ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Improving the Background Awareness for Identifying Child Begging by Integrating Foreground and Background Dual Image Classifiers

Midhu Jean Joseph[1], Virach Sornlertlamvanich[2] and Thatsanee Charoenporn[3]

ABSTRACT

Image recognition technologies have found widespread application in domains such as object detection, medical imaging, and autonomous driving. Beyond these applications, they offer promising potential in tackling social challenges, such as identifying children involved in begging activities through intelligent analysis of visual data. The existing data on child begging is scarce and often presents outdated information. Our research demonstrates that image classification models, such as CNN, VGG16, and EfficientNet, can be effectively trained on images captured from public cameras to identify children engaged in begging. This data can be used for quicker and more effective interventions. To further improve detection, we integrated background learning into our approach. The classification models may struggle to distinguish between similar features across environments (for example, misidentifying a poorly dressed child in a ghetto as a beggar). Incorporating background learning can help mitigate such errors by providing contextual understanding. We further proposed an "Integrated Dual Image Classifier" to learn the background and foreground separately and then subsequently combine both model prediction probabilities. In this method, background understanding is incorporated with the foreground prediction for recognition. The test accuracy results from the integrated dual model approach showed a reduction in false negatives and false positives (Failure to detect actual instances and incorrect identification of false instances as accurate, respectively), with test accuracy above 70%.

Article information:

Keywords: Image Classification, Artificial Intelligence (AI), Convolutional Neural Network (CNN), Child Begging, Transfer Learning, VGG16, EfficientNet, Deep learning models, Image Background Learning, Image Foreground Learning

Article history:
Received: June 30, 2024
Revised: November 21, 2024
Accepted: July 7, 2025
Published: July 26, 2025
(Online)

DOI: 10.37936/ecti-cit.2025193.257362

1. INTRODUCTION

Artificial Intelligence (AI) is a field of computer science technology that focuses on the simulation of human intelligence by machines. Machine models are trained with real-world data, enabling them to perform tasks that traditionally require human intelligence. AI can streamline complex, time-intensive tasks, delivering rapid results with high accuracy, and is transforming industries like healthcare, finance, and agriculture. In healthcare, AI is used for disease diagnosis and treatment recommendations based on clinical data. In finance, AI is used for fraud detection and stock market predictions by analyzing patterns and utilizing AI-driven algorithms. In agriculture, crop monitoring using drone imagery is a valuable application. AI has the potential to be a transformative tool for social good, addressing challenges and creating meaningful solutions that benefit society. For example, in disaster response, AI is used for flood mapping to identify disaster-affected areas using satellite images, enabling faster response and efficient resource allocation. In this research, we are leveraging AI to address child begging, a social issue in many developing nations.

According to the Registrar General of India, ap-

[1] The author is with the Faculty of Data Science, Musashino University, Japan., Email: g2551002@stu.musashino-u.ac.jp
[2] The author is with the Asia AI Institute (AAII), Faculty of Data Science, Musashino University, Japan; Faculty of Informatics, Burapha University, Thailand., Email: virach@musashino-u.ac.jp
[3] The author is with the Asia AI Institute (AAII), Faculty of Data Science, Mussashino University, Japan., Email: thatsane@musashino-u.ac.jp
[3] Corresponding author: thatsane@musashino-u.ac.jp

proximately 45,000 children are involved in begging [1]. Studies show that many beggars in Bangladesh are children, and begging is often organized [2]. In many developing nations, it is common to see children begging on the streets and at traffic signals. Begging activities, usually part of organized operations, are sometimes linked to serious social crimes such as child abduction, forced labour, and exploitation, severely endangering a child's safety and well-being. Traditionally, data on begging is collected through surveys, which require manual effort and result in time delays. AI-powered data collection can automate the identification process while addressing challenges related to limited data availability and the need for time-sensitive data acquisition. The data collected can enable effective interventions and rehabilitation by the respective authorities, which is the motivation for our research.

Computer vision can be used to identify children involved in begging by analyzing image data from public cameras, which can be achieved through image classification models trained on task-specific datasets. In this study, CNN (Convolutional Neural Network) and variations of CNNs (VGG16, EfficientNet) were trained using publicly available images (Begging and Normal child), and the results were compared to identify the best fit model for the task (EfficientNet). We conducted additional experiments for background/context learning by training the model using image categories and subcategories based on the image background. Background learning can mirror human behavior by analyzing the context and recognizing complex scenarios, which has demonstrated improved results, highlighting the effectiveness of background learning. Finally, we introduced an integrated dual-model approach, where background analysis was performed using background-trained models, followed by foreground analysis using foreground-trained models. Then we combined both judgments to detect the begging child. This dual-layered approach showed an enhancement in detection accuracy by considering both subject characteristics and environmental context.

This paper outlines three sets of experiments in detail –

(1) To identify the best-fit image classification model among CNN, VGG16, and EfficientNet for this task.

(2) To fine-tune the best model (EfficientNet), incorporating background learning, by categories and subcategories using training data categorized by background types associated with begging and non-begging children (normal).

(3) Training the best-fit model to learn backgrounds and foreground features separately and subsequently integrating both for final prediction, enhancing the reliability and accuracy.

For the first two experiments, a dataset of 600 images was taken from publicly available sources for training the image classification models. For the third experiment, we utilized two distinct sets of images for training the image classification models: a set of 200 foreground-only images depicting “Begging” and “Normal” activities, and a set of 300 background images sourced from publicly available internet images, representing various environments. The experiments are designed and conducted with multiple combinations of different numbers of neural network layers, datasets, and architectures.

The paper is organized into the following sections. Firstly, a comprehensive literature review and social impact analysis are presented to contextualize the societal challenges that can be addressed through image classification methods, followed by an introduction to the methodology used in the study, including the experimental setup and underlying data. The following section explains the detailed experimentation results and provides a comparison of the experimental outcomes. The Final section touches upon the discussions, implications, and future work.

2. LITERATURE REVIEW AND SOCIAL IMPACT

Many developing countries are facing an increase in begging activities; at times, children are forced to take up begging due to various vulnerability factors. There are reports on recruitment agents misusing the opportunity as a business. Beggary points to a social disorganization. Studies based on Bangladeshi society show that a significant amount of the beggar population is children, and this has been executed as an organized activity [2].

Child begging falls under the categorization of child abuse as per WHO, which defines child maltreatment as those under the age of 18 years suffering from physical and/or emotional ill treatment, negligence, and commercial or other exploitation, which results in actual or potential harm to the child's health, survival, and development [3]. At the root level, poverty and social vulnerabilities, for example, unemployment, mental health issues, discrimination, illiteracy, and disability, are identified as factors leading to begging [4]. To reduce this, the government and policymakers must adopt approaches to prioritize social assistance and support welfare programs for rehabilitation. Children who are forced to beg are at high risk, which calls for immediate action. To prevent child begging, proper planning and regulations must be in place, along with rehabilitation efforts. For the rehabilitation process, human actions are required; the process of identifying children involved in begging can be assisted and well-automated using an AI-supported machine. Building on image detection, AI models can be trained to identify instances of child begging through background-aware learning. Categorizing the contextual environments of begging children into distinct groups enhances the model's training efficiency and accuracy.

There are studies in various domains in which deep learning, particularly CNNs, have been applied to address complex classification and detection tasks. For instance, Farhat *et al.* (2020) developed a deep learning technical solution for child labor detection using a CNN and Transfer learning method, studying child labor videos for classification [5]. Their research used CNNs to recognize instances of child labor in various industrial settings by analyzing video frames. The similarity between this paper and our research area lies in the nature of tasks involving the detection of vulnerable children in challenging environments, for which CNNs are well-suited. In our research, we adapted this architecture in the initial experiments and designed our CNN models with additional layers, adjusted hyperparameters, and background learning to improve accuracy in identifying begging children.

Paredes *et al.* (2023) utilized CNNs to develop an emotion recognition algorithm specifically tailored for individuals with Down syndrome, achieving higher accuracy through transfer learning and microexpression analysis [6]. If compared to our research area of identifying begging children, although the context differs, the underlying challenge of identifying patterns is similar. Our research aims to leverage these capabilities, addressing the unique challenges posed by child begging and further enhancing the model. Building upon advancements in deep learning, a study by Oluwalade *et al.* (2021) explores the deployment of neural network architectures for activity recognition using data from smartphones and smartwatches [7]. The research evaluates models such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Convolutional LSTM (ConvLSTM), concluding that convolutional models, particularly when utilizing accelerometer data from smartwatches, exhibit superior performance in classifying a range of activities. Mekala *et al.* (2016) explored foreground-background classification for activity detection in surveillance videos, comparing seven background subtraction methods [8]. Gradient-based techniques achieved the highest accuracy (97.44%), while ViBe ensured spatial coherence. Challenges included ghost elimination, shadow removal, and parameter tuning. The study highlights the importance of optimizing background subtraction for improved ROI detection in surveillance-based activity recognition. Poor background modeling can lead to false positives (e.g., noise detection) and false negatives (missed objects). Xiao *et al.* (2020) analyze how object recognition models rely on image backgrounds [9]. Their findings show that models can classify images using backgrounds alone, but often misclassify when backgrounds change adversarially (resulting in up to 87.5% error), and that more accurate models depend less on backgrounds. They introduce ImageNet-9 (IN-9) to study foreground-background separation and propose a challenge to improve model robustness. The paper suggests that models should be designed to balance background dependence appropriately, rather than ignoring it entirely.

The above background learning utilized training images with variations such as "only background" with the foreground separated, "only foreground" with the background separated, "mixed backgrounds", and "no foregrounds", where the foreground is removed and replaced with black. These may result in less representation of real-world scenarios.

This paper discusses the automatic detection of a begging child using an image classification model trained on background learning. Achieved by introducing a dual-model approach, to learn foreground and background features separately and then integrating both. The background model focused on background features, and the Foreground model on foreground activities. The background model categorizes environments such as markets, streets, and traffic, while the foreground model identifies specific actions like begging postures. By integrating the probabilistic outputs of these models, we achieve a more robust and context-aware classification due to improved generalization, which results in better accuracy in complex environments. Training FG and BG separately helps mitigate adversarial misclassification risks by preventing over-reliance on one feature set. This model adaptation can be utilized by communities, institutions, and the government to identify children who beg and then streamline the rehabilitation process for these children, ultimately bringing about a significant change in the lives of the affected children and, consequently, the broader society.

3. METHODOLOGY

A technical solution for identifying children involved in begging is proposed using a CNN (Convolutional Neural Network). The study has extended the application of child begging identification by incorporating standard CNN techniques, including data augmentation, hyperparameter tuning, background learning techniques, and integrating foreground and background learning.

CNN is a general term that refers to a class of Deep Neural Networks specifically designed for grid-like structured data and uses convolutional layers as a base. A convolutional layer performs convolutional operations to detect patterns and features in an image, enabling the identification of a begging child. Convolutional operations are, in simple terms, a dot product between two functions, where one function (filter) is a learnable parameter 'Kernel', and the other metric is the restricted portion of the image—the Kernel slides across the image, producing an image representation [10]. CNN can have multiple

Layers of Convolutional layers (Conv), Pooling Layers (PL), and Fully Connected layers (FC), with the flexibility of having different numbers of these layers and different filter sizes (Kernels). VGG16 has a specific CNN architecture with 16 layers (13 conv, 3 FC layers, a fixed 3x3 convolutional filter). This model captures complex patterns and fine-grained features for image identification. On the other hand, the EfficientNet model's architecture optimizes in a principled manner by scaling the depth, width, and resolution simultaneously. This model is more efficient in terms of memory and computational requirements.

CNN and its variants, such as VGG16 and EfficientNet, are utilized in the experiments to evaluate their effectiveness for the task and identify the best-fit model. EfficientNet was recognized as the best-fit model for this task. The selection of these models was based on their proven capabilities to deliver high accuracy and efficiency.

Data augmentation was strategically applied during model training, enhancing images through horizontal flipping, shearing, and zooming, thereby increasing the dataset's diversity and improving model performance. Demonstrating that even established methods can be powerful when applied thoughtfully and in novel contexts.

Hyperparameter tuning helps optimize the model output and maximize the performance of the CNN. Hyperparameters are predefined values assigned to the algorithm's parameters before the training process, defining the architecture used in the research. In our experiments, the hyperparameters were adjusted to evaluate the model's performance using multiple combinations, aiming to identify the optimal configuration. Several hyperparameters, including image dimensions, batch size, and filter sizes across convolutional layers (32, 64, 128, 256 filters), as well as specifying the number of units in dense layers (512, 256, 14, 2), were utilized in the study.

The model was trained using Google Colab, leveraging a T4 GPU with 12 GB of VRAM for accelerated computations. The system environment included Python 3.1 and TensorFlow/PyTorch (version 2.15). Training was conducted on Google Colab, a cloud-based environment, ensuring efficient processing while managing computational constraints.

Further enhancement of the image classification model is achieved by introducing background learning through "Subcategories" and subsequently through "Sub-labelling", which was completed by categorizing the training images according to the background, for example, (Begging child in traffic, street, and market; normal child in traffic, street, and market). Context learning helps the classification models to detect child begging in public places with improved efficiency. Additionally, a dual-model approach was introduced by training the best-fit image classification model to learn foreground and background features separately, and subsequently combining their predictions to enhance classification accuracy.

A dataset of thumbnails of images featuring begging children, normal children, and various backgrounds was manually classified using keyword searches in the Google search bar, such as "Begging child India", "Begging child Mumbai", and "School areas in Mumbai", among others, and manually annotated for each image. These keywords were used to ensure the dataset includes region-specific photos that reflect the real-world context of child begging in India and also helped to capture images that are more representative of the problem, including variations in age, posture, surroundings, and social conditions. This approach ensures a diverse and contextually relevant dataset for training an accurate model.

The biggest challenge in the research was manually labelling the images to create the dataset for training, through manual observation, such as gestures or postures that might resemble begging.

(i) Group Behaviour: Identifies whether the individual is alone or part of a group, disabled (e.g., multiple children performing similar actions), often observable in public areas like traffic signals, streets, markets

(ii) Key Items: Bowls, crutches, or containers being used for collecting money or food. Absence of Play Materials: Lack of toys or props that signify play or casual activities.

(iii) Condition of Clothes: Torn, dirty, or underdressed clothing that may indicate economic hardship. Body Hygiene: Signs of neglect, such as unkempt hair or visible malnutrition.

The study was structured around three key categories of experimental evaluation. The accuracy (training & validation) was compared to conclude from the results (Tables 3 to 8). Accuracy gives a general sense of performance. Training accuracy is a measure of how accurately a machine learning model predicts outcomes on the data on which it is trained, allowing it to learn patterns and make predictions. It reflects how well the model has learned patterns and relationships within the training data. Validation accuracy is a measure of how well the trained model performs on a separate "validation" dataset that model has not seen during training and is monitored during training. This metric is crucial because it indicates how well the model generalizes to new, unseen data. Accuracy was chosen as a simple and effective evaluation method. Finally, we conducted testing on 38 independent images, and test accuracy was measured. Test accuracy is checked after training to evaluate how well the model generalizes to new, real-world data (Table 1).

3.1 Proposed Experiments and Methods

Experiment (1): Identifying the best image classification model: Established CNN and Variations of CNNs (VGG16, EfficientNet) are used to train and learn the features of an image to identify a begging child. The best-fit Model identified (EfficientNet Functional Architecture) was further used in experiments (2) and (3).

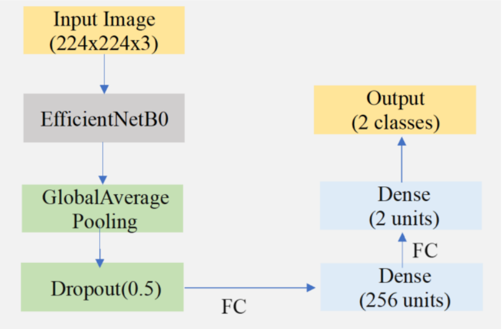


Fig.1: *EfficientNet_2 Functional architecture.*

Experiment (2): The best fit model, EfficientNet to learn image background techniques: EfficientNet model identified in experiment (1) is further trained to distinguish the backgrounds (such as traffic signal, crowded markets etc.) using sub categorizing and sub-labelling of the training images as per the backgrounds, where the child begging is prominent and least common. Learning about the surroundings of a begging child helps with contextual understanding.

Experiment (3): The best fit model to separately train the backgrounds and foregrounds and integrate both models: The foreground alone ("Begging", "Normal") and multiple backgrounds where the child begging is highly prominent and less likely ('Market', 'Street', 'Traffic', 'Over Bridge', 'Ghettos', 'House', 'KidsPlayArea', 'School'), where trained separately to create separate EfficientNet models (Foreground model and Background model). The probabilistic output of these two models was integrated for the final classification of the image.

3.2 Underlying Data

There were no pre-existing research datasets to learn from or build upon for this issue (Child begging identification). The ground truth is the manually verified labels assigned to each image in the dataset. These images were used solely for academic research purposes to extract visual features for model development, ensuring ethical use in compliance with research standards.

(i) A unique thumbnail images dataset featuring Begging child, Normal child, different environments like traffic areas, market, schools were created manually by using keyword search in google search bar like "Begging child India", "Normal child India", "Begging child Mumbai", "Begging child Delhi", "Normal child in street India", "children in traffic areas India", "traffic areas in Delhi", "School areas in Mumbai" etc. (Fig. 2 - 5 below)

(ii) Each image was annotated as either 'begging' or 'normal' and surrounding categories (like school, market, traffic), based on objective observations and human verification. This ground truth is essential for training and evaluating image classification models to ensure they accurately identify children who are begging.

(iii) For the first two experiments, a data set of 600 images (300 of begging child and 300 of Normal child) was created, and for the third experiment, 500 images (200 foreground only and 300 different background images) were used; however, different sample sizes of images were used in each experiment.

(iv) For the third experiment to create the 200-image data set for foreground-only images, an online platform (remove.bg) was used. (Fig. 6 below).

(v) In the experiments model architecture, the image was resized to 150*150 pixels for CNN and 224*224 pixels for VGG16 and EfficientNet.

(vi) 38 unseen images during training and validation were tested for different models' comparison (Table 1).

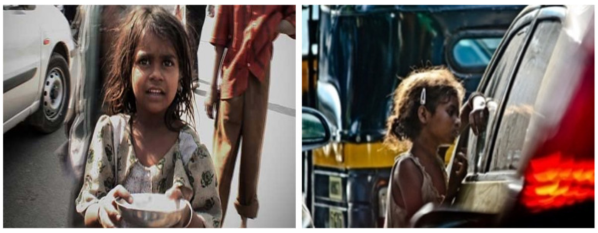


Fig.2: *Sample Images of a begging child.*

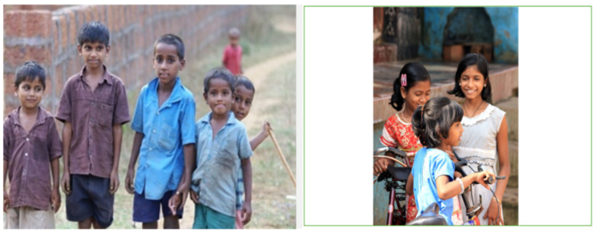


Fig.3: *Sample Images of a normal child.*

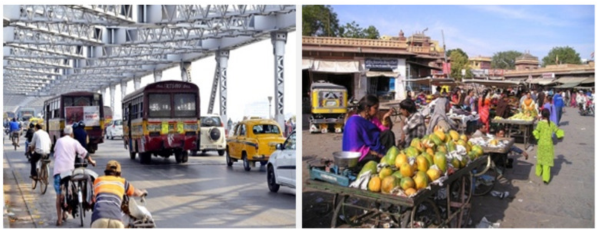
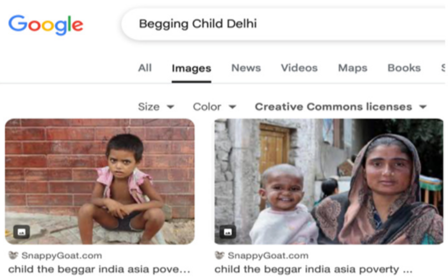


Fig.4: *Sample Images of different backgrounds.*



Begging Child Delhi

SnappyGoat.com
child the beggar india asia pove...

SnappyGoat.com
child the beggar india asia poverty ...

Fig.5: *Sample Images of keyword search.*

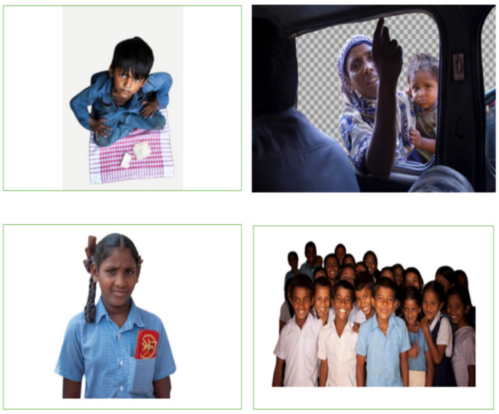


Fig.6: *Sample Images of foreground-only images (2 of begging and 2 of normal, respectively). Image Sources: which permits free use - non-commercial purposes.*

4. EXPERIMENTS AND RESULTS

4.1 Experiment (1): Identification of the best image classification model for the identification of the begging child.

Three established image classification models with architectural variations were trained and tested under this experiment: a) CNN, b) VGG16, c) EfficientNet.

4.1.1 Experiment (1a) Basic CNN Model:

Three different sub-experiments of CNN with multiple combinations of Conv, PL, FC, and Batch Normalization(BN) were used for training and validation to choose the best CNN model (Details in Table 3)

CNN_1: Initialized the experiment with 3 Conv, 3 PL, 1 FC layers and 200 image sample size (100 training and 100 validation)

CNN_2: From the previous experiment, increased by an additional layer of Conv, PL, and FC along with Batch Normalization to improve feature extraction and accelerate training. 200-image sample size (100 training and 100 validation)

CNN_3: Increased to 8 convolutional layers, the training image size was increased to 500 to improve the model's generalization with more diverse examples, and 100 samples were used for validation. Tested the model on 38 unseen images during training and validation, and the test accuracy is shown in Table 1.

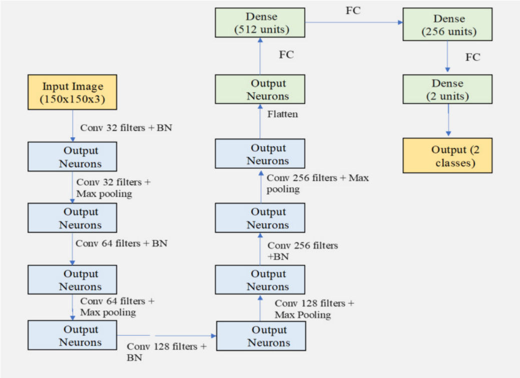


Fig.7: *CNN architecture (Experiment CNN_3).*

<u>Observations from CNN model experiments:</u> Increasing conv, FC layers, and training data size improved validation accuracy and reduced validation loss; however, these models showed overfitting (Overfitting: Training accuracy > Validation accuracy; Validation loss: Discrepancy between the prediction and the actual value). CNN experiments gave a validation accuracy range of 70% to 77% (Training and validation accuracy in Table 3).

4.1.2 Experiment (1b) VGG16 Model:

VGG16 used transfer learning and varied training size for three different experiments for training and validation to choose the best VGG16 model (Details in Table 4)

VGG16 Base: The base model of 13 Conv and 3 FC layers was used to initialize the Experiment, and a 600-image sample size (500 training and 100 validation)

VGG16 Transfer 1 (CNN): The VGG16 pre-trained model serves as the base model (transfer learning), leveraging the pre-trained model and adding custom classifier layers to improve performance, with a 600-image sample size (500 training and 100 validation).

VGG16 Transfer 2 (CNN): Like the previous transfer learning experiment, but with reduced training image size to study and compare the outcome. 200 image sample size (100 training and 100 validation). Tested the Model on 38 independent images unseen during training and validation, and the test accuracy is in Table 1.

<u>Observations from VGG16 model experiments:</u> All VGG16 models achieved 99% training accuracy and 85% validation accuracy (after 25 epochs run); However, overfitting continues even for VGG16. Transfer Learning with reduced training samples showed reduced validation loss (Training and validation accuracy in Table 4)

4.1.3 Experiment (1c) EfficientNet Model:

There were two architectures tested: Sequential and Functional architecture, and different layers for

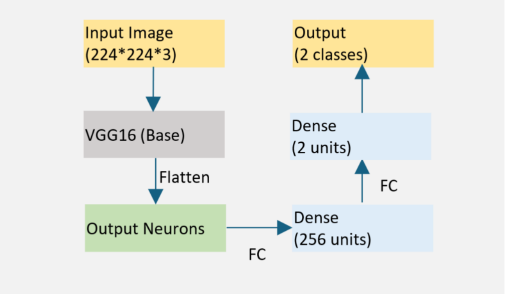


Fig.8: *VGG16 architecture (Transfer Learning).*

three different experiments for training and validation to choose the best EfficientNet model (Details in Table 5)

EfficientNet (sequential): The experiment with the sequential architecture is using a simple linear stack of layers and a 600-image sample size (500 training and 100 validation)

EfficientNet_1 (Functional): The Functional architecture used in this model is more complex, allowing more flexibility to the model. And a 600-image sample size (500 training and 100 validation)

EfficientNet_2 (Functional): Increased the FC layer in the functional architecture to find the difference in the outcome and 600 image sample size (500 training and 100 validation). Tested the Model on 38 independent images unseen during training and validation, and the test accuracy is in Table 1.

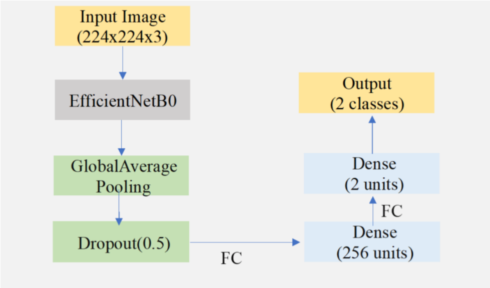


Fig.9: *EfficientNet_2 Functional architecture.*

Observations from EfficientNet model experiments: The EfficientNet model with Sequential architecture removed the overfitting issue with an accuracy at 50%; However, EfficientNet model with functional architecture and additional FC layer achieved 98% training and 99% Validation accuracy and this Model shows underfitting (Training accuracy < Validation accuracy) (Training and validation accuracy in Table 5)

Result of experiment (1): The EfficientNet model (Functional) gave the best results. Hence, chose this model to upskill by image background learning in further experiments.

4.2 Experiment (2): Training the best-fit Model to learn Backgrounds and Foregrounds to enhance the accuracy

For image background learning, the technique used was to classify the images into 14 subcategories, based on the location/surroundings associated with both begging and non-begging child activities in public, and then train the model to learn features from those classifications. This method will provide more background learning of the two classes of children (begging and normal), where they are often observed. There are areas, such as traffic and markets, where both categories of children are seen, and some surroundings where begging activities are less prevalent and the probability of finding a begging child is very low, for example, playgrounds and school areas.

(i) Begging: Market Areas, Public places, Traffic areas, Street, and over bridges.

(ii) Normal: Market Areas, Public places, Traffic areas, Street, over bridge, ghettos, House, School, and Kids' play areas.

The above 14 subcategories were used to train the model through two distinct approaches: "Learn with Subcategories" and "Learn with Subcategories and Sub-labelling" (Fig. 10 and Fig. 12, shown below). The model EfficientNet-Functional architecture, the best from Experiment (1), was used here for the image background learning method.

4.2.1 Experiment (2a) Learn with Sub-categories:

The images are categorized into 14 subcategories based on the various backgrounds within Begging and Normal circumstances (Fig. 10,11), with a sample size of 600 images (500 for training and 100 for validation). Tested the model on 38 independent images unseen during training and validation, the test accuracy is in Table 1.

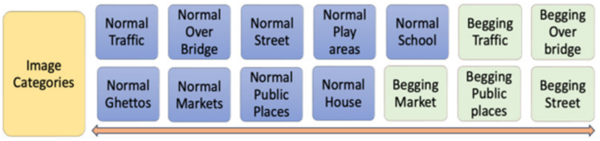


Fig.10: *The image sub-categories structure as per the background.*

Observations from subcategories model experiment: The model achieved 85% training and 97% validation accuracy (100 epochs), and underfitting is seen in this model (Training accuracy < Validation accuracy). (Training and validation accuracy provided in Table 6).

4.2.2 Experiment (2b)_Learn with Subcategories and Sub-labelling:

In the second method, an extra layer for learning the image background is added by sub-labelling

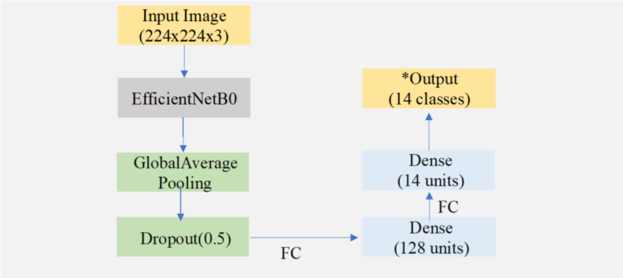


Fig.11: *EfficientNet Learn with sub-categories architecture.*
**The highest probability amongst sub-categories is used to determine begging or normal child*

the images into "Begging" and "Normal", and further sub-categorizing them according to the 14 backgrounds (Fig. 12,13), with a sample size of 600 images (500 for training and 100 for validation). Tested the model on 38 independent images unseen during training and validation, the test accuracy is in Table 1.

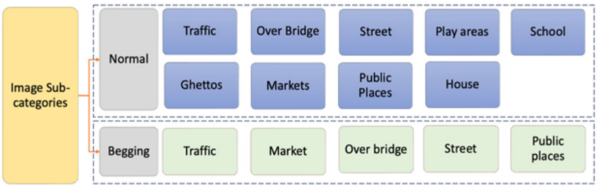


Fig.12: *The Image subcategories sub-labelling background structure.*

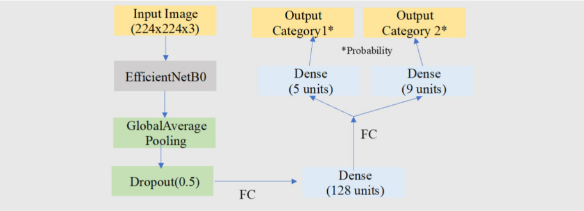


Fig.13: *EfficientNet Learn with sub-categories architecture.*
**Category 1 and Category 2 refer to two subcategories (Begging or Normal)*

Observations from subcategories and sub-labelling model experiment: The models achieved 98% training accuracy, and the validation accuracy touched 99% (after 100 epochs). There was less underfitting observed. (Training and validation accuracy provided in Table 7)

Result of Experiment (2): The model architecture utilized image background learning through subcategories and sub-labelling, resulting in improved overall accuracy.

4.3 Experiment (3): Training best fit Model to separately learn Backgrounds and Foregrounds, and integrating both model probabilities to enhance the prediction by adding background context awareness (Integrated Dual Image Classifier)

The model EfficientNet, the best model from experiment (1), was used here for the image background and foreground learning. For training backgrounds and Foregrounds separately, two sets of training images were used (foreground alone images and multiple background images) (Fig. 14 and Fig.15 below).

(i) Foreground alone images of 2 classes of children.

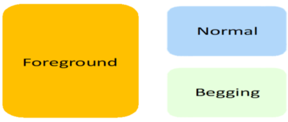


Fig.14: *The foreground image classes for the foreground model.*

(ii) Background images of eight different locations/surroundings where the child was found begging and not begging. [For example, there are locations where begging and normal children can be found (example: traffic areas), similarly, there are places where begging children are rarely seen (example: school area)]. Here, the training images were a mix of mostly the entire image and a few foreground-separated background images.

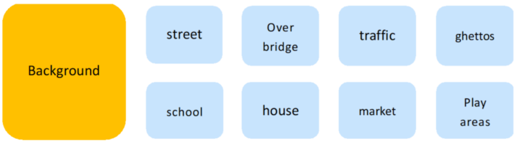


Fig.15: *The background image classes for the background model.*

4.3.1 Experiment (3a)_Training Background model and Foreground model:

The first part of the "Integrated Dual Image Classifier" involves training models for foreground and background only. The training dataset was divided into background and foreground sets to train the models independently (Fig. 14,15 above and architecture in Fig. 16,17 below).

Observations from Background model experiment: The models achieved 85% training accuracy and 90% of validation accuracy (after 100 epochs), and underfitting is seen in this Model (Training accuracy < Validation accuracy). (Training and validation accuracy provided in Table 8)

Observations from the Foreground model experiment: The models achieved 80% training accuracy and 85% validation accuracy (after 100 epochs), and

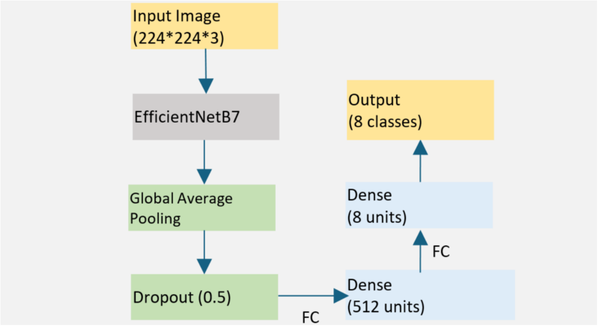


Fig.16: *EfficientNet Background model architecture.*

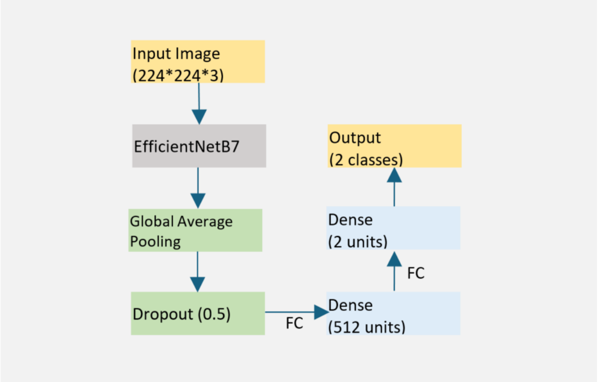


Fig.17: *EfficientNet Foreground model architecture.*

slight overfitting is seen in this Model (Training accuracy > Validation accuracy). (Training and validation accuracy provided in Table 8)

4.3.2 Experiment (3b)_Combining Background and Foreground Models Training and Probability Analysis for Background and Foreground (Integrated Dual Image Classifier)

The second part of the "Integrated Dual Image Classifier" is initiated after building separate models for the background and foreground. The background model prediction is leveraged for foreground prediction, ensuring a context-aware decision-making process.

Firstly, to obtain probabilities for background and foreground, the two pre-trained models from Experiment (3a), the background and foreground models, were used to generate predictions for the input image. Each Model outputs the probability scores for relevant classes (Example, the probability of "Begging", "Normal", "Street", "House", and so on) for the input image. Classes shown in Fig. 16 and Fig. 17.

Followed by calculation of background judgment and foreground judgments: The classification process first involves background judgment, followed by foreground analysis conditioned on the background result. Based on the foreground prediction probability above a threshold, images are categorized as either "begging" or "normal" (Fig. 18).

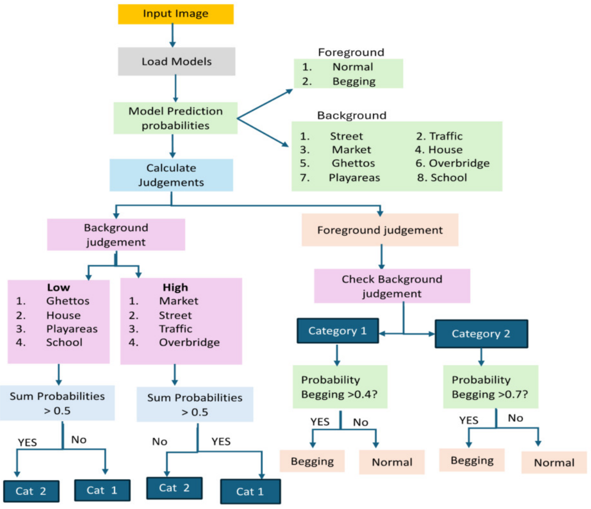


Fig.18: *Integrated Dual Image Classifier (combining background and foreground models).*

Calculation background and foreground judgments:

(i) The background judgment process categorizes an image based on the probabilities assigned to various background elements. The categories are divided into 'high' and 'low' based on their significance ('High' are those where both classes of children are commonly seen, and 'low' are those where the begging children are less probable to be found). 'High' background categories include 'Market', 'Street', 'Traffic', and 'Over Bridge', while 'low' categories include 'Ghettos', 'House', 'KidsPlayArea', and 'School'. The function calculates the sum of probabilities for each set of categories. If the sum for the 'high' categories exceeds 0.5, the background is judged as 'Cat 1'; otherwise, it is judged as 'Cat 2'. This classification helps in understanding the environment where the activity is taking place, providing context for further analysis.

(ii) The foreground judgment process focuses on determining the nature of the features of the foreground in the image, specifically whether it is 'Begging' or 'Normal'. Foreground judgment is adjusted based on the background judgment. If the background is classified as 'Category 1', the foreground threshold for classifying the foreground as 'Begging' is set lower at 0.4, meaning the model is more sensitive to detecting 'Begging' in active environments. Conversely, if the background is 'Category 2', the foreground threshold is raised to 0.7, making the model more stringent in classifying 'Begging' in less active environments. This adjustment ensures that the foreground classification considers the context provided by the background.

To optimize classification accuracy, the model applies a 40% FG threshold in high-probability areas (e.g., traffic) to capture moderate begging indicators. In comparison, a stricter 70% threshold is used in low-probability regions (e.g., playgrounds) to avoid misclassifying normal activities as begging. These thresholds were established through empirical dataset analysis, ensuring alignment with real-world patterns, and refined through contextual analysis and experimental validation. Multiple threshold values were tested and fine-tuned during experiments to achieve the best balance between false positives and false negatives.

4.4 Comparison of performance output of various models (including existing and proposed)

In this study, we first trained existing image classification models (CNN, VGG16, EfficientNet) with enhanced data augmentations, hyperparameter tuning, and transfer learning methods to select the best-fit model for classifying the begging children. Further, the EfficientNet model was trained for background learning by categorizing and labelling the two classes (begging and normal) according to the surroundings. The final Experiment introduced the dual classifier model to learn the foreground and background separately and integrate both model predictions for final classification to identify a begging child. For a comparative study of these models, 38 distinct test images were used (test accuracy).

In the first set of experiments (Experiment 1), the EfficientNet model showed improved test accuracy. The test accuracies were 64%(CNN), 50% (VGG16) and 74%(EfficientNet). The findings revealed that similar foreground subjects, when placed in different backgrounds, led to misclassifications (for example, a child in shabby clothing within a ghetto environment was sometimes incorrectly classified as begging). Which might not be a real case. To resolve this, background learning was considered (Experiment 2).

EfficientNet trained for background learning by Subcategorizing and sublabelling images based on their background, showed higher test accuracies of 78% and 81%, respectively.

To further enhance background awareness, enabling better understanding of the environment improved foreground classification. The background model and foreground model were trained separately and then integrated; their performance was subsequently studied (Experiment 3). We referred to this as the Integrated Dual Image Classifier. Although the overall accuracy remained at a similar level, the false negatives (failing to recognize the begging child) were drastically reduced. Further work is required to improve the efficiency of the foreground and Background model.

Table 1: *Performance comparison of models.*

Methods	Number of test datasets	Test Accuracy%
CNN	38	64%
VGG16 (Transfer Learning)	38	50%
EfficientNet	38	74%
EfficientNet Learn with subcategories	38	78%
EfficientNet Learn with Subcategories and Sub-labelling	38	81%
Integrated Dual Image Classifier	38	74%

Note: For the results in Table I. Independent 38 images were separately used for performance comparison

Table 2 compares performance of all models across 3 sample images shown in Fig. 19.

5. CONCLUSION

The proposed 'Integrated Dual Image Classifier' method effectively reduces the false negatives in the detection of begging children due to a context-aware decision-making process and is an improvement over singular classification. While foreground-background separation enhances detection, further improvements are needed in cases where foreground and background features are highly intertwined (e.g., children sitting on the ground in a way that blends with their surroundings). Additionally, while optimizing efficiency, the model needs to be improved for real-time deployment, reducing the need for high processing power and minimizing computational costs.

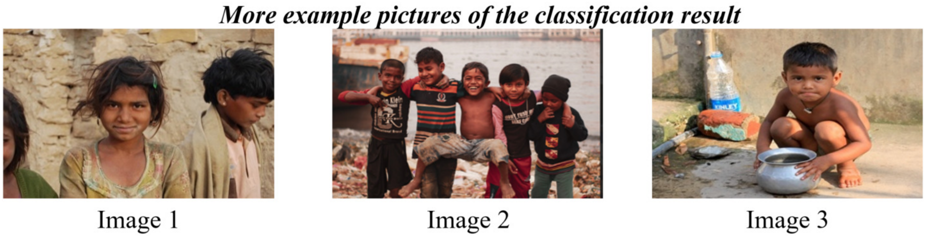


Fig.19: *Images (Image 1 to 3) used for testing for the results in Table 2 below. Image Source: Licensed under the Pixabay and Unsplash License, which permits free use and non-commercial purposes.*

Table 2: *Comparison of the test results of the three images shown in Fig. 19 above.*

Model	**Input Image**	**Predicted Label**	**Ground Truth**	**Correct/Incorrect**
CNN	Image 1	Normal Child	Normal Child	Correct
	Image 2	Begging Child	Begging Child	Correct
	Image 3	Begging Child	Normal Child	Incorrect (False positive)
VGG16	Image 1	Begging Child	Normal Child	Incorrect (False positive)
	Image 2	Normal Child	Begging Child	Incorrect (False negative)
	Image 3	Begging Child	Normal Child	Incorrect (False positive)
EfficientNet	Image 1	Normal Child	Normal Child	Correct
	Image 2	Normal Child	Begging Child	Incorrect (False negative)
	Image 3	Begging Child	Normal Child	Incorrect (False positive)
EfficientNet Learn with SubCategories	Image 1	Normal Child	Normal Child	Correct
	Image 2	Normal Child	Begging Child	Incorrect (False negative)
	Image 3	Begging Child	Normal Child	Incorrect (False positive)
EfficientNet Learn with Subcategories and Sub-labelling	Image 1	Normal Child	Normal Child	Correct
	Image 2	Normal Child	Begging Child	Incorrect (False negative)
	Image 3	Normal Child	Normal Child	Correct
EfficientNet Integrated Dual Image Classifier	Image 1	Normal Child	Normal Child	Correct
	Image 2	Begging Child	Begging Child	Correct
	Image 3	Normal Child	Normal Child	Correct

Note: Images 1 and 3 depict normal children; however, they are particularly tricky as the presence of a begging feature could lead to being mistaken for a begging child. Image 2 (begging) presents a significant challenge due to its combination of shabby clothing and happy expressions in a public setting, as it blends both poverty and non-begging cues.

- *False Positive (FP): Model incorrectly classifies a normal child as a begging child*
- *False Negative (FN): Model incorrectly classifies a begging child as a normal child*

Table 3: *CNN Experiments and Their Performances.*

S.No	Number of Conv Layers	Number of Pooling Layers	Number of FC Layers	Number of BN Layers	Training Accuracy	Validation Accuracy
CNN_1	3	3	1	0	99%	70%
CNN_2	4	4	2	4	99%	70%
CNN_3	8	4	3	4	99%	77%

Table 4: *VGG16 Experiments and Their Performances.*

S.No	Number of Conv Layers	Number of FC Layers	Training Accuracy	Validation Accuracy
VGG16 Base	13	3	99%	86%
VGG16_Transfer1_(CNN)	13	2	99%	85%
VGG16_Transfer2_(CNN)	13	2	99%	85%

Table 5: *EfficientNet experiments and their performances.*

S.No	Number of Conv Layers	Number of FC Layers	Extra Number of FC Layers	Training Accuracy	Validation Accuracy
EfficientNet (sequential)	20	1	1	50%	50%
EfficientNet_1 (Functional)	20	1	0	96%	90%
EfficientNet_2 (Functional)	20	1	1	98%	99%

Table 6: *Background Learning EfficientNet with Subcategories experiments and their performances.*

S.No	Number of Conv Layers	Number of FC Layers	Extra Number of FC Layers	Number of epochs	Training Accuracy	Validation Accuracy
EfficientNet Learn with Subcategories	20	1	1	100	85%	97%

Table 7: *Background Learning EfficientNet with Subcategories and Sub-labelling experiments, and their performances.*

S.No	Number of Conv Layers	Number of FC Layers	Extra Number of FC Layers	Number of epochs	Training Accuracy	Validation Accuracy
EfficientNet Learn with Subcategories and Sub-labelling	20	1	1	200	98%	99%

Table 8: *Separate learning of Backgrounds and Foregrounds experiments and their performances.*

S.No	Number of Conv Layers	Number of FC Layers	Extra Number of FC Layers	Number of epochs	Training Accuracy	Validation Accuracy
EfficientNet Background Model	218	2	1	200	85%	90%
EfficientNet Foreground Model	218	2	1	200	80%	85%

6. DATA USAGE DISCLAIMER

This study utilizes publicly available thumbnails of children to extract features for model development, sourced from the internet solely for academic research purposes, specifically for training and evaluating the proposed machine learning model. We acknowledge the copyright limitations and associated risks. The images are not redistributed or used for commercial purposes and are not intended for misuse, exploitation, or the identification of individuals. The use of such data is solely aimed at addressing socially relevant challenges in a research context.

7. IMPLICATIONS & FUTURE WORK

A neural network model with image background learning demonstrates broader potential. It demonstrates that a robust AI-assisted system can be developed for identifying children involved in begging, supporting social intervention efforts such as rehabilitation processes. These AI-assisted systems (AI Cameras) can be utilized in locations where begging activities are more common, such as traffic signals, crowded tourist spots, and busy markets. The data collected from these AI systems can be utilized by governments, NGOs, and other organizations for informed actions and policies aimed at reducing child begging. In the Future, the following can be explored as an extension of this study:

(i) Improving current models with improved learning on image foreground detection by adding other feature mappings like posture trained using MediaPipe, which will be an added advantage to the current Model.

(ii) Semantic foreground detection by utilizing background model predictions, aiming to improve the accuracy of detection through a combined score of background and foreground analysis.

AUTHOR CONTRIBUTIONS

Midhu Jean Joseph: Conceptualization, Methodology, validation, investigation, data curation, writing-original draft preparation, visualization.

Virach Sornlertlamvanich: Conceptualization, methodology, formal analysis, validation, supervision, writing-review.

Thatsanee Charoenporn: Methodology, validation, investigation, formal analysis, writing-review and editing.

References

[1] Government of India, Press Information Bureau, Delhi, “Child Begging,” [Press release], Dec. 7, 2021. [Online]. Available: `https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1778853`.

[2] M. A. Al Helal and K. S. Kabir, “Exploring Cruel Business of Begging: The Case of Bangladesh,” *Asian Journal of Business and Economics*, vol. 3, no. 3.1, 2013.

[3] World Health Organization, “Child maltreatment,” Sep. 9, 2022. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/child-maltreatment`.

[4] M. V and K.T. Geetha, “Socio-Economic Causes of Begging,” *International Research Journal of Human Resources and Social Sciences*, vol. 3, no. 2, pp. 243-258, Aug. 2014.

[5] F. T. Progga, M. T. Shahria, A. Arisha and M. U. A. Shanto, “A Deep Learning Based Approach to Child Labour Detection,” *2020 6th Information Technology International Seminar (ITIS)*, Surabaya, Indonesia, pp. 24-29, 2020.

[6] N. Paredes, E. Caicedo-Bravo and B. Bacca, “Emotion Recognition in Individuals with Down

Syndrome: A Convolutional Neural Network-Based Algorithm Proposal," *Symmetry*, vol. 15, no. 7, pp. 1435, 2023.

[7] B. Oluwalade, S. Neela, J. Wawira, T. Adejumo and S. Purkayastha, "Human Activity Recognition using Deep Learning Models on Smartphones and Smartwatches Sensor Data," in *Proc. 14th Int. Conf. Health Informatics (HEALTHINF)*, 2021. [Online]. Available: `https://arxiv.org/pdf/2103.03836`.

[8] D. Mekala, K. Anand, A. Ghosh and S. Kataria, "Foreground-background classification and ROI detection in surveillance videos," 2016.

[9] K. Y. Xiao, L. Engstrom, A. Ilyas and A. Madry, "Noise or Signal: The Role of Image Backgrounds in Object Recognition," *ICLR 2021*, 2020.

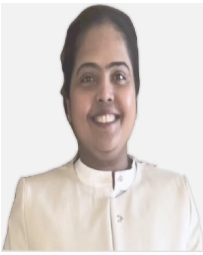
[10] M. Mayank, "Convolutional Neural Networks, Explained," *Towards Data Science*, Aug. 2020.

[11] Y. Yu and W. Liu, "Ethical Issues of Child Abuse-A Cultural Comparison," *2020 International Conference on Public Health and Data Science (ICPHDS)*, Guangzhou, China, pp. 52-55, 2020.

[12] S. M. Ayoob, "Beggary in the Society: A Sociological Study in the Selected Villages in Sri Lanka," *Journal of Xi'an University of Architecture & Technology*, vol. 11, no. 12, pp. 1725, 2019.

[13] M. Moayeri, P. Pope, Y. Balaji and S. Feizi, "A Comprehensive Study of Image Classification Model Sensitivity to Foregrounds, Backgrounds, and Visual Attributes," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 19065-19075, 2022.

[14] O. Pronina and O. Piatykop, "The recognition of speech defects using a convolutional neural network," *CoSinE 2022*, CITEd Kyiv, Ukraine, Dec. 2022.



Midhu Jean Joseph is a doctoral researcher in Data Science at Musashino University, Tokyo, Japan, specializing in context-aware neural networks. She holds a Master's degree in Data Science from the same university, with a focus on Human Behaviour Technology and Social Innovation (2023–2025). Her academic journey is complemented by an Executive Program in Applied Financial Risk Management from the Indian Institute of Management (IIM) Kashipur, India (Batch of 2017), and a Bachelor of Science in Chemistry from St. Teresa's College, MG University, Kerala, India (Class of 2006).

With over 16 years of professional experience in the investment banking sector, Midhu has worked extensively in Market Risk, Market Risk Reporting, Regulatory Compliance, and Project Management. Her work has spanned global markets and legal entity-level risk management across Asia and Europe. She has led initiatives in risk engine enhancements, report automation, and regulatory limit monitoring, and has actively contributed to senior management reporting and quarterly risk reviews.

Her interdisciplinary research interests lie at the intersection of artificial intelligence and societal impact, with a particular emphasis on AI-based human behavior recognition, context-aware neural networks, social innovation, and risk analytics. She is also deeply interested in the ethical applications of AI, especially for vulnerable populations.

Virach Sornlertlamvanich was esteemed "The Researcher of the Year 2001" by the Nation Newspaper (Thailand). Observing his continuous contributions in the field of Computer Engineering, he was conclusively awarded "National Distinguished Researcher Award 2003" in Information Technology and Communication by the National Research Council of Thailand, "ASEAN Outstanding Engineering Achievement Award 2011" by ASEAN Federation of Engineering Organizations (AFEO), and followed by "Outstanding Alumni Award, Tokyo Tech Alumni Association (Thailand Chapter)" in 2021.

He received a doctoral degree in Computer Engineering from the Tokyo Institute of Technology in 1998. He worked with NEC Corporation as a sub-project leader for Thai language processing in the Multi-lingual Machine Translation Project. He joined the National Electronics and Computer Technology Center (NECTEC) in 1992. His research interests are in the area of Natural Language Processing, Machine Translation, Information Retrieval, Knowledge Engineering, and Artificial Intelligence. His recent efforts are on the research and development of technology for Digitized Thailand 2009 which is aimed to establish a service platform for digital content and applications to accomplish the creative industry. He is also a pioneer in establishing an AI research platform for Thammasat AI City 2020 at Rangsit Campus during 2020-2023.

Thatsanee Charoenporn s a Professor in the Faculty of Data Science at the Asia AI Institute, Musashino University. She holds a B.A. and M.A. in Linguistics and a Ph.D. in Information Technology. She has previously held research and project leadership roles at the National Institute of Information and Communications Technology (Japan), the National Electronics and Computer Technology Center (Thailand), and Burapha University (Thailand). Prof. Charoenporn has extensive experience in multidisciplinary research and projects. Her current research focuses on language resources, cross-cultural communication, smart education, healthcare technology, and social innovation.