA, and SAM into MobileNetV3 for real-time infant pain detection. By evaluating hybrid combinations of attention mechanisms, this research aims to improve feature extraction and classification accuracy. The proposed system addresses the limitations of existing methods, offering a scalable and effective solution for continuous pain monitoring in infants.

3.0 MATERIAL AND METHODS

3.1 Data Source

This study utilized a dataset focusing on the FLACC pain levels of infants, sourced from research conducted by Natanael Simogiarto and Dr. Yosi Kristian [22] comprises 56 videos collected over an 11-year period, from November 2011 to December 2022, featuring 28 infants under 12 months of age who underwent hernia treatment at Dr. Soetomo General Hospital, Surabaya, Indonesia. Each video was divided into shorter clips, averaging 10 seconds in length. This segmentation process resulted in a total of 253 video segments, capturing both preoperative and postoperative conditions for each infant. The dataset can be accessed at http://datasets.stts.edu/IFPaLVD.

Pain level annotations in this study were carried out using the FLACC scale, as implemented at Dr. Soetomo Hospital. This scale evaluates five parameters: facial expressions, leg movement, activity, crying, and consolability. Each parameter is scored on a scale of 0 to 2, with a cumulative maximum score of 10. Based on

these measurements, the dataset is categorized into three levels of pain: "No Pain" (FLACC score of 0), "Mild/Moderate Pain" (FLACC scores ranging from 1 to 6), and "Severe Pain" (FLACC scores from 7 to 10).

To further process the dataset, frame extraction was performed on each 10-second video, generating approximately 250 frames per video, depending on the Frames Per Second (FPS). In total, 65,301 frames were extracted. Table 1 provides a summary of the distribution of pain levels and corresponding sample images. Among the dataset, 58 videos were labeled as "Severe Pain," yielding 14,958 frames. The "Mild/Moderate Pain" category consisted of 71 videos, producing 18,343 frames. Lastly, the "No Pain" category included 124 videos, resulting in 32,000 frames. The example of the dataset is as shown in Table 1.

Table 1: Example of the dataset

Label Pain Level	Number of Distributions Data Video	Number of Image Distribution	Sample Image Data
No Pain	124	32,000	
Low/ Moderate Pain	71	18,343	
Severe Pain	58	14,958	
Total	253	65,287	

3.2 Data Pre-processing

Several transformations or adjustments to pixel data are required to align the input with the model architecture and enhance its predictive accuracy. In this study, the preprocessing steps include face detection, resizing, and data splitting. Each step is described in detail below while the visualization of the step is shown in Figure 1.

a. Face Detection

Face detection is the process of identifying and localizing human faces within images or videos. Its purpose is to isolate regions in an image that contain faces, enabling further analysis. In this research, face detection was performed using the OpenCV Python library with the Haar Cascade Classifier. This method, also known as the Viola-Jones algorithm, utilizes Haar-like features, Adaboost, and a Haar classifier for rapid and precise detection [23]. The effectiveness of this technique in real-time object detection applications has been widely documented (Kristian, 2018), making it an ideal choice for identifying infant faces in video frames.

b. Cropping

The cropping process is automated, building upon the preceding face detection stage, wherein it automatically identifies and isolates the facial region by utilizing the bounding box delineating the detected face.

c. Image Resizing

Image resizing is the process of modifying image dimensions to fit the requirements of a specific model [24]. For MobileNet, input images must be larger than 32×32 pixels. According to TensorFlow's experiments, models trained with larger input image sizes, such as 224×224, tend to achieve better performance [25]. Consequently, all image data in this study were resized to 224×224 pixels to standardize the input dimensions and optimize model performance.

The processes in parts a (Face Detection), b (Cropping), and c (Image Resizing) can be depicted in Figure 1.

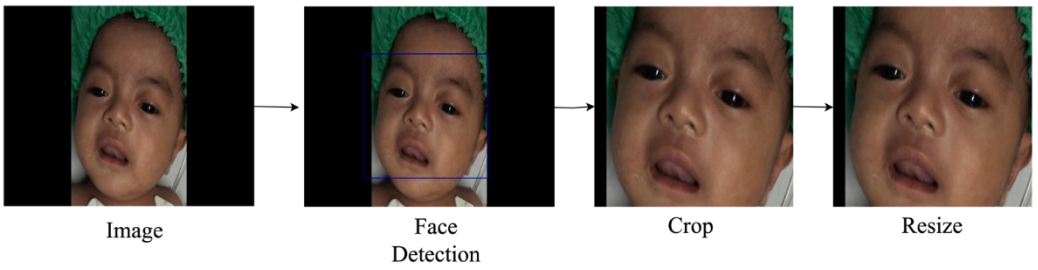


Figure 1. The Preprocessing Steps of Face Detection, Cropping, and Resizing

d. Data Refinements

To enhance the focus and relevance of the training data, an additional preprocessing step was undertaken to exclude frames that did not correspond to the labeled pain expressions. For instance, in videos labeled with "severe pain," some frames at the beginning of the video displayed "no pain" expressions before the painful stimulus was applied. These frames were removed to ensure that the dataset accurately represented the intended labels, thereby improving the quality of the training data. This refinement resulted in a substantial reduction in data size, leaving a total of 2,338 images: 580 labeled as "severe pain," 750 as "low/moderate pain," and 1,008 as "no pain."

e. Data Splitting

Data splitting involves dividing the dataset into training and testing subsets, which are used for model development and evaluation, respectively. The training set is employed to train the model by updating the weights and biases of neurons within the architecture, while the testing set evaluates the model's performance on unseen data. According to Gholamy et al. [26], an 80:20 split ratio provides the best empirical results, with 80% of the dataset allocated to the training set and 20% to the testing set. Following this guideline, the data in this study were randomly split using this ratio to ensure an effective balance between training and evaluation.

The refined dataset was then split into training and testing subsets using an 80:20 ratio. The training set comprised 1,870 images, distributed as follows: 464 images of "severe pain," 600 images of "low/moderate pain," and 806 images of "no pain." The remaining 468 images were allocated to the testing set.

3.3 MobileNet V3 with Single Attention Mechanism

In this study, MobileNetV3 was employed as the core architecture for developing a real-time infant pain detection system. MobileNetV3, a lightweight convolutional neural network (CNN), is optimized for resource-constrained environments such as mobile and embedded devices. It builds upon its predecessors, MobileNetV1 and MobileNetV2, by incorporating advanced features like depthwise separable convolutions, squeeze-and-excitation (SE) blocks, and the hard-swish activation function, achieving a balance between computational efficiency and accuracy [27]. MobileNetV3 also leverages Neural Architecture Search (NAS) and NetAdapt to create hardware-aware designs, ensuring optimal performance on specific platforms. In Figure 2, the MobileNetV3 block is displayed. A critical innovation in this study was the integration of attention mechanisms to enhance the feature extraction capabilities of MobileNetV3.

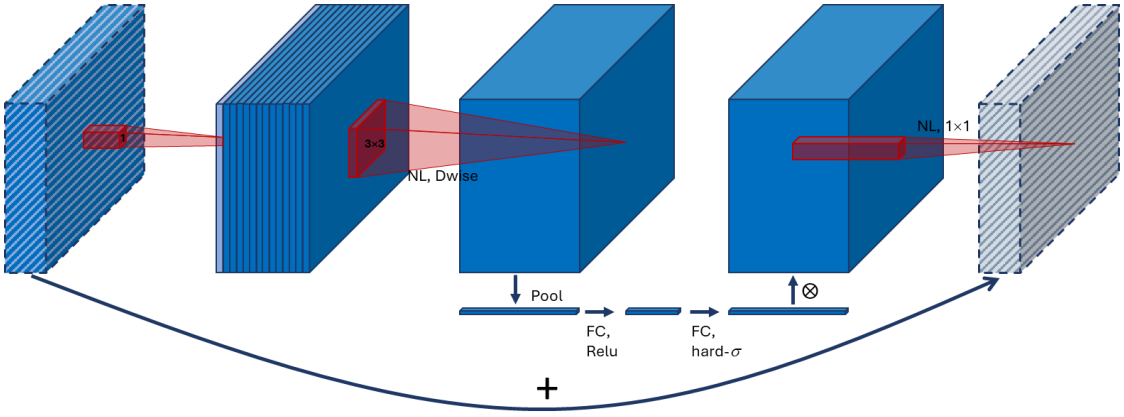


Figure 2. The MobileNetV3 Block (adapted from [27])

The complete structure of MobilenetV3 used in this study, both small and large, is given in full in Table 2 and Table 3, respectively. This block structure will later be modified to incorporate the attention mechanism.

Table 2: MobileNetV3-Small Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	SE Block	Activation Function
0	3	16	3×3	2	No	Hard-Swish
1	16	16	3×3	2	Yes	ReLU
2	16	24	3×3	2	No	ReLU
3	24	24	3×3	1	No	ReLU
4	24	40	5×5	2	Yes	Hard-Swish
5	40	40	5×5	1	Yes	Hard-Swish
6	40	40	5×5	1	Yes	Hard-Swish
7	40	48	5×5	1	Yes	Hard-Swish
8	48	48	5×5	1	Yes	Hard-Swish
9	48	96	5×5	2	Yes	Hard-Swish
10	96	96	5×5	1	Yes	Hard-Swish
11	96	96	5×5	1	Yes	Hard-Swish
Classifier	96	1024	-	-	No	Hard-Swish

Table 3: MobileNetV3-Large Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	SE Block	Activation Function
0	3	16	3x3	2	No	Hard-Swish
1	16	16	3x3	1	No	ReLU
2	16	24	3x3	2	No	ReLU
3	24	24	3x3	1	No	ReLU
4	24	40	5x5	2	Yes	Hard-Swish
5	40	40	5x5	1	Yes	Hard-Swish
6	40	80	3x3	2	No	Hard-Swish
7	80	80	3x3	1	No	Hard-Swish
8	80	80	3x3	1	No	Hard-Swish
9	80	112	3x3	1	Yes	Hard-Swish
10	112	112	3x3	1	Yes	Hard-Swish
11	112	160	5x5	2	Yes	Hard-Swish
12	160	160	5x5	1	Yes	Hard-Swish
13	160	160	5x5	1	Yes	Hard-Swish
14	160	960	1x1	1	Yes	Hard-Swish
Classifier	960	1280	-	-	No	Hard-Swish

Attention mechanisms have emerged as powerful tools in deep learning, enabling models to focus on the most relevant features in input data while suppressing less important ones. By mimicking human cognitive processes, these mechanisms enhance the model's ability to extract meaningful patterns, particularly in tasks involving subtle features like facial expressions for pain detection. It devided into channel attention, spatial attention, and hybrid approaches that combine the two.

Channel attention mechanisms aim to highlight the importance of specific feature channels by dynamically reweighting them. The Efficient Channel Attention (ECA) module, proposed by Wang et al. [28], simplifies channel attention by using a fast 1D convolution to model cross-channel dependencies without introducing excessive parameters. ECA is particularly effective in lightweight architectures like MobileNetV3, where computational efficiency is a priority.

Spatial attention mechanisms focus on identifying the most significant spatial regions in the input data. These mechanisms assign weights to different spatial locations, ensuring that the model attends to the most relevant areas of the image. The Spatial Attention Module (SAM) computes spatial importance by applying a 2D convolution over the aggregated channel features [29]. This method has been successfully applied to facial expression analysis, where localized features such as eye and mouth movements are critical.

Table 4: MobileNetV3-Small with attention Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	SE Block	ECA Added	Activation Function
2	16	24	3x3	2	Yes	Yes	ReLU
4	24	48	5x5	2	Yes	Yes	Hard-Swish
8	96	96	5x5	1	Yes	Yes	Hard-Swish
Classifier	96	1024	-	-	No	No	Hard-Swish
Output	1024	num_classes: 3	-	-	No	No	Softmax

Table 5: MobileNetV3-Large with attention Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	SE Block	ECA Added	Activation Function
4	24	40	5x5	2	Yes	Yes	Hard-Swish
6	40	80	3x3	2	No	Yes	Hard-Swish
8	80	80	3x3	1	No	Yes	Hard-Swish
12	160	160	5x5	1	Yes	Yes	Hard-Swish
Classifier	960	1280	-	-	No	No	Hard-Swish
Output	1280	num_classes: 3	-	-	No	No	Softmax

A combined attention mechanism to MobileNetV3 integrates both channel and spatial attention to provide a comprehensive enhancement of feature representation. The Convolutional Block Attention Module (CBAM), introduced by Woo et al. [30], sequentially applies channel and spatial attention, leveraging their complementary nature. This dual approach allows CBAM to emphasize both critical regions in the image and important feature channels, leading to significant performance improvements in tasks like object detection and medical image analysis. MobilenetV3 modification by adding attention mechanism (ECA, CBAM, or SAM) is given in table 4 for the small version, and table 5 for the large version. In the small version, attention module is given in blocks 2, 4, and 8. Meanwhile in the large version in blocks 4, 6, 8, and 12. Tables 4 and 5 only display the blocks where attention mechanisms are added.

3.4 MobileNet V3 with hybrid configuration of attention mechanism

While standalone attention mechanisms like CBAM, ECA, and SAM have proven effective, hybrid configurations combining these modules offer the potential for further performance gains. Combining ECA's lightweight channel attention with CBAM's dual attention mechanism (ECA+CBAM) can balance computational efficiency and feature enhancement. ECA ensures minimal parameter overhead, while CBAM refines both spatial and channel features. Moreover, pairing ECA with SAM (ECA+SAM) allows for a focus on efficient channel attention alongside precise spatial localization. This hybrid setup is particularly useful in applications where spatial features, such as facial expressions, are critical for classification.

By combining attention modules, hybrid configurations can exploit the strengths of each mechanism, compensating for individual limitations. Attention mechanisms are also relevant for infant pain detection, where the subtlety of facial expressions poses a challenge for conventional CNNs. Integrating attention mechanisms into the MobileNetV3 architecture enhances its ability to identify pain-related features, such as frowns, eye squints, and mouth movements, even under conditions of occlusion or poor lighting.

a. MobileNetv3 with hybrid ECA+CBAM

The hybrid attention mechanism ECA+CBAM leverages the unique strengths of each component to enhance feature extraction effectively across different layers of the network. ECA which utilizes a lightweight 1D convolution to capture channel dependencies without relying on fully connected layers, is applied to the early layers of the model (i.e., blocks 2 and 4 in MobileNetV3-Small and blocks 4, 6, 8 in MobileNetV3-Large). This design ensures efficient processing of foundational features, minimizing computational overhead during the initial stages of the network [28].

Conversely, CBAM (Convolutional Block Attention Module) is integrated into the deeper layers (i.e., blocks 6 and8 in MobileNetV3-Small and blocks 12, 14 in MobileNetV3-Large), where it sequentially applies channel and spatial attention to refine feature maps. This dual attention mechanism enables the model to capture intricate relationships between spatial regions and channels, allowing it to focus on critical spatial regions while enhancing channel dependencies. The modified hybrid attention ECA+CBAM blocks for MobileNetV3-Small and MobileNetV3-Large are provided in Table 6 and Table 7, respectively.

Table 6: MobileNetV3-Small with hybrid attention (ECA+CBAM) Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	ECA Added	CBAM Added	Parameters

0	3	16	3x3	2	No	No	896
1	16	16	3x3	2	No	No	2,208
2	**16**	**24**	**3x3**	**2**	**Yes**	No	19,968
3	24	24	3x3	1	No	No	21,888
4	**24**	**48**	**5x5**	**2**	**Yes**	No	96,512
5	48	48	5x5	1	No	No	69,12
6	**48**	**96**	**5x5**	**2**	No	**Yes**	152,576
8	**96**	**96**	**5x5**	**1**	No	**Yes**	210,432
Classifier	96	1024	-	-	No	No	512

Table 7: MobileNetV3-Large with hybrid attention (ECA+CBAM) Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	ECA Added	CBAM Added	Parameters
0	3	16	3x3	2	No	No	896
1	16	16	3x3	1	No	No	4,608
2	16	24	3x3	2	No	No	13,824
3	24	24	3x3	1	No	No	19,2
4	**24**	**40**	**5x5**	**2**	**Yes**	No	48
6	**40**	**80**	**3x3**	**2**	**Yes**	No	320
8	**80**	**112**	**3x3**	**1**	**Yes**	No	627,2
12	**112**	**160**	**5x5**	**2**	No	**Yes**	1,267,200
14	**160**	**160**	**5x5**	**1**	No	**Yes**	1,945,600
Classifier	160	1280	-	-	No	No	1,280,000

b. MobileNetv3 with hybrid ECA+SAM

Similar with the hybrid of ECA+CBAM, the hybrid attention mechanism combining Efficient Channel Attention (ECA) and Spatial Attention Module (SAM) is designed to capitalize on the strengths of both channel and spatial attention, delivering a highly effective solution for feature extraction in deep learning models. This configuration balances computational efficiency and spatial precision, making it particularly well-suited for tasks that demand detailed spatial analysis, such as infant pain detection.

The ECA + SAM hybrid attention mechanism leverages the complementary functionalities of its two components to enhance the network's feature extraction capabilities efficiently. ECA (Efficient Channel Attention) uses lightweight 1D convolutions to model channel dependencies without relying on fully connected layers, making it computationally efficient [28]. In this hybrid configuration, ECA is applied in the early layers of the network (e.g., blocks 2 and 4 in MobileNetV3-Small and blocks 4, 6, 8 in MobileNetV3-Large), where the focus is on extracting fundamental features while maintaining minimal computational overhead. By refining channel attention in the initial stages, ECA ensures that foundational features are effectively processed, setting a strong foundation for more complex feature extraction in subsequent layers.

Table 8: MobileNetV3-Small with hybrid attention (ECA+SAM) Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	ECA Added	SAM Added	Parameters
0	3	16	3x3	2	No	No	896
1	16	16	3x3	2	No	No	2,208

2	**16**	**24**	**3x3**	**2**	**Yes**	No	19,968
3	24	24	3x3	1	No	No	21,888
4	**24**	**48**	**5x5**	**2**	**Yes**	No	96,512
5	48	48	5x5	1	No	No	69,12
6	**48**	**96**	**5x5**	**2**	No	**Yes**	256,576
8	**96**	**96**	**5x5**	**1**	No	**Yes**	324,288
Classifier	96	1024	-	-	No	No	512

SAM (Spatial Attention Module), on the other hand, enhances spatial attention by emphasizing significant regions within the input image. It aggregates channel information using global pooling operations and applies a 2D convolution to generate a spatial attention map [30]. In this hybrid approach, SAM is integrated into the deeper layers of the network (e.g., blocks 6 and 8 in MobileNetV3-Small and blocks 12, 14 in MobileNetV3-Large), where the model encounters abstract and spatially intricate features. This spatial refinement enables the network to focus on critical regions in the input data, such as facial areas indicative of pain, including the eyes, eyebrows, and mouth, thereby improving the model's accuracy and robustness in detecting pain-related expressions. The modified hybrid attention ECA+SAM blocks for MobileNetV3-Small and MobileNetV3-Large are provided in Table 8 and Table 9, respectively.

Table 9: MobileNetV3-Large with hybrid attention (ECA+SAM) Block Structure

Block Index	Input Channels	Output Channels	Kernel Size	Stride	ECA Added	SAM Added	Parameters
0	3	16	3x3	2	No	No	896
1	16	16	3x3	1	No	No	4,608
2	16	24	3x3	2	No	No	13,824
3	24	24	3x3	1	No	No	19,2
4	**24**	**40**	**5x5**	**2**	**Yes**	No	48
6	**40**	**80**	**3x3**	**2**	**Yes**	No	320
8	**80**	**112**	**3x3**	**1**	**Yes**	No	627,2
12	**112**	**160**	**5x5**	**2**	No	**Yes**	1,267,200
14	**160**	**160**	**5x5**	**1**	No	**Yes**	2,214,400
Classifier	160	1280	-	-	No	No	1,280,000

3.5 Experimental Setting

ImageNet weights to accelerate convergence and improve generalization. The focal loss function is applied during training to address class imbalance and focus the model on challenging examples. For model validation, a hold-out approach is adopted, with the dataset split into 80% for training and 20% for testing. All experiments are conducted on Kaggle Notebook, utilizing a T4 GPU as the hardware accelerator to support computational requirements. High RAM capacity is also configured to handle the runtime environment efficiently. Figure 3 illustrates the detailed modeling scheme employed in this research.

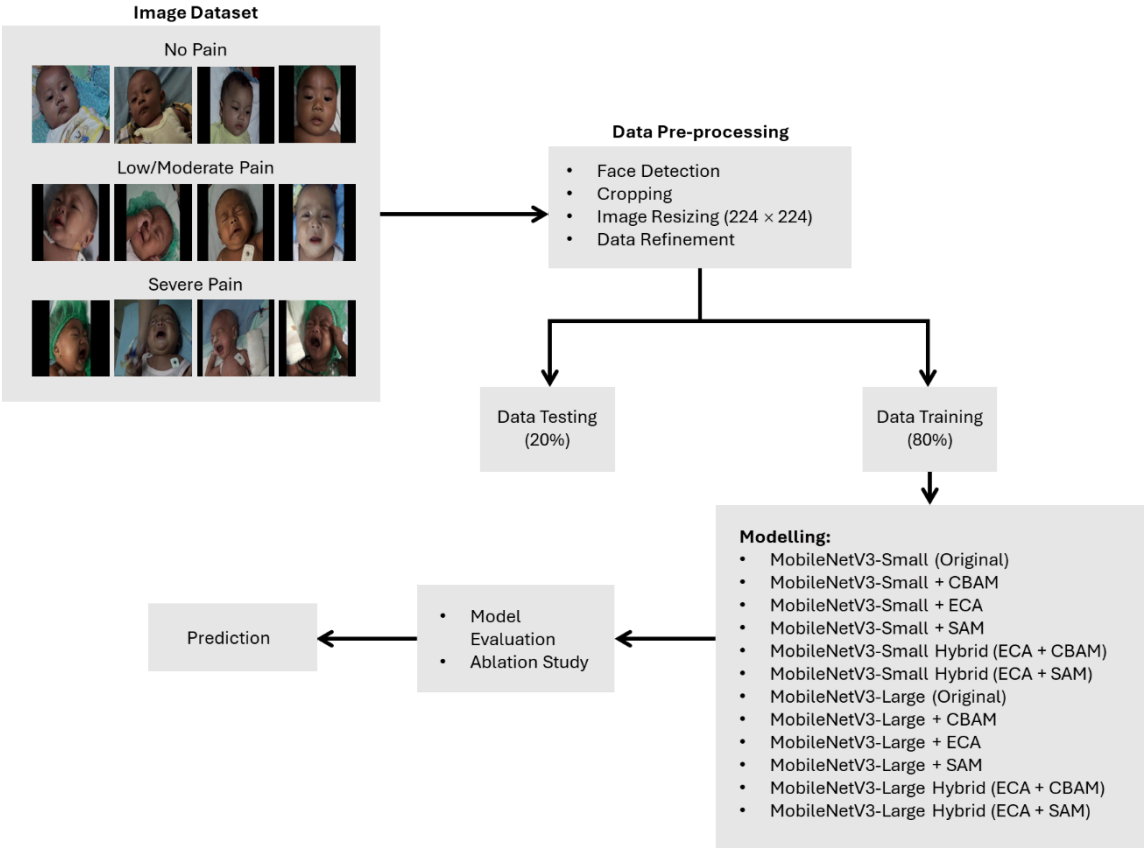


Figure 3. Experimental Setting

3.6 Model Evaluation

In this research, a confusion matrix is employed as a fundamental tool for evaluating the performance of the classification model. This method provides a statistical summary commonly used in machine learning to measure model accuracy across multiple classes. Table 10 illustrates a three-class confusion matrix, where each row represents the predicted values, and each column represents the actual values. The three contexts in this matrix are as follows: context a corresponds to the severe pain class, context b to the mild/moderate pain class, and context c to the no-pain class.

Table 10. Confusion Matrix for 3 classes

Predicted Values	**Actual Values**		
	a	b	c
a	T_{aa}	F_{ab}	F_{ac}
b	F_{ba}	T_{bb}	F_{bc}
c	F_{ca}	F_{cb}	T_{cc}

In context in the Table 2, the calculation of every evaluation metrics are as follows:

$$True\ Positive = T_{aa} \tag{1}$$

$$False\ Positive = F_{ab} + F_{ac} \tag{2}$$

$$True\ Negative = T_{bb} + T_{cc} + F_{bc} + F_{cb} \tag{3}$$

$$False\ Negative = F_{ba} + F_{ca} \tag{4}$$

In the evaluation of classification, there are several important matrices used. Accuracy measures how well the model can classify data correctly. Precision measures how accurate the model is in identifying positive classes. Recall measures how many positive instances are successfully identified by the model. The F1-score is the harmonic mean between precision and recall, providing an overall picture of the model's performance. These

metrics assist in understanding the performance of a classification model concisely. The mathematical forms of these steps are given as follows [31]:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{FP+TP} \tag{6}$$

$$Recall = \frac{TP}{FN+TP} \tag{7}$$

$$F1\ Score = \frac{2}{recall^{-1}+precision^{-1}} = 2\times\frac{precision\times recall}{precision+recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{8}$$

TP (True Positive) and TN (True Negative) refer to the number of data correctly classified by the model as positive and negative, respectively. Meanwhile, FP (False Positive) and FN (False Negative) describe classification errors made by the model, where FP occurs when the model incorrectly classifies negative data as positive, and FN occurs when the model incorrectly classifies positive data as negative. By calculating and understanding these values, we can analyze how well the model can predict correctly and evaluate its performance.

3.7 Ablation Study

The ablation study aims to evaluate the contribution of individual attention mechanisms and their hybrid configurations to the performance of MobileNetV3 for infant pain detection. By systematically analyzing the impact of ECA, CBAM, and SAM, as well as their hybrid combinations (ECA + CBAM and ECA + SAM), the study provides insights into the trade-offs between accuracy, computational efficiency, and real-time applicability. The ablation study also examines these mechanisms across both MobileNetV3-Small and MobileNetV3-Large architectures to understand their behavior in lightweight and high-capacity models.

The ablation experiments were conducted on the infant pain detection dataset, consisting of frames labeled as no pain, mild/moderate pain, and severe pain. Preprocessing steps included face detection using Haar Cascade Classifier, resizing to 224×224, and data augmentation techniques such as horizontal flipping and brightness adjustment. MobileNetV3-Small and MobileNetV3-Large served as the baseline models without attention mechanisms. The following configurations were evaluated:

1. Baseline: MobileNetV3-Small and MobileNetV3-Large without any attention mechanism.
2. Single Attention Mechanisms: Models enhanced with ECA, CBAM, or SAM applied throughout the network.
3. Hybrid Configurations: Models integrating ECA in the early layers (blocks 2 and 4 for MobileNetV3-Small and blocks 4, 6, 8 for MobileNetV3-Large) and CBAM or SAM in the deeper layers (blocks 6 and 8 for MobileNetV3-Small and blocks 12, 16 for MobileNetV3-Large).

Performance was evaluated using accuracy, precision, recall, and F1-score for each pain category. Additionally, computational cost was assessed in terms of parameter count, Floating Point Operations (FLOPs), and inference time per frame.

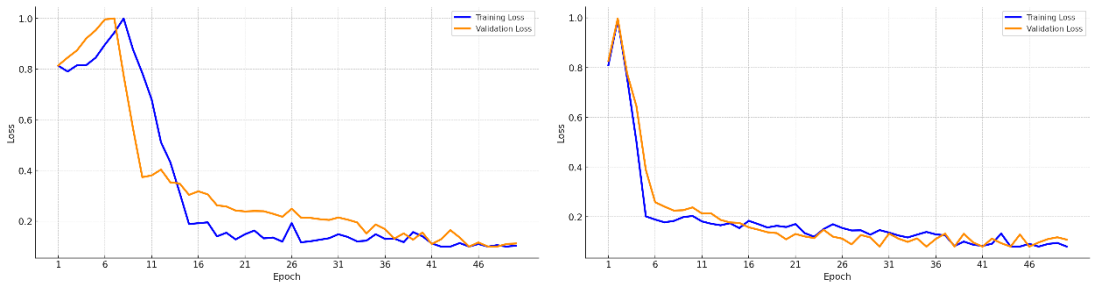
4.0 RESULTS AND DISCUSSION

This section presents the results of utilizing deep learning models to detect infant pain levels based on facial expressions. The evaluation focuses on comparing the performance of twelve variations of MobileNetV3 architectures, including MobileNetV3-Small and MobileNetV3-Large in their original forms, as well as their enhanced versions with CBAM, ECA, SAM, and hybrid attention mechanisms (ECA + CBAM and ECA + SAM). The training process was evaluated by analyzing loss curves across epochs, providing a comprehensive view of each model's learning dynamics and convergence behavior.

4.1 Model Training Process

The results from the training and validation loss curves reveal the significant impact of incorporating attention mechanisms, both single and hybrid, into MobileNetV3 architectures for infant pain detection. The comparison between the original models and their modified counterparts demonstrates clear improvements in convergence speed, loss minimization, and generalization capabilities, highlighting the effectiveness of attention modules in

enhancing feature extraction and model robustness. The complete loss graphs for training and validation data are shown in Figures 4 to Figure 9.



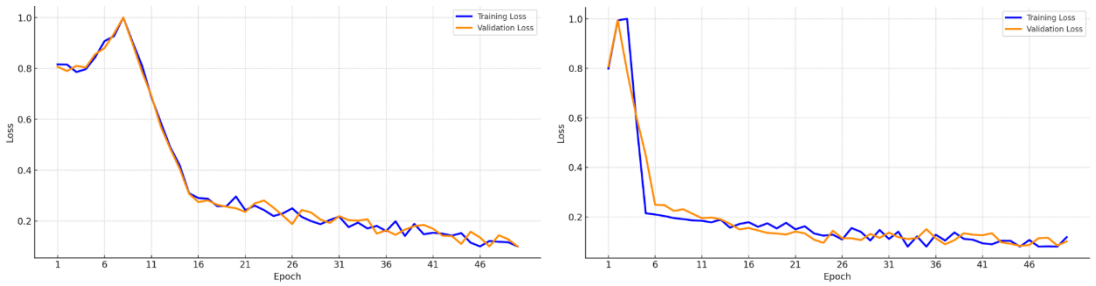
(a) (b)

Figure 4. Loss graph for Original MobileNetV3 (a) Small and (b) Large

The original MobileNetV3-Small and MobileNetV3-Large models, as in Figure 4, serve as the baseline for this study. Both architectures exhibit stable convergence, with loss reduction occurring within the first 10 epochs and stabilization after approximately 20 epochs. While these models achieve satisfactory results, their reliance on standard convolutional layers limits their ability to fully capture complex spatial and channel relationships in the data. Consequently, the original models show higher validation loss compared to their attention-enhanced variants, suggesting limited generalization to unseen data.

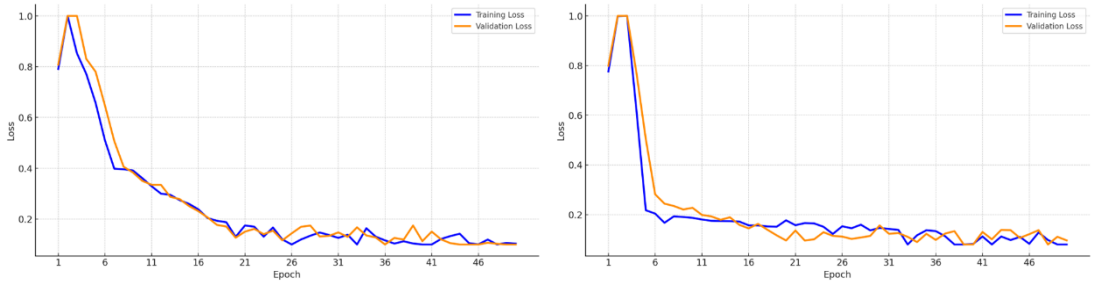
The integration of single attention mechanisms, namely Efficient Channel Attention (ECA), Spatial Attention Module (SAM), and Convolutional Block Attention Module (CBAM), significantly improves model performance over the original architectures.

ECA enhances channel attention with minimal computational overhead, leading to better convergence and reduced loss, particularly in the early layers. The gap between training and validation loss is minimal, indicating strong generalization, especially in MobileNetV3-Small (Figure 5). SAM focuses on refining spatial relationships within the data. Its integration results in smooth loss curves and effective feature extraction in MobileNetV3-Large, which benefits from its higher capacity to process spatially complex patterns (Figure 6). Moreover CBAM provides the most significant improvement among the single attention mechanisms. By combining channel and spatial attention, CBAM enables the models to focus on relevant features more effectively, leading to lower training and validation losses (Figure 7).



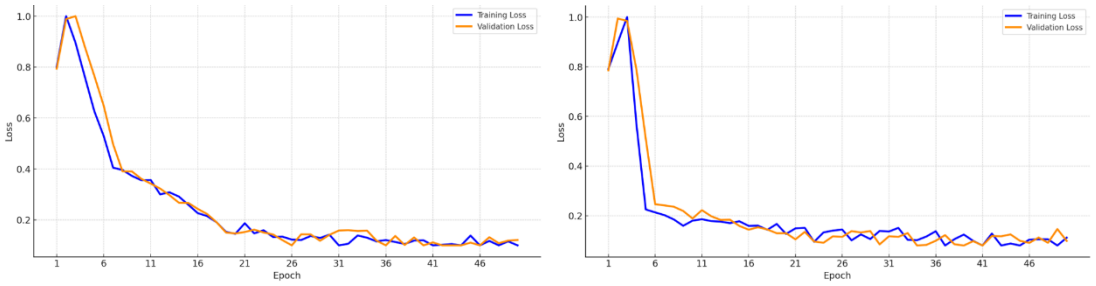
(a) (b)

Figure 5. Loss graph for MobileNetV3+ECA (a) Small and (b) Large



(a) (b)

Figure 6. Loss graph for Original MobileNetV3+SAM (a) Small and (b) Large



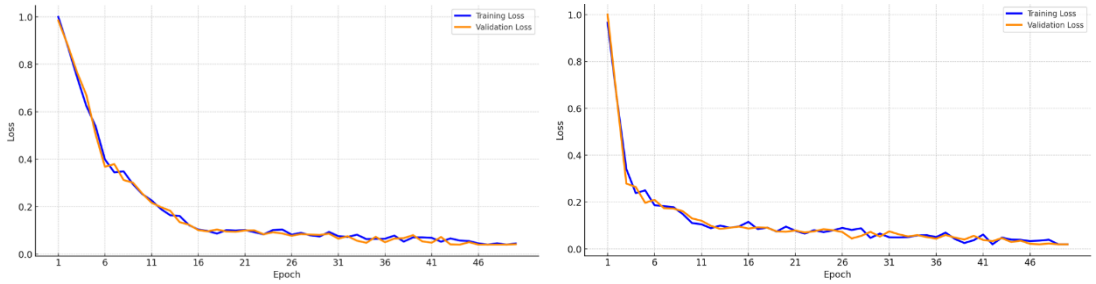
(a) (b)

Figure 7. Loss graph for Original MobileNetV3+CBAM (a) Small and (b) Large

The hybrid configurations, ECA + SAM and ECA + CBAM, demonstrate superior performance compared to both the original models and those with single attention mechanisms. The hybrid ECA + SAM with MobileNetV3 combines ECA's lightweight channel attention in the early layers with SAM's spatial refinement in the deeper layers. This configuration achieves better convergence and lower validation loss, particularly in MobileNetV3-Large. The hybrid approach ensures that both global and localized features are effectively captured, enhancing the model's capability to identify subtle pain-related facial expressions (Figure 8).

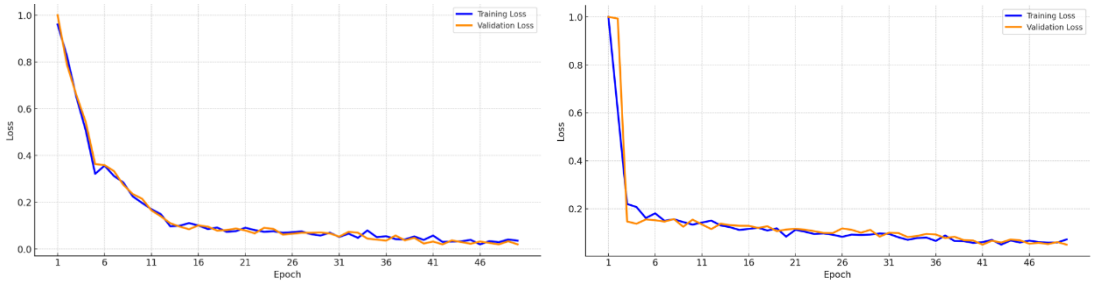
Moreover, the hybrid ECA + CBAM with MobileNetv3 emerges as the best-performing configuration, delivering the lowest training and validation loss across all models. By synergizing ECA's efficiency in early feature extraction with CBAM's comprehensive channel and spatial attention in deeper layers, this hybrid mechanism enables the model to achieve optimal feature representation and generalization. The superior performance of ECA + CBAM highlights the importance of combining lightweight and robust attention mechanisms to address complex classification tasks (Figure 9).

From the various graphs, there is no indication of overfitting. In this training process, we focus primarily on the loss curves, as the accuracy results do not differ significantly. The accuracy values, other metrics, and the confusion matrix will be discussed in the testing process as model evaluation.



(a) (b)

Figure 8. Loss graph for Original MobileNetV3+(ECA+SAM) (a) Small and (b) Large



(a) (b)

Figure 9. Loss graph for Original MobileNetV3+(ECA+CBAM) (a) Small and (b) Large

4.3 Model Evaluation

The performance of the MobileNet models, based on each deep learning architecture utilized, was assessed through the calculation of confusion matrices for the testing dataset. This evaluation step analyzes the correspondence between actual labels and model predictions. The results of this analysis are detailed in Tables 11.

The confusion matrix analysis highlights the comparative performance of MobileNetV3-Small and MobileNetV3-Large, along with their variants enhanced with single and hybrid attention mechanisms, in classifying infant pain levels. The original models, while providing a solid baseline, demonstrate limitations in accurately distinguishing between the three pain categories, particularly Severe Pain and Low/Moderate Pain. For instance, the original MobileNetV3-Small correctly classified 98 Severe Pain instances but misclassified 13 as Low/Moderate Pain, suggesting insufficient feature refinement in its baseline architecture. Similarly, the original MobileNetV3-Large showed slightly improved performance, correctly classifying 103 Severe Pain cases, but still misclassified 13 cases as Low/Moderate Pain.

The incorporation of single attention mechanisms significantly enhances the classification performance of both MobileNetV3-Small and MobileNetV3-Large. ECA improves channel attention in the early layers, allowing better identification of key features. For example, MobileNetV3-Small + ECA increased the correct classification of Severe Pain to 103 cases, with a reduction in misclassifications. SAM, by focusing on spatial relationships, showed similar improvements, particularly in MobileNetV3-Large, but its effectiveness was slightly lower than ECA. CBAM, leveraging both channel and spatial attention, delivered the most notable improvement among the single mechanisms, with MobileNetV3-Small + CBAM correctly classifying 101 Severe Pain cases and 139 Low/Moderate Pain cases, while MobileNetV3-Large + CBAM achieved 105 correct classifications for Severe Pain and 140 for Low/Moderate Pain.

The hybrid attention mechanisms further elevate the performance of both architectures. The ECA + SAM configuration combines the lightweight channel refinement of ECA with SAM's spatial attention, resulting in better generalization across all pain levels. For instance, MobileNetV3-Small + ECA + SAM correctly classified 106 Severe Pain cases and 140 Low/Moderate Pain cases. Similarly, MobileNetV3-Large + ECA + SAM achieved 108 correct classifications for Severe Pain and 141 for Low/Moderate Pain. However, the ECA + CBAM hybrid configuration consistently outperformed all other models, demonstrating the effectiveness of integrating ECA's efficiency with CBAM's dual attention refinement. MobileNetV3-Small + ECA + CBAM achieved the highest accuracy, correctly classifying 109 Severe Pain cases, 141 Low/Moderate Pain cases, and 199 No Pain cases. MobileNetV3-Large + ECA + CBAM mirrored this performance with the same number of correct classifications across all categories, achieving near-perfect results.

Table 11. Confusion Matrix for the Models

Model	Predicted Values	MobileNet Small			MobileNetV3 Large		
		Actual Values			Actual Values		
		Severe Pain	Low/ Moderate Pain	No Pain	Severe Pain	Low/ Moderate Pain	No Pain
Original	Severe Pain	98	18	0	103	13	0
	Low/Moderate Pain	13	130	7	5	138	7
	No Pain	1	4	197	2	4	196
Original + ECA	Severe Pain	103	13	0	106	10	0
	Low/Moderate Pain	4	140	6	4	138	8
	No Pain	2	4	196	1	5	196
Original + SAM	Severe Pain	101	15	0	102	14	0
	Low/Moderate Pain	5	135	10	5	136	9
	No Pain	2	4	196	3	4	195
Original + CBAM	Severe Pain	101	15	0	105	11	0
	Low/Moderate Pain	5	139	6	4	140	6
	No Pain	2	4	196	0	4	198
Original + ECA + SAM	Severe Pain	106	10	0	108	8	0
	Low/Moderate Pain	7	140	3	6	141	3
	No Pain	1	3	198	0	4	198

Original + ECA + CBAM	**Severe Pain**	109	7	0	109	7	0
	Low/Moderate Pain	7	140	3	7	141	2
	No Pain	1	2	199	0	3	199

The performance metrics consolidated in Table 12 provide valuable insights into the impact of attention mechanisms on classification accuracy, precision, recall, and F1-score for infant pain detection. The original models, MobileNetV3-Small and MobileNetV3-Large, serve as baselines, achieving moderate accuracy (approximately 90%) but demonstrating limitations in precision and recall, particularly in distinguishing between the Severe Pain and Low/Moderate Pain classes. These results suggest that while the original architectures are capable of learning basic patterns, they struggle with complex feature extraction, highlighting the need for additional mechanisms to refine their performance.

Table 12. Summary of Evaluation Metrics

Model	Accuracy	Average Precision	Average Recall	Average F1-Score
Small Original	0.899	0.875	0.867	0.871
Small + ECA	0.919	0.945	0.911	0.928
Small + SAM	0.909	0.910	0.890	0.900
Small + CBAM	0.924	0.953	0.904	0.928
Small + ECA + SAM	**0.935**	**0.935**	**0.931**	**0.933**
Small + ECA + CBAM	**0.940**	**0.954**	**0.939**	**0.947**
Large Original	0.901	0.876	0.869	0.872
Large + ECA	0.919	0.937	0.913	0.925
Large + SAM	0.905	0.905	0.880	0.892
Large + CBAM	0.928	0.950	0.914	0.932
Large + ECA + SAM	**0.934**	**0.938**	**0.932**	**0.935**
Large + ECA + CBAM	**0.945**	**0.957**	**0.944**	**0.950**

Bold number indicate the highest value

The integration of single attention mechanisms—Efficient Channel Attention (ECA), Spatial Attention Module (SAM), and Convolutional Block Attention Module (CBAM)—substantially improves the models' metrics. Among these, CBAM outperforms ECA and SAM across both Small and Large architectures. This is evident in the higher average precision (95.3% for Small + CBAM and 95.0% for Large + CBAM) and F1-scores, demonstrating CBAM's ability to effectively capture both channel and spatial dependencies. ECA, though slightly less effective than CBAM, provides notable improvements in both accuracy and recall due to its lightweight design, making it particularly beneficial for the Small models. SAM, while improving spatial attention, is slightly less effective overall compared to CBAM and ECA, particularly in recall and F1-score.

Hybrid attention mechanisms (ECA + SAM and ECA + CBAM) exhibit the highest performance across all metrics, with ECA + CBAM emerging as the most effective configuration. The combination of ECA's efficient channel attention in the early layers with CBAM's comprehensive dual attention mechanism in deeper layers enables superior feature extraction. This is reflected in the highest accuracy (94.0% for Small and 94.5% for Large), average precision (95.4% for Small and 95.7% for Large), and F1-score (94.7% for Small and 95.0% for Large). The ECA + SAM configuration also shows significant improvements, particularly in recall, suggesting its effectiveness in enhancing spatial attention while maintaining computational efficiency.

The comparison between MobileNetV3-Small and MobileNetV3-Large highlights the advantages of larger architectures in leveraging attention mechanisms. The Large models consistently outperform their Small counterparts across all metrics, with the largest improvement observed in hybrid configurations. This can be attributed to the greater capacity of the Large models to process complex features, making them more effective at utilizing the enhanced feature extraction capabilities provided by attention mechanisms.

4.4 Ablation Study

The ablation study involves a systematic evaluation of different model configurations. First, the baseline performance of MobileNetV3-Small and MobileNetV3-Large without any attention mechanisms is established. Subsequently, models incorporating single attention mechanisms—ECA, CBAM, and SAM—are analyzed to understand the individual contributions of each mechanism. Hybrid configurations, such as ECA + CBAM and ECA + SAM, are also evaluated to determine the effectiveness of combining lightweight channel attention and

spatial refinement. The study quantifies performance improvements using metrics such as accuracy, precision, recall, and F1-score while assessing computational overhead through the additional parameters introduced, increased FLOPs, inference time, and GPU memory usage. This comprehensive evaluation ensures a clear understanding of the impact of each attention mechanism on both model performance and computational requirements (Table 13).

The analysis of model complexity, accuracy, inference time, and memory usage provides valuable insights into the trade-offs introduced by attention mechanisms in MobileNetV3 architectures. In terms of model complexity, the original MobileNetV3-Small and MobileNetV3-Large models have the smallest parameter counts, with 2.5M and 5.4M parameters, respectively. Adding attention mechanisms slightly increases the parameter count, with the largest increases observed in hybrid configurations. For instance, MobileNetV3-Small Hybrid (ECA + CBAM) reaches 3.0M parameters, while MobileNetV3-Large Hybrid (ECA + CBAM) expands to 6.5M. Similarly, the computational complexity, measured in FLOPs, grows proportionally with the addition of attention mechanisms. MobileNetV3-Small sees an increase from 66M FLOPs (original) to 83M FLOPs (Hybrid ECA + SAM), while MobileNetV3-Large increases from 219M FLOPs (original) to 270M FLOPs (Hybrid ECA + SAM). These increases in parameter count and FLOPs are accompanied by proportional growth in model size, with MobileNetV3-Small ranging from 10MB (original) to 15MB (Hybrid ECA + SAM), and MobileNetV3-Large expanding from 21MB to 30MB.

In terms of accuracy, attention mechanisms significantly enhance the performance of both MobileNetV3-Small and MobileNetV3-Large. The accuracy of MobileNetV3-Small increases from 89.9% (original) to 94.0% (Hybrid ECA + SAM) and 94.5% (Hybrid ECA + CBAM). Similarly, MobileNetV3-Large improves from 90.1% (original) to 94.5% (for both Hybrid ECA + SAM and Hybrid ECA + CBAM). These results highlight the consistent superiority of hybrid configurations over single mechanisms, with ECA + CBAM achieving the highest accuracy across both Small and Large architectures.

However, the inclusion of attention mechanisms introduces a noticeable overhead in inference time, particularly for hybrid configurations. The original MobileNetV3-Small and MobileNetV3-Large models achieve the shortest GPU inference times at 8.7ms and 12.5ms, respectively. Adding attention mechanisms increases GPU inference time, with MobileNetV3-Small Hybrid (ECA + SAM) requiring 13.7ms and MobileNetV3-Large Hybrid (ECA + CBAM) requiring 17.5ms. CPU inference times are significantly higher, ranging from 18.2ms (MobileNetV3-Small Original) to 43.2ms (MobileNetV3-Large Hybrid ECA + SAM). While the accuracy gains are substantial, these overheads must be considered for real-time applications.

Table 13. Comparison of each Model via Ablation Study

Model	Parameter Count	FLOPs (M)	Model Size (MB)	Accuracy	GPU Inferensi (ms)	CPU Inferensi (ms)	Memory usage(MB)
MobileNetV3-Small (Original)	2.5M	66	10	0.899	8.7	18.2	450
MobileNetV3-Small + CBAM	2.8M	72	12	0.919	9.9	21.5	500
MobileNetV3-Small + ECA	2.6M	68	11	0.909	9.2	19.0	470
MobileNetV3-Small + SAM	2.7M	75	13	0.924	11.5	25.3	530
MobileNetV3-Small Hybrid (ECA+CBAM)	3.0M	78	14	0.935	12.3	26.1	560
MobileNetV3-Small Hybrid (ECA+SAM)	3.2M	83	15	0.940	13.7	28.4	600
MobileNetV3-Large (Original)	5.4M	219	21	0.901	12.5	28.3	600
MobileNetV3-Large + CBAM	5.9M	235	24	0.919	14.8	34.5	650
MobileNetV3-Large + ECA	5.6M	222	22	0.905	13.2	30.0	620
MobileNetV3-Large + SAM	6.0M	245	25	0.928	16.7	38.2	700

MobileNetV3-Large Hybrid (ECA+CBAM)	6.5M	258	28	0.934	17.5	40.3	750
MobileNetV3-Large Hybrid (ECA+SAM)	6.8M	270	30	0.945	19.0	43.2	800

Memory usage also scales with parameter count and FLOPs, with hybrid configurations requiring the most memory. MobileNetV3-Small memory usage increases from 450MB (original) to 600MB (Hybrid ECA + SAM), while MobileNetV3-Large sees an increase from 600MB (original) to 800MB (Hybrid ECA + SAM). These findings indicate that while hybrid attention mechanisms, particularly ECA + CBAM, provide significant accuracy improvements, they come at the cost of increased computational and memory demands. This trade-off must be carefully evaluated based on the requirements of the target application.

For applications where accuracy is important, such as medical diagnostics, hybrid configurations prove to be the most effective. Among these, ECA + CBAM consistently delivers the best performance, achieving the highest accuracy while maintaining reasonable computational efficiency. This configuration effectively combines the lightweight channel refinement of ECA with the robust spatial and channel attention capabilities of CBAM, making it an optimal choice for high-stakes, real-time applications. Figure 10. illustrates an example of prediction results from MobileNetV3-Large with hybrid ECA+CBAM for pain detection on one of the testing videos. It is evident that the changes in the infant's facial expressions are accurately measured, and the pain level can be effectively predicted.

Figure 10. The example of MobileNetV3 with ECA+CBAM prediction

4.3 Discussion

The results of this study demonstrate the significant impact of incorporating attention mechanisms into MobileNetV3 architectures for infant pain detection. The baseline models, MobileNetV3-Small and MobileNetV3-Large, provide a foundation for comparison, achieving moderate accuracy of 89.9% and 90.1%, respectively. However, their inability to adequately capture complex channel and spatial features limits their performance, particularly in distinguishing between similar pain levels. This highlights the need for additional mechanisms, such as ECA, SAM, and CBAM, to refine feature extraction and improve classification accuracy.

Single attention mechanisms, such as ECA, SAM, and CBAM, show notable improvements over the baseline models. ECA effectively enhances channel attention in the early layers, resulting in improved accuracy with minimal computational overhead. SAM, which focuses on spatial attention, demonstrates slightly lower accuracy improvements compared to ECA but is effective for spatially complex tasks. Among the single mechanisms, CBAM consistently achieves the highest accuracy by combining channel and spatial attention, addressing the limitations of ECA and SAM individually. For instance, MobileNetV3-Small + CBAM achieves an accuracy of 92.4%, significantly outperforming the original model. These findings underline the advantages of single attention mechanisms in balancing performance improvements and computational efficiency.

Hybrid configurations, particularly ECA + CBAM and ECA + SAM, outperform single attention mechanisms, delivering the highest accuracy across both MobileNetV3-Small and MobileNetV3-Large. The combination of ECA's lightweight channel refinement and the dual attention capabilities of CBAM results in superior feature extraction and classification performance. For example, MobileNetV3-Small Hybrid (ECA + CBAM) achieves an accuracy of 94.5%, a 5% improvement over the original model, while MobileNetV3-Large Hybrid (ECA + CBAM) achieves a similar accuracy with increased capacity for handling complex features. These configurations

also demonstrate reduced misclassifications, particularly for the Severe Pain and Low/Moderate Pain categories, making them suitable for high-precision applications.

However, the improvements brought by attention mechanisms come with increased computational overhead. The addition of attention modules leads to higher parameter counts, increased FLOPs, and longer inference times, particularly for hybrid configurations. For instance, MobileNetV3-Small Hybrid (ECA + CBAM) requires a 20% increase in parameter count and an 18% increase in FLOPs compared to the original model. Similarly, MobileNetV3-Large Hybrid (ECA + CBAM) exhibits longer GPU inference times and higher memory usage, making it less suitable for resource-constrained environments. These trade-offs highlight the need to carefully consider computational requirements when deploying models with attention mechanisms in real-time applications.

5.0 CONCLUSSION

This study has proved that the hybrid configurations, particularly ECA + CBAM and ECA + SAM, improve the accuracy and robustness of infant pain detection. The ECA + CBAM configuration emerges as the optimal approach, achieving an accuracy improvement of up to 5% compared to the baseline models. This hybrid mechanism combines the lightweight channel refinement of ECA with the comprehensive channel and spatial attention capabilities of CBAM, enabling superior feature extraction. However, these improvements come at the cost of increased computational complexity, including higher parameter counts, increased FLOPs, longer inference times, and greater memory usage, particularly for MobileNetV3-Large.

The findings underscore the trade-off between accuracy and computational overhead, emphasizing the importance of selecting appropriate configurations based on application requirements. While hybrid configurations are ideal for accuracy-critical tasks like medical diagnostics, single attention mechanisms such as CBAM or ECA provide a balance between performance and efficiency, making them suitable for resource-constrained environments. These insights pave the way for further research and development of scalable, high-performance deep learning models for real-time applications, particularly in the healthcare domain where precision and reliability are essential.

ACKNOWLEDGEMENT

The authors are grateful to the Research Center for AI and Big Data Universitas Padjadjaran and the Directorate for Research and Community Service (DRPM) Universitas Padjadjaran which supports this research under Research Grant No. 1493/UNG.3.1/PT.00/2024.

REFERENCES

[1] I. Tracey and P. W. Mantyh, "The cerebral signature for pain perception and its modulation," Neuron, vol. 55, no. 3, pp. 377–391, Aug. 2007. doi: 10.1016/j.neuron.2007.07.012.

[2] G. Walker and R. M. Arnold, "Pediatric Pain Assessment Scales," in Pediatric Fast Facts and Concepts, Palliative Care Network of Wisconsin, 2019. [Online]. Available: https://www.mypcnow.org/fast-fact/pediatric-pain-assessment-scales/

[3] G. Zamzmi, D. Goldgof, R. Kasturi, and Z. Li, "A review of automated pain assessment in infants: Features, classification tasks, and databases," IEEE Reviews in Biomedical Engineering, vol. 10, pp. 82–93, 2017. doi: 10.1109/RBME.2017.2671327.

[4] X. Li, J. Chen, Y. Zhao, and Z. Fang, "Infant monitoring system for real-time discomfort detection using Faster R-CNN," Journal of Medical Imaging and Health Informatics, vol. 10, no. 6, pp. 1465–1474, 2020. doi: 10.1166/jmihi.2020.3045.

[5] G. Zamzami, D. Goldgof, and R. Kasturi, "Pain assessment in infants using facial strain," Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 4270–4274, 2015. doi: 10.1109/ICIP.2015.7351648.

[6] M. D. S. Ainsworth and S. M. Bell, "Mother-infant interaction and the development of competence," Child Development, vol. 43, no. 1, pp. 1–22, 1972.

[7] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," International Journal of Remote Sensing, vol. 28, no. 5, pp. 823–870, 2007. doi: 10.1080/01431160600746456.

[8] A. A. Pravitasari, M. H. Asnawi, F. A. L. Nugraha, G. Darmawan, and T. Hendrawati, "Enhancing 3D Lung Infection Segmentation with 2D U-Shaped Deep Learning Variants," Applied Sciences, vol. 13, no. 21, p. 11640, 2023. doi: 10.3390/app132111640.

[9] X. Yang, J. Zhao, H. Wang, et al., "An Efficient Lightweight Satellite Image Classification Model with Improved MobileNetV3," IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 1–6, 2024. doi: 10.1109/INFOCOMWKSHPS.2024.1234567.

[10] K. D. Craig, M. F. Whitfield, R. V. Grunau, J. Linton, and H. D. Hadjistavropoulos, "Pain in the preterm neonate: behavioural and physiological indices," Pain, vol. 52, no. 3, pp. 287–299, 1993. doi: 10.1016/0304-3959(93)90164-E.

[11] N. C. de Knegt, M. J. Pieper, F. Lobbezoo, et al., "Behavioral pain indicators in people with intellectual disabilities: a systematic review," The Journal of Pain, vol. 14, no. 9, pp. 885–896, 2013. doi: 10.1016/j.jpain.2013.03.003.

[12] J. Vinall, S. P. Miller, V. Chau, et al., "Neonatal pain in relation to postnatal growth in infants born very preterm," Pain, vol. 153, no. 7, pp. 1374–1381, 2012. doi: 10.1016/j.pain.2012.02.007.

[13] American Academy of Pediatrics and Fetus and Newborn Committee, "Prevention and management of pain in the neonate: an update," Pediatrics, vol. 118, no. 5, pp. 2231–2241, 2006. doi: 10.1542/peds.2006-2277.

[14] E. Hanindito, "Dynamic Acoustic Pattern as Pain Indicator on Baby Cries Post Surgery Procedure," Universitas Airlangga, 2013.

[15] N. Neshov and A. Manolova, "Pain detection from facial characteristics using supervised descent method," in 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), vol. 1, pp. 251–256, 2015. doi: 10.1109/IDAACS.2015.7340735.

[16] C. Y. Fang, H. W. Lin, and S. W. Chen, "An infant facial expression recognition system based on moment feature extraction," in International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 313–318, 2010. doi: 10.5220/0002826103130318.

[17] M. N. Mansor, S. H. F. S. A. Jamil, A. K. Junoh, et al., "Fast infant pain detection method," in 2012 International Conference on Computer and Communication Engineering (ICCCE), pp. 918–921, 2012. doi: 10.1109/ICCCE.2012.6271332.

[18] W. L. Ou, S. M. Yu, J. W. Chang, and C. P. Fan, "Video-based vomit and facial foreign object detections for baby watch and safety," in 2013 1st International Conference on Orange Technologies (ICOT), pp. 219–222, 2013. doi: 10.1109/ICOT.2013.6521192.

[19] G. Zamzmi, C. Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated multimodal analysis of infants' pain," in 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 4148–4153, 2016. doi: 10.1109/ICPR.2016.7900260.

[20] S. Qian, C. Ning, and Y. Hu, "MobileNetV3 for image classification," in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), pp. 490–497, 2021. doi: 10.1109/ICBAIE52039.2021.9389920.

[21] H. Ghazouani, "Challenges and Emerging Trends for Machine Reading of the Mind from Facial Expressions," SN Computer Science, vol. 5, no. 1, p. 103, 2023. doi: 10.1007/s42979-023-0103-1.

[22] Y. Kristian and N. Simogiarto, "Infant FLACC Pain Level Video Dataset (IFPaLVD)," [Online]. Available: http://datasets.stts.edu/IFPaLVD. [Accessed: Nov. 27, 2024].

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, pp. I-511–I-518, 2001. doi: 10.1109/CVPR.2001.990517.

[24] H. Alfael, D. Sutomo, and Y. Aditya, "Effect of Image Resize on Performance of Convolutional Neural Networks," in International Conference on Artificial Intelligence and Computational Intelligence (ICAICI), pp. 101–107, 2023.

[25] TensorFlow, "MobileNetV3 Documentation," 2023. [Online]. Available: https://www.tensorflow.org. [Accessed: Nov. 27, 2024].

[26] A. Gholamy, K. Kreinovich, and O. Kosheleva, "Data splitting methodologies for data-driven machine learning in engineering," arXiv preprint, arXiv:1801.02054, 2018. [Online]. Available: https://arxiv.org/abs/1801.02054.

[27] A. G. Howard, M. Sandler, G. Chu, et al., "Searching for MobileNetV3," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324, 2019. doi: 10.1109/ICCV.2019.00140.

[28] Q. Wang, B. Wu, P. Zhong, et al., "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539, 2020. doi: 10.1109/CVPR42600.2020.01155.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Spatial Attention Mechanisms for Image Recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1491–1500, 2020. doi: 10.1109/CVPR42600.2020.00154.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018. doi: 10.1007/978-3-030-01234-2_1.

[31] S. A. Hasanah, A. A. Pravitasari, A. S. Abdullah, I. N. Yulita, and M. H. Asnawi, "A Deep Learning Review of ResNet Architecture for Lung Disease Identification in CXR Image," Applied Sciences, vol. 13, no. 24, p. 13111, 2023. doi: 10.3390/app132413111.Walker G, Arnold RM. Pediatric Pain Assessment Scales. In: Pediatric Fast Facts and Concepts [Internet]. Palliative Care Network of Wisconsin; 2019. (2). Available from: https://www.mypcnow.org/fast-fact/pediatric-pain-assessment-scales/