UNIVERSIDADE D
COIMBRA

Gabriela de Melo Simões

RGB-Based Automatic Recognition of the Engagement of Children Interacting with a NAO Humanoid Robot

Dissertation supervised by Professor Doctor Urbano José Carreira Nunes and by Professor Doctor Ana Cristina Barata Pires Lopes and submitted to the Electrical and Computer Engineering Department of the Faculty of Science and Technology of the University of Coimbra, in partial fulfillment of the requirements for the Master Degree in Electrical and Computer Engineering, specialization in Robotics, Control and Artificial Intelligence with subareas: Computational Learning, Robotics and Autonomous Systems.

February of 2023

UNIVERSIDADE D
COIMBRA

Faculty of Sciences and Technology
Department of Electrical and Computer Engineering

RGB-Based Automatic Recognition of the Engagement of Children Interacting with a NAO Humanoid Robot

Gabriela de Melo Simões

Master Degree in Electrical and Computer Engineering

Coimbra, February of 2023

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE Đ
COIMBRA

RGB-Based Automatic Recognition of the Engagement of Children Interacting with a NAO Humanoid Robot

Supervisors:
Prof. Dr. Urbano José Carreira Nunes

Co-Supervisor:
Prof. Dra. Ana Cristina Barata Pires Lopes

Jury:
Prof. Dr. Paulo Jorge Carvalho Menezes
Prof. Dr. Nuno Miguel Mendonça da Silva Gonçalves
Prof. Dr. Urbano José Carreira Nunes

Master Degree in Electrical and Computer Engineering

Coimbra, February of 2023

Agradecimentos

A concretização desta dissertação marca o fim de um ciclo de cinco anos e meio onde fui posta à prova diversas vezes. Muitas dessas vezes duvidei se estaria à altura do desafio, mas nessas alturas tinha na minha vida uma série de pessoas especiais que sempre me ajudaram e fizeram acreditar que seria capaz, às quais expresso o meu profundo agradecimento.

Gostaria de começar por agradecer aos meus orientadores, Prof. Dr. Urbano Nunes e Prof. Dra. Ana Lopes, pela vossa dedicação e apoio incansável durante todo o processo de pesquisa, desenvolvimento e escrita desta dissertação. Estou profundamente grata pela vossa partilha de sabedoria, paciência, incentivo e pelas excelentes condições de trabalho que me proporcionaram. Gostaria de agradecer também ao Prof. Dr. Carlos Carona que partilhou o seu conhecimento, experiência e me guiou durante a realização deste trabalho.

Ao projeto RoboTherapist - "Child-Robot Interaction (cHRI) in Pediatric Settings", com financiamento suportado pela FCT/MCTES, financiado inteiramente por fundos nacionais através do UIDB/0048/2020.

Ao Instituto de Sistemas e Robótica (ISR) que me acolheu e aos meus colegas de laboratório Ricardo Pereira, João Paulo e Luís Garrote, pela partilha de conhecimento, troca de ideias e contributos para o meu trabalho.

Aos meus colegas de curso, a minha gratidão pela vossa amizade, apoio e por todo o trabalho em equipa ao longo destes desafiantes anos de estudo, sem a vossa ajuda teria sido sem dúvida muito mais difícil.

Ao meu namorado, Diogo, por me ter acompanhado, motivado, encorajado e ter sido o meu porto de abrigo ao longo desta jornada. À sua família que me acolhe, acarinha e incentiva sempre.

À minha equipa, Secção de Badminton da Associação Académica de Coimbra, que me viu crescer. Durante todo este processo ajudou-me a aprender a gerir o meu tempo, a carregar baterias, a descontrair e a desenvolver outras competências sociais e de trabalho em equipa.

Ao meu amigo de quatro patas e bigode Snow, por ser uma fonte constante de amor e conforto.

O maior dos agradecimentos vai para a minha família em especial para os meus pais

Paula e Pedro, avós e padrinhos pelo amor incondicional, carinho e motivação. Um obrigado aos meus pais que me concederam as melhores condições para eu poder crescer e estudar. Por terem estado sempre presentes na minha vida e incentivado a seguir os meus sonhos e a traçar o meu próprio caminho. Às minhas avós Lena e Tila, que me criaram e cuidaram de mim com o amor que só uma avó consegue dar. Por essas razões a eles dedico este trabalho.

Abstract

The role of social robots has enhanced relevance in modern societies. The interaction with humanoid robots proved to be extremely helpful and empowered on multiple levels while interacting with children, with a therapeutic function to improve social, educational, and learning skills by mimicking human behaviour and encouraging activities. The robots can also be a useful tool in the mitigation of negative states and behaviours during therapeutic sessions with specialists, such as psychologists. In this sense, engagement recognition of the child while performing a task with the robot is fundamental to evaluate whether the child is interested, and maintains interest, in the child-robot activity.

This work focuses on the development of an engagement classification system based on emotion recognition through facial expression, which hopefully will be useful to online and offline analysis of therapy sessions. For this purpose, the emotion classification was developed based on CNN architectures fed by children's image faces provided from different datasets (EmoReact, FER2013, and CAFE datasets), and two methods of engagement classification were explored. Method-1 relies on emotion classification to infer the levels of engagement based on the direct relationship between levels of engagement and emotions, whereas Method-2 employs the same network of Method-1 to directly classify four classes of engagement. The affective model proposed by [1] is used to relate emotions and engagement. Method-1 performs better than Method-2, achieving a accuracy of 88.14%.

A child-robot activity, based on the NAO humanoid robot, was designed with the assistance of Professor Carlos Carona, professor at the Faculty of Psychology of University of Coimbra, with the initial phase focusing on engaging the child with dances and motions, followed by a breathing-based relaxation exercise.

Preliminary results with three children performing the child-robot activity, and the offline use of the engagement classification system, showed promising results and can be further explored and implemented in psychology areas in the future.

Keywords: Engagement Classification, Emotion Recognition, Convolutional Neural Networks, NAO Robot

Resumo

O papel dos robôs sociais tem aumentado a sua relevância na sociedade moderna. A interação com os robôs humanóides provou ser extremamente útil e capacitada a vários níveis enquanto interagem com crianças, com uma função terapêutica para melhorar as competências sociais, educacionais e de aprendizagem, imitando o comportamento humano e encorajando atividades. Os robôs podem também ser uma ferramenta útil na mitigação de estados e comportamentos negativos durante sessões terapêuticas com especialistas, tais como psicólogos. Neste sentido, o reconhecimento do envolvimento da criança enquanto executa uma tarefa com o robô é fundamental para avaliar se a criança está interessada, e mantém o interesse, na actividade do robô criança.

Este trabalho concentra-se no desenvolvimento de um sistema de classificação do envolvimento baseado no reconhecimento de emoções através da expressão facial, que esperamos que seja útil para a análise online e offline de sessões terapêuticas. Com este porpósito, a classificação das emoções foi desenvolvida com base em arquitecturas CNN alimentadas por imagens faciais de crianças provenientes de diferentes conjuntos de dados (EmoReact, FER2013 e CAFE datasets) sendo explorados dois métodos de classificação do envolvimento. O Método-1 depende da classificação das emoções para inferir diretamente os níveis de envolvimento com base na relação direta entre os níveis de envolvimento e as emoções, enquanto que o Método-2 aplica a mesma rede do Método-1 para classificar directamente quatro classes de envolvimento. O modelo afectivo proposto por [1] é utilizado para relacionar as emoções e o envolvimento. O Método-1 funciona melhor que o Método-2, alcançando uma boa precisão de 88.14%.

Uma atividade criança-robô, baseada no robô humanóide NAO, foi concebida com a assistência do Professor Carlos Carona, Professor na Faculdade de Psicologia da Universidade de Coimbra, com a fase inicial centrada no envolvimento da criança com danças e movimentos, seguida de um exercício de relaxamento baseado na respiração.

Resultados preliminares com três crianças da realização da atividade criança-robô e da utilização offline do sistema, mostrou resultados promissores que poderão vir a ser explorados e implementados em áreas da psicologia no futuro.

Keywords: Classificação do Envolvimento, Reconhecimento de Emoções, Redes Neurais Convolucionais, Robô NAO

Contents

List of Acronyms

AF	Activation Function
ANN	Artificial Neural Networks
CAFE	The Child Affective Facial Expression Dataset
CNN	Convolutional Neural Network
CUDA	Compute Unified Device Architecture
DA	Data Augmentation
DL	Deep Learning
FACS	The Facial Affective Coding System
FER2013	Facial Emotion Recognition 2013 Dataset
FPS	Frames Per Second
FC	Fully Connected
GPU	Graphics Processing Unit
GELU	Gaussian Error Linear Units
ML	Machine Learning
MTCNN	Multi-Task Cascaded Convolutional Neural Networks
ReLU	Rectified Linear Units
TL	Transfer Learning

List of Figures

List of Tables

1 Introduction

This chapter describes the context and motivation, goals, proposed general framework and main contributions of the work described in this dissertation, focusing on child-robot interaction and how to recognise whether the child is engaged in the task/activity with a robot.

1.1 Context and motivation

In recent times the role of social robots has enhanced relevance in society. They proved to be extremely helpful and empowered on multiple levels while interacting with children [9][10]. Frequently, they have a therapeutic function to improve social, educational, and learning skills by mimicking human behaviour and encouraging activities. The robots can be a useful tool for improving communication, providing positive stimulation through entertainment and therapy, and contributing to the mitigation of negative states and behaviours during the therapeutic sessions with specialists such as psychologists.

In a relationship between a child and, in this case, a robot, one of the main difficulties is to gain and keep the child's attention and focus in a certain task. Therefore, in therapeutic tasks based on child-robot interaction, it is important to find out if the child maintains interest in the robot's activity and cooperates with it.

The ability to recognise emotions leads to the development of social skills and effective communication, and is vital for perception and decision-making [11]. In the research work presented in this dissertation, recognising a person's emotion based on the facial expression is considered as an ideal starting point for inferring their level of engagement with the robot activity, since it is a reliable predictor of how a person (in this case a child) is feeling.

The research work described in this dissertation may also contribute as a supplementary assessment tool to be used technical specialists, such as psychologists, after therapeutic sessions, to determine whether the child was engaged and attentive during the therapy session and robot activity. In the future, it will be advantageous to capture the child's engagement in real time so that the robot may adjust its behaviour to the child's engagement level, and emotional state.

To reach the intended goals, that will culminate in the design of effective child-robot

applications, the work described in this dissertation faces multiple research challenges. Although many progresses have been made in the child-robot interaction field in recent years [12][13][14], there are still many open questions remaining, such as: i) During the design of effective robot therapy applications, which type of sensing allows effective perception of child engagement in the robot task? ii) How can effective Deep Learning (DL) methods for automatic engagement perception can deal with limited and unbalanced resources in terms of available data; iii) Which features should be considered to accurately perceive engagement?

It is intended that the work carried out within the scope of this dissertation will help to respond to some of these problems. A child-robot activity, using a humanoid robot NAO, is proposed with the primary goal of researching and using Machine Learning (ML) techniques (methods based on Convolutional Neural Networks (CNN)) for recognising the level of engagement, based on the recognition of emotions by using facial expressions of a child performing a task with the humanoid robot.

1.2 Goals and proposed general framework

The main purpose of this dissertation was to design and develop a classification engagement system to be employed as an auxiliary tool during a therapeutic child-robot activity.

The global framework presented in Fig.1.1 can be decomposed into four principal modules, which represent the four main objectives of this dissertation, in particular: (1) - Datasets - to research and establish representative sets of data to be used in training, test and validation of the proposed engagement classification system; (2) - Engagement classification system - to research and develop appropriate methodologies of engagement classification based on deep learning techniques, in particular CNNs; (3) - The child robot activity - to research and develop a child-robot activity that could be used in therapeutic settings; (4) - Recorded experimental tests - acquisition of video images during child-robot-activity experimental setting for further offline analysis and validation of the proposed engagement classification methodology.

An engagement classification system as the one depicted in Fig. 1.1 requires a large amount of data, so a representative dataset of emotions was selected and pre-processed to serve as the input dataset.

The engagement classification system was developed using CNN that were fed with images of children's facial expressions. Two strategies were explored using CNNs: In the first approach, illustrated in Fig.1.2, the CNN received seven classes of images of children's facial expressions as input and returned an emotion-based prediction from which four levels of engagement could be directly inferred based on the affective model [1]. On the second approach, illustrated in Fig.1.3, the CNN received the direct inference of the engagement

level based on the affective model [1] as an input, and returned the four engagement classes. In both approaches, a training and a testing phase were necessary. During the training phase, the feature extraction and classification module is trained on a labelled training dataset. During the testing phase, the classification accuracy of the model is determined by testing it on a labelled testing set. The training and testing are both subsets of the same dataset.

The child-robot activity, based on the NAO humanoid robot, was designed with the assistance of a psychology professor at the University of Coimbra, Carlos Carona, with the initial phase focusing on engaging the child with dances and motions, followed by a breathing-based relaxation exercise.

The child-robot activity proposal and the engagement classification system were validated by three children aged 5, 6, and 7 years. A camera on a mobile phone was used to record the interactions between the robot and the children, and this data was then pre-processed and fed into the engagement classification model.

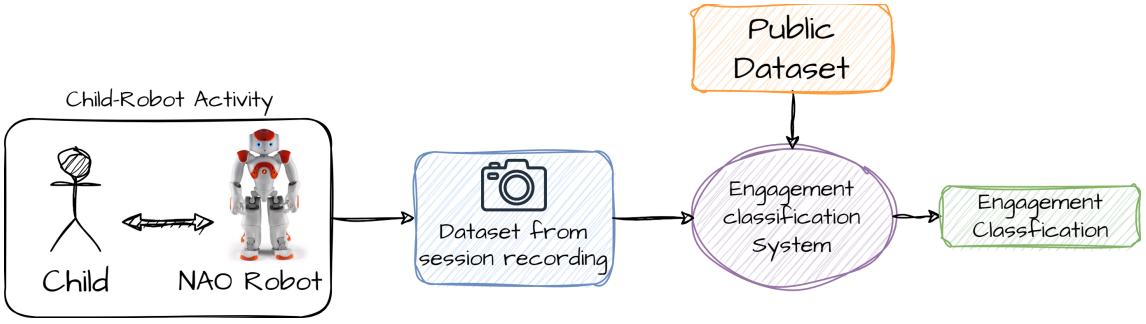

Figure 1.1: Proposed Engagement Classification Framework.

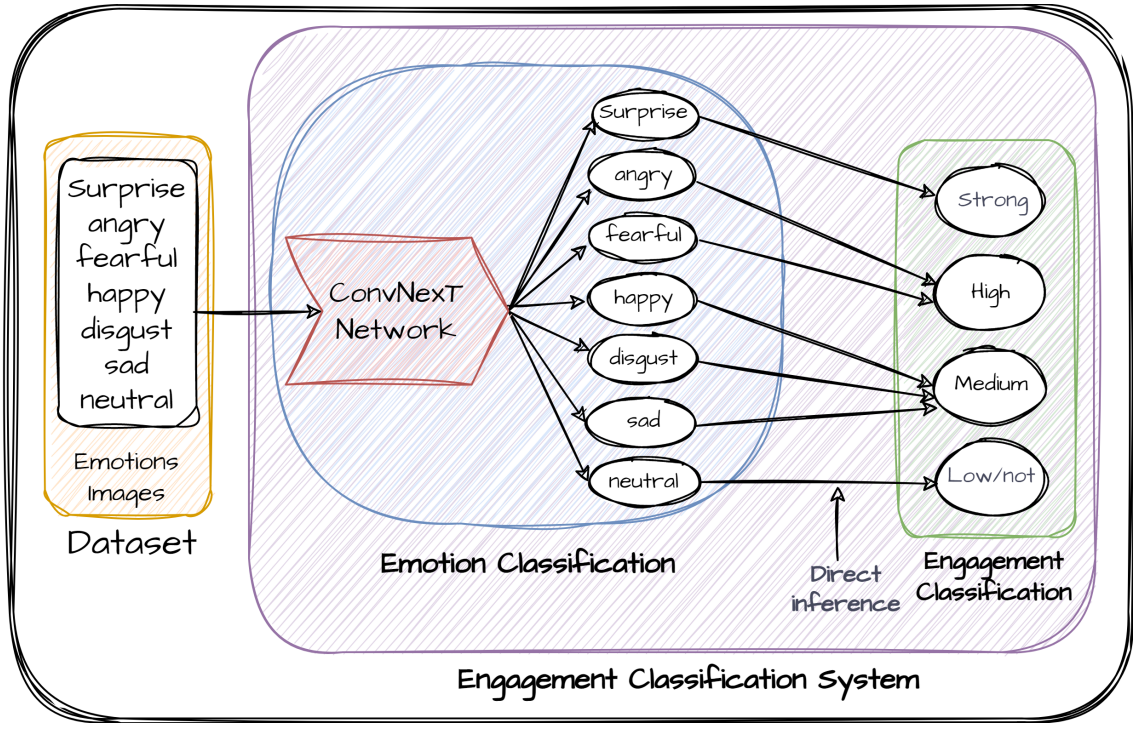

Figure 1.2: Engagement Classification System - Method-1.

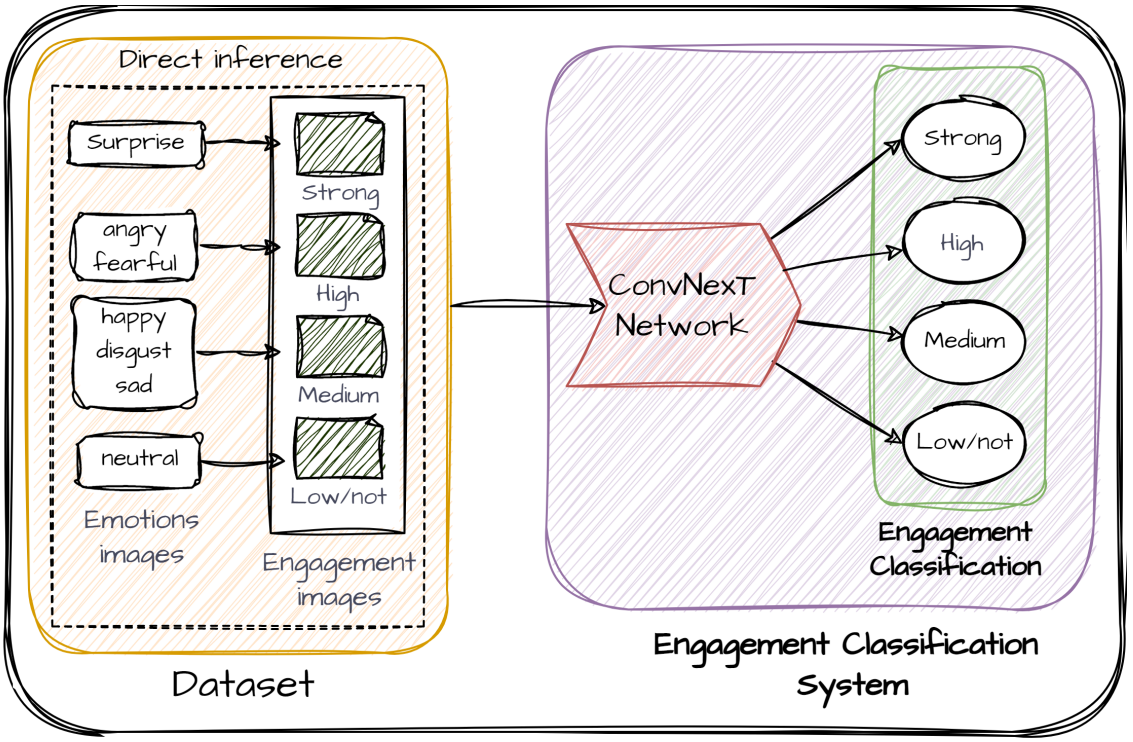

Figure 1.3: Engagement Classification System - Method-2.

1.3 Implementation and Key Contributions

To accomplish the proposed engagement classification framework, the following key contribution were attained:

1. Selection, analysis and pre-processing of three public datasets: EmoReact (children video dataset with 8 emotions); FER2013; and CAFE datasets;
2. Evaluation of different CNN architectures using the three datasets mentioned in 1), such as ResNet18, ResNet50, VGG16, ResNeXt50, and ConvNeXt, with variations on different parameters, aiming to choose the architecture with the best performance in classifying emotions;
3. Design and development of an engagement classification system employing the CNN with best performance in 2), and using two different approaches to infer engagement levels based on an affective model [7] that relates emotions with engagement;
4. Design of a child-robot activity with the guidance of a psychology therapist professor of the Psychology Faculty of the University of Coimbra;
5. Experimental results with three children that validate the effectiveness of the proposed child-robot activity;
6. Acquisition of child-robot activity data for posterior evaluation of the engagement classification system using data from the recorded videos.

1.4 Dissertation Organization

The dissertation is organised into six chapters as follows:

- **Background Material (Chapter 2):** Contains the theoretical concepts necessary to understand the proposed engagement classification system frameworks;
- **State-of-the-Art (Chapter 3):** Review of the most pertinent State-of-the-Art publications on detecting and classifying engagement and its use in robotherapy;
- **Materials and Methods (Chapter 4):** Describes the software and hardware used to achieve the objectives of the work described in this dissertation;
- **Developed Work (Chapter 5):** Describes dataset pre-processing as well as the research and development methodologies for each component of the proposed classification engagement system and child-robot activity;
- **Results and Discussion (Chapter 6):** Analysis of the experimental outcomes collected throughout the development of the engagement classification system;
- **Conclusion and Future Work (Chapter 7):** Draws conclusions and discusses possible future work.

2 Background Material

This chapter addresses fundamentals of DL, CNN, and other important theoretical topics applied in this dissertation.

2.1 Deep Learning

DL is a ML subset embedded in Artificial Intelligence (AI) that enables computers to do what comes naturally to people, namely, operate like a human brain and acquire specific knowledge. For that, they are composed of multilayered neural networks that learn from huge amounts of data. DL is a significant component of data science, which also covers statistics, predictive modelling, speech recognition [15], computer vision [16], natural language processing [17] and solve tasks like classifications [18] among others.

2.1.1 Convolutional Neural Networks

CNNs [2][19] are a type of neural network that are also based on the human brain. In particular, they are based on the visual cortex of a cat's brain, which is a complex set of cells that the CNN mimics. CNNs are mostly used for image classification and pattern identification within images, and their primary advantage is their capacity to identify relevant features without human interference.

Supervised learning was employed to train computer models for image processing and classification using a CNN, consisting on an approach that uses training datasets composed of labelled examples. Computing with appropriate label data target aims to teach the algorithm to classify unlabelled data based on its training with labelled data. The CNN model uses a three-dimensional input with height, width, and depth, and its primary layers are convolutional layers, pooling layers, and fully connected layers, which are described in more detail below.

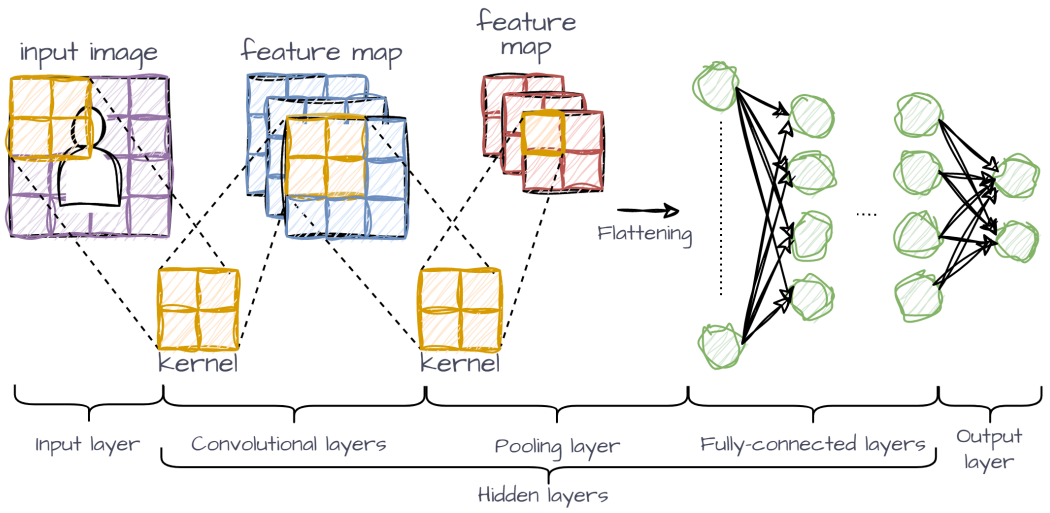

Figure 2.1: A simple CNN architecture (adapted from [2][3]).

Convolutional Layers

The initial layers of a CNN are convolutional layers [2][3] that work as feature extractors provided its ability to extract high-level features from input data. Each layer conducts a convolution between its input and a particular kernel. The convolution procedure is carried out by superimposing the kernel on the input image and calculating the dot product between the pixel values and the kernel's weights.

Nonetheless, before explaining what occurs in these layers, it is important to understand **kernel**, **stride**, and **padding**, among other crucial parameters. The kernel is characterised by a grid of discrete values designated kernel weights. At the start of the CNN training method, the kernel's weights are initialised at random and are adjusted during the training phase, being a learnable parameter. For each iteration of the kernel, the data is shifted by a certain number of rows or columns, called the stride. Zero-padding is the straightforward process of padding the input's border, and it is an efficient approach for providing further control over the output volumes' dimensions.

The convolutional layer performs a dot product between the input image and the kernel. The kernel has a smaller height and width than the height and width image, but the same depth. For instance, if the picture is in RGB format, the kernel will also have three channels.

During the forward pass, the kernel walks with a step equal to the stride size along the image's height and width. In addition, the dot product between the input image and the kernel is determined, where their respective values are multiplied and then summed

concurrently to provide a single scalar result. The entire procedure is then repeated until sliding is no longer feasible. The results of dot product represents the feature map.

If we have an input of size $[W \times W \times D]$ and a kernel with a spatial dimension of F, stride S, and padding P, then the following formula may be used to compute the output volume size (W_{out}):

$$W_{out} = \frac{W - F + 2P}{S} + 1 \tag{2.1}$$

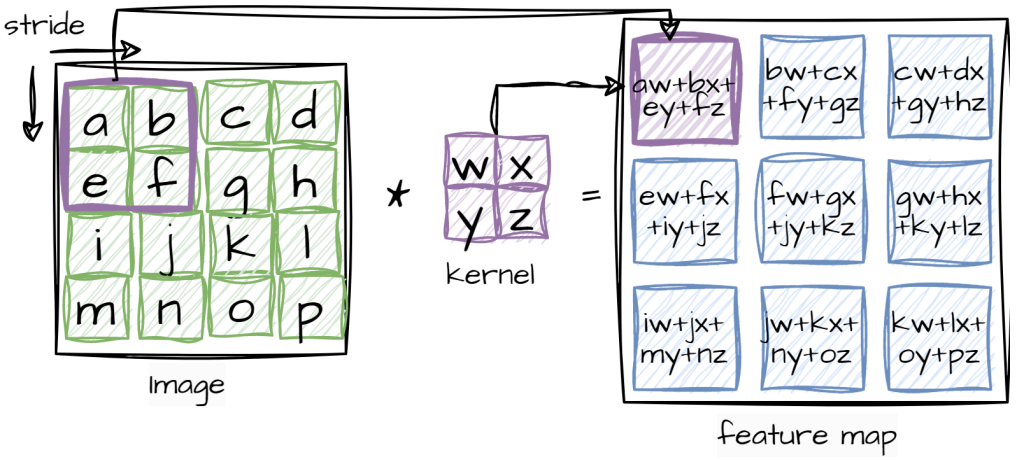

Figure 2.2: An example of convolutional operation with an input image $[4 \times 4 \times 1]$, a kernel $[2 \times 2 \times 1]$ and a stride of 1 in both axis (adapted from [4][3]).

Pooling Layers

The major purpose of pooling layers, which are applied after convolutional layers, is to subsample the feature maps. This technique lowers large-scale feature maps to smaller-scale feature maps. Similar to the convolutional technique, both the stride and the kernel are assigned sizes before to the execution of the pooling process, and the majority of the dominating information (or characteristics) are maintained at each iteration of the pooling process.

There are several accessible pooling algorithms; however, the most used are the Maximum [20][21], Average [21], Global Average[22] and Adaptive Average [23] Pooling methods, which select the highest value, calculate the average of the region selected by the kernel, or compute a weighted average based on the distance to the central element, respectively. Adaptive Average Pooling is just an average pooling operation that calculates the right kernel size based on the size of the input and the size of the output. With this operation, the search space should become much easier to use and more expressive.

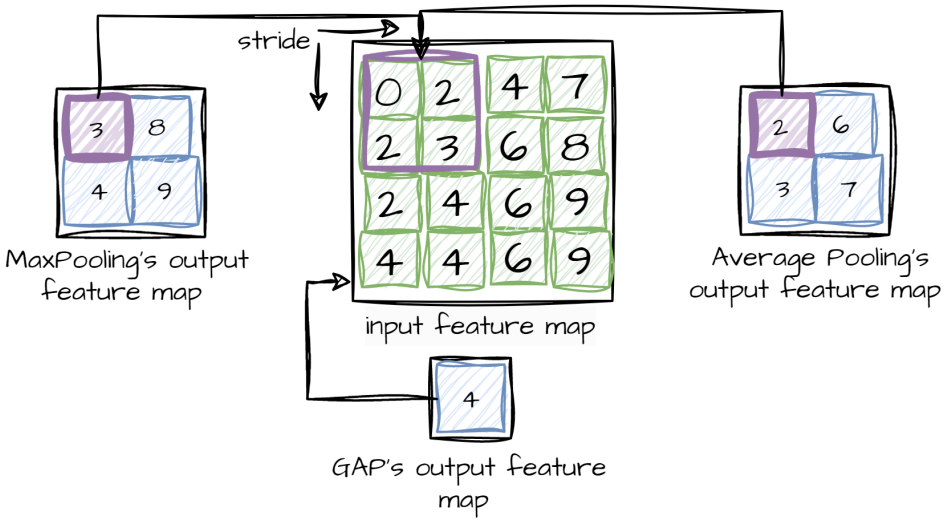

Figure 2.3: An example of a max, average, and global average pooling operation with an input feature map $[4 \times 4 \times 1]$, a kernel $[2 \times 2]$ and a stride of 2 in both axis (adapted from [4][5]).

Fully Connected Layers

At the end of each CNN architecture is a Fully Connected (FC) layer [24]. Each neuron in this layer is coupled to all neurons in the preceding layer and receives input from the final pooling or convolutional layer. This input is a vector formed from the feature maps after they have been flattened. CNN classification happens at this layer; hence, the output of the FC layer provides the final CNN output.

Activation Function (AF)

Non-linear layers are generally placed right after the convolutional layer and FC layers to bring non-linearity to the activation map and give them the capacity to learn extra-complicated things. The input value is computed by adding the weighted sum of the neuron's input and its bias. This implies that the activation function determines whether or not a neuron should activate in response to a certain input by generating the associated output. CNN and other deep neural networks most frequently employ the activation functions Rectified Linear Units (ReLU) [25], Sigmoid [26], Hyperbolic Tangent (Tanh) [26] and the more recently explored Gaussian Error Linear Units (GELU) [27], with corresponding graphics and equations shown in Fig. 2.4 and Fig. 2.5.

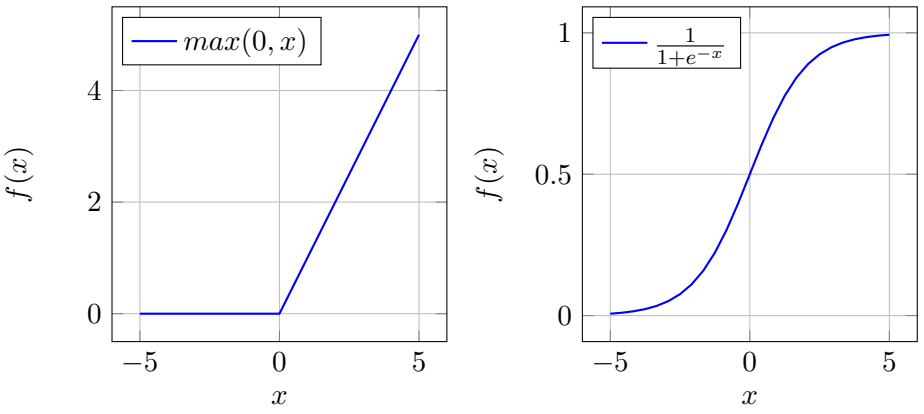

Figure 2.4: ReLU and Sigmoid Activation Functions, respectively.

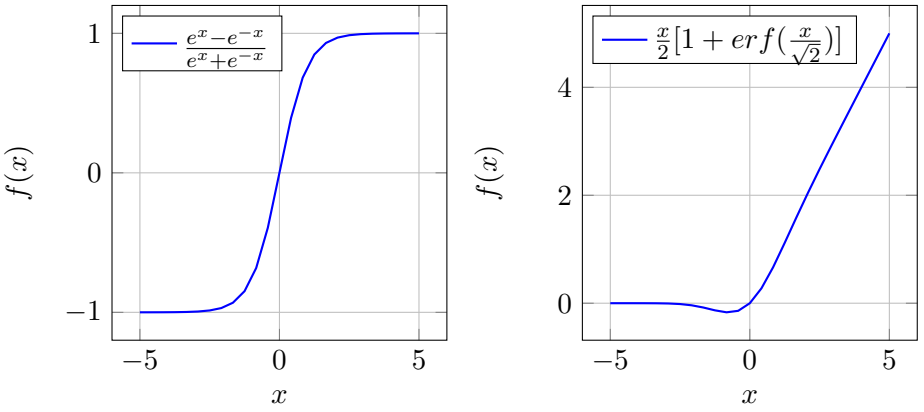

Figure 2.5: Tanh and GELU Activation Functions, respectively.

Loss Function

Some loss functions are applied in the output layer of the CNN model to compute the predicted error generated by the training data. This deviation reflects the disparity between the label and the predicted output, or estimation.

The Cross-Entropy Loss Function [28] mathematically described in 2.2, one of the more commonly used loss functions, is used to evaluate the CNN model's performance. The difference between the predicted and true probability distributions is computed, with the true distribution represented by one-hot encoded labels.

$$\text{Loss} = -\sum_{i=1}^{n} t_i \log(p_i) \tag{2.2}$$

where n denotes the number of classes, t_i denotes the ground truth label and p_i denotes the softmax probability for the i_{th} class.

A Softmax Classifier is used for multi-class classification problems. The softmax function is used to obtain the predicted probabilities for each class by computing the exponential

of each input, normalising the results to obtain a probability distribution, and predicting the class with the maximum probability.

Dropout and Stochastic Depth Technique

Dropout [29] is a commonly employed method of generalization. Throughout each training epoch, neurons are lost at random. In this manner, the feature selection power is divided uniformly over the whole group of neurons, and the model is forced to learn different independent features.

Stochastic Depth [30] aims to reduce the depth of a network during training, while leaving it untouched during testing; it is a method for residual networks that removes or deactivates residual blocks at random during training. This technique is similar to dropout, except whereas dropout eliminates neurons, Stochastic Depth eliminates blocks (roughly the layers of a residual network).

CNN Learning Process

A massive set of images labelled with their corresponding class labels, e.g. (sad, disgust, happiness, fear, etc) is fed into the neural network during CNN training. The CNN begins with random weights, and it processes each image with these random weights, before making comparisons with the input image's class label. If the output does not match the class label, the CNN makes a small change to the weights of its neurons so that the output will exactly match the class label image and this lowers the loss function.

Using a process called **backpropagation**[31], the values of weights are adjusted. Backpropagation improves the tuning procedure and facilitates more precise changes. Each training session of the picture dataset is referred to as an **epoch**. The **learning rate** is defined as the parameter update step size. The training epoch is a complete repetition of the parameter update, including the whole training dataset. Note that, although it is a hyper-parameter, the learning rate must be chosen with care so that it does not impact the learning process inaccurately.

The Gradient Descent or Gradient-Based Learning approach repeatedly modifies the network parameters throughout each training session in order to reduce the training error. To precisely update the parameters, the objective function gradient (slope) must be computed using a first-order derivative with respect to the network parameters. To decrease error, the parameter is then modified in the opposite direction of the gradient. Using network back-propagation, the gradient at each neuron is back-propagated to all neurons in the previous layer during the process of parameter updating. This procedure can be resumed mathematically as follows:

$$w_{ijt} = w_{ijt-1} - \Delta w_{ijt} \tag{2.3}$$

where,

$$\Delta w_{ijt} = \eta \times \frac{\delta L}{\delta w_{ij}} \tag{2.4}$$

and w_{ijt} denotes the final weight in the present training epoch, whereas w_{ijt-1} denotes the weight in the previous $t-1$ training period. The learning rate is η, and the error in prediction (Loss) is L.

There are several variants of the gradient-based learning algorithm that are widely used, including the following: Batch Gradient Descent, Stochastic Gradient Descent, Mini-Batch Gradient Descent, Momentum, and Adaptive Moment Estimation (Adam).

During its training, CNN goes through a number of epochs, during which its weights change by the right amount. After each epoch step, the neural network gets marginally more accurate at identifying and accurately predicting the category of training pictures. As the CNN gets better, the changes to the weights get smaller.

After training the CNN, its accuracy is validated using a test dataset. The test dataset consists of labelled images that were omitted from the training procedure. CNN is fed with each image, and the output is compared to the test image's real class designation. In essence, the test dataset analyses CNN's prediction performance. If a CNN performs well on its training data but poorly on its test data, this is referred to as **overfitting**. On the other hand, an **underfit** model results when the model does not learn enough from the training data. **Justfitted** refers to a model that performs well on both training and testing data.

2.1.2 Fine-Tuning Transfer Learning

Deep CNN models need to be trained on a significant amount of data in order to achieve satisfactory results while addressing a variety of classification problems. The absence of sufficient training data is the most often encountered obstacle connected with the utilisation of such models. The Transfer Learning (TL) methodology [32] is consequently now being used to overcome the problem of small datasets and tackling the problem of inadequate training data.

During its training, the DL network is exposed to a massive amount of data and also learns the bias as well as the weights associated with each node. After that, these weights are moved to various networks in order to retrain them or test a fresh model that is comparable to the original. As a result, the cutting-edge approach makes it possible to pre-train weights rather than necessitating training from the ground up.

A lot of CNN models, like VGG [33], ResNet [34] and ConvNeXt [35], have been trained

on huge datasets like ImageNet [36] to recognise images. When there is not enough information to go on, these models are very helpful in a number of ways. To start, training large models on large datasets requires expensive processing power and is time-consuming. Lastly, a model that has already been trained can make it easier for the network to become more general and speed up the rate of convergence. As a result, TL eliminates the need to start from scratch when learning how to perform a new task.

Resuming, there are two types of transfer learning [3]: (1) fine-tuning the whole pre-trained network model so that the weights of every layer are optimised for the target dataset; or (2) using the pre-trained network model as a feature extractor, where the layers' weights are frozen and are used as feature extractors while the last layers' weights are optimised for the target dataset, acting as classifiers.

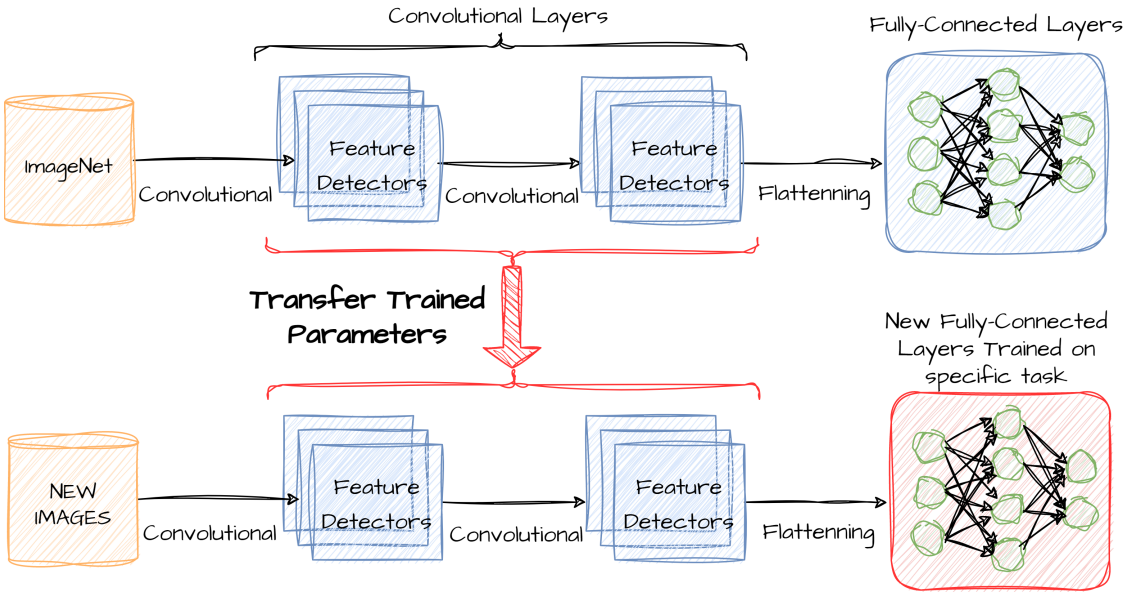

Figure 2.6: An example of the conceptual TL technique (adapted from [5] [6]).

2.1.3 Data Augmentation Techniques

The best strategy to prevent the model from becoming over-fit to the data and to get a decent generalisation is to train the model using a substantial quantity of data. Data Augmentation (DA) [37] refers to the process by which we utilise certain artificial techniques to increase the size of the training set in order to accomplish this goal. Some examples of possible techniques include flipping the image both vertically and horizontally, rotating it by a certain number of degrees, cropping the image, translating a few pixels, injecting noise, and changing the colour space.

3 State of Art

This chapter provides an overview of the most relevant works in Engagement Classification, being organised in five main sections: Robotherapy and Child-Robot Relation, Engagement, Models to Detect Engagement Levels, Emotion Recognition, and Engagement Recognition.

3.1 Robotherapy and Child-Robot Interaction

According to Lytridis et al. [38], the use of social robots in education has been a key focus of robotics research in recent years. Numerous studies have demonstrated the benefits of using robots as instructors or teaching assistants. These studies focus mostly on activities in which the child interacts with the robot to attain a particular educational or therapeutic objective. Children tend to be more interested in the educational process when a robot is involved, which is the primary reason why robots in education are reported to have a beneficial influence.

Robotherapy investigates possibilities and the impact of using robots in therapy sessions. On sensory-motor, affective, cognitive, and social levels, Robotic Psychology examines the compatibility between humans and robotic organisms. Robotherapy may be defined as a framework of human–robotic creature interactions designed to rebuild a person's bad experiences via the development of coping methods, mediated by technological instruments, in order to offer a platform for the creation of new positive life skills. These concepts are addressed in the article Libin et al. [7].

From David et al. [39]'s perspective, interactive/social robots often have an anthropomorphic look or resemble animals and imagined entities, engaging with humans for different objectives including educating, entertaining, and/or giving therapeutic assistance because they can exhibit human-like behaviour or other social interaction capabilities. They can be used as companions, for psychotherapy, or for physical and cognitive rehabilitation. The same work suggests that robots might play one or more of the following functions in robotherapy: Robo-Therapist, Robo-Mediator, and Robo-Assistant. The Robot-Therapist role provides a new method for psychotherapy but not a new type of psychotherapy, as their actions are specified and controlled by clinicians. The robots can essentially operate

as psychotherapists and potentially fully replace them in their absence. This may be the result of excessive costs, a lack of service providers, or the inability of psychotherapists to attend to patients at all times. The Robot-Mediator acts when the classical psychotherapeutic techniques cannot be applied to specific populations (e.g., when the clinical disorder renders patients less sensitive to human interactions, such as in the case of autism spectrum disorder (ASD)) and/or the target population would respond better to a treatment specifically implemented with a robotic mediating agent. The Robot-Mediator functions as a specialised and required "catalyst" that facilitates or accelerates therapeutic progress by mediating the interaction between the therapist and his or her patients. Similar to chemistry, if the catalyst is absent, the impact either does not occur or happens less effectively. In addition, the robot can serve as a motivator, making the intervention more understandable and appealing. The robot facilitates the therapist's role performance as a Robot-Assistant. Psychotherapists employ robots to complement and/or facilitate their traditional tactics during interventions. The robot is not required for psychotherapy, but its use can facilitate or enhance conventional therapeutic procedures. In David et al. [39] trial, the NAO robot was employed to cheer up young patients who required isolation in sterile hospital rooms after receiving a bone marrow transplant. Even though only qualitative input was collected from children, medical professionals, and parents, the results were largely favourable.

A few characteristics describe a robot friend as a good human companion, per Libin et al. [8]: It mimics a natural (human or animal-like) activity. It models motor, emotional, and cognitive behaviours commonly experienced by animals or people. It interacts with a person on several levels: tactile–kinesthetic, sensory, emotional, cognitive, and social.

3.1.1 Child-Robot Activity

Hadfield et al. [40] proposed a child-robot interaction in their research work. They used a NAO robot that approached one of a bricks and acted as if it were going to pick it up, but was unable to do so. It attempted to attract the kid's attention and induce the child to hand over the brick through a sequence of gestures. These movements included pointing at the brick, opening and closing its hand, shifting its attention between the child and the block, and a combination of hand and head movements. If the child did not comprehend the robot's intent after a specific amount of time, the robot would proceed to ask the kid vocally. After receiving the brick, the robot thanked the child and, at times, looked for another brick to hold.

Rudovic et al. [41] implemented a NAO robot-assisted therapy for children with Autism Spectrum Condition (ASC) that teaches them about emotional expressions. A therapist uses pictures of facial and physical expressions of basic emotions (e.g., sadness, happiness, and fear). The therapist next asks the youngster to identify the emotion displayed by

the robot. The kid is then encouraged to mimic the robot's facial expressions during the mirroring stage. If the child is successful, the therapist advances to the next level by telling a story and asking the kid to imagine how a robot would feel in a specific situation.

Javed et al. [42] used the Darwin Mini humanoid robot to explore five sensory stations set up on a table, each of which presented a specific sensory stimulation to which the robot responded interactively and in a social appropriate manner. In the seeing station: The robot approaches a box with a lid that opens to reveal a flashlight shining in its direction. The robot hides its eyes from the light by turning away. In the hearing station, the robot approaches a speaker that begins to play music, and the robot starts to dance. In the smelling station, the robot approaches a flowerpot, sniffs the flowers, and sneezes. In the tasting station, the robot approaches two dishes holding various types of food. It tastes one and expresses pleasure; then it tastes the other and expresses distaste. In the last station, touching station, the robot approaches a soft red blanket and touches it to indicate that it likes the blanket's soft texture.

Feng et al. [43] used a NAO robot to guide an autistic child to perform an interaction task. The interaction was as follows: first, upon detecting a kid, the robot changes its posture towards the child and moves at a proper distance; second, the robot encourages the child to communicate through language and action. Third, the engagement is evaluated, and according to the result, the robot adapts its behavior. If the evaluation of engagement reveals that the kid is interested, the robot will instruct him or her to replicate its actions. If the outcome of the evaluation of engagement is neutral, the robot will enhance current interactions and attempt to capture the kid's attention. If the evaluation reveals aversion, the robot will attempt to stimulate the child's interest by speaking and behaving. If the feeling of aversion persists for a predetermined amount of time, the robot will modify its behaviour to comfort the kid.

The activities described above are summarised in the Table 3.1 and the Table 3.2 describes in generic terms the child-robot interaction, relating the needs with the benefits.

Figure 3.1 relates child-robot interaction with engagement recognition and subsequent adaptation of the robot's behaviour to the child using reinforcement learning, not covered in this dissertation.

Table 3.1: Child-Robot activities table.

Article	Robot	Activity
Hadfield et al. [40]	NAO Robot	The robot went up to one of the bricks and tried to pick it up, but it couldn't. It tried to get the child's attention and get the child to hand over the brick by making a series of gestures or by speaking if the child did not comprehend the robot's intent
Rudovic et al. [41]	NAO Robot	Robot-assisted therapy for autism that teaches children with ASC about emotional expressions by encouraging them to mimic the robot's facial expressions during the mirroring stage
Javed et al. [42]	Darwin Mini Robot	The robot explored five sensory stations set up on a table, each of which presented a specific sensory stimulation to which the robot responded interactively and socially appropriately
Feng et al. [43]	NAO Robot	The robot guides the autistic kid to perform an interaction task by imitating its actions

Table 3.2: Classification of Interactive Engaging Robots with Regard to Human Needs and Benefits, adapted from [7][8]

Interactive Engaging Robot		**PERSON**	
Type	**Behavioural configuration**	**Need**	**Benefit**
Social Robots	Imitation of human facial expressions and complex gestures with social meaning or modelling basis emotional states and life-like behaviours	To provide company	Communication and companionship
Educational Robots		To entertain	Enrichment of learning skills
Robots with therapeutics potential		To alleviate negative mental states and psychological dysfunction	Therapy of negative states and behaviours

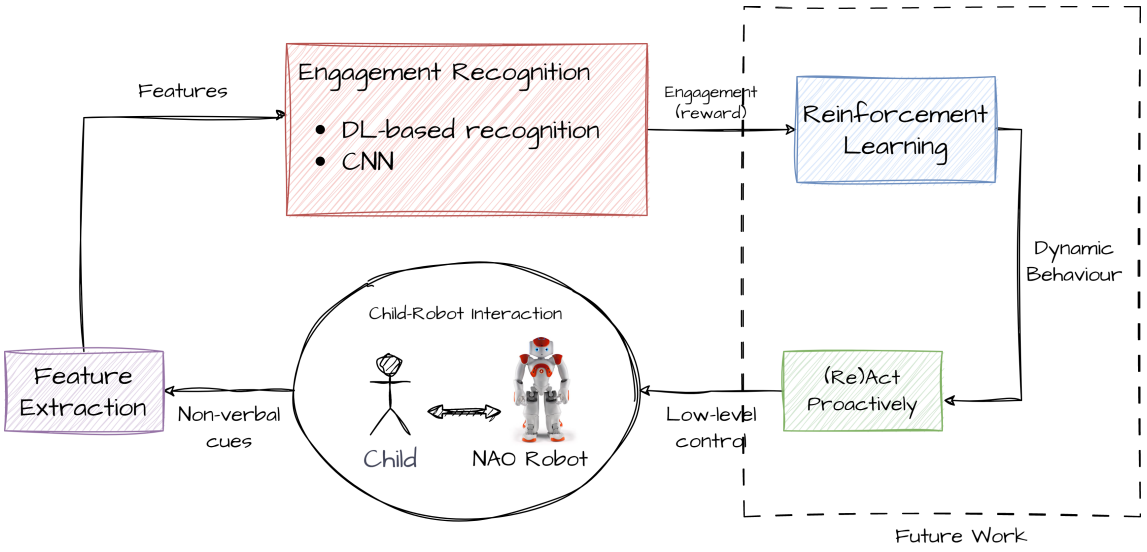

Figure 3.1: Child-Robot Interaction.

3.1.2 Breathing Techniques into Psychotherapy

According to Salomen et al. [44] breathing control is a technique for regulating and controlling both the pattern and depth of breathing, thereby facilitating shoulder relaxation. Slow and deep breathing enhances parasympathetic activity, which sends signals to the brain to relax the body and controls the body's response to stress, advocated by Jerath et al. [45], Magnon et. al [46] and Russo et al. [47]. Other breathing control methods have been studied by Mason et al. [48], Steffen et al. [49] and Vierra et al. [50] to have direct effects on health issues such as oxygen saturation, blood pressure, and heart rate variability. There are numerous strategies for controlling breathing for relaxation. One is Weil A. Weil's 4-7-8 breathing regulation, a breathing rhythm created by an American physician. The 4-7-8 method of breathing regulation involves inhaling, holding air, and expelling for the respective counts of 4, 7, and 8. The 4-7-8 breathing control is based on an ancient yogic technique called pranayama. It is meant to help you feel less anxious and sleep better at night.

These breathing methods are applicable to the child robot activity proposed in this dissertation, in which the robot can teach one of these techniques to the child.

3.2 Concept of Engagement

The concept of engagement has been investigated from a number of different viewpoints. One of these points of view refers that engagement is the process by which people start to feel connected, keep feeling connected, and then stop feeling connected to each other. It mixes verbal communication with non-verbal actions, all of which enhance the feeling

of connectivity between interactors, and it does this by combining the two [51]. Another definition describes engagement as a quality of the user experience that is characterised by attributes such as challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety or novelty, interactivity, and perceived user control. It states that there are four distinct phases during the process of engagement: the point of engagement, the period of sustained engagement, disengagement, and re-engagement [52].

For engagement recognition, some features can be used, such as gaze direction, emotions, speech detection, audio, contextual analysis, electroencephalography, posture [38]. In this dissertation, the engagement was classified based on emotion through facial expressions.

According to Anagnostopoulou et al. [53], the recognition of student engagement can be an important task because it is a factor in the improvement of the learning process as well as a qualitative indicator. From D'Errico et al. [54] perspective, to accomplish quality interaction between children and social robots, it is crucial that robots can adjust their behaviour to the cognitive state of children; for that reason, the recognition of engagement is an important task.

3.3 Models to Detect Engagement Levels

Several works [55][56][57][58][59][60] emphasise the significance of recognising emotions to determine a person's engagement in a certain task. There are several models of emotions, the majority of which show a direct relationship that establishes the connection between emotions and engagement levels. Emotions are a major focus of affective computing research. Emotions are related to complex internal states that include emotional, cognitive, physiological, expressive, and motivational components.

Khawlah et al. [1] builds a new engagement model with many levels based on [55]-[60] models: These stages are: Strong Engagement, High Engagement, Medium Engagement, Low Engagement, and Not Engaged and each of these levels correspond to different emotions that can be recognise through facial expression analysis.

Watson and Tellegen's [55] approach maps two levels of engagement to various emotions. Russel and Feldman [61] have since modified the Watson and Tellegen model with a 45-rotational theory. Consequently, they modified the Russel and Feldman model. First, they retained just intellectual feelings and eliminated all others, and second, incorporated additional academic emotions for which Remington et al. [59] assessed their placement (0-360 degrees).

Strong engagement and disengagement have maintained the same position as predicted by Watson and Tellegen. The high degree of participation (engagement) was characterised by anger, enthusiasm, and excitement. This level has both significant positive and negative effects that are related with a positive (attractive emotions) or negative valence (repulsive

emotions). Thus, when a child has a high amount of positive or negative emotions, he or she may have a high level of engagement. The medium degree of engagement comprised emotions of contentment, happiness, pleasure, delight, and sadness. This level contains both pleasant and unpleasant feelings. Consequently, a student's degree of involvement may be moderate while he or she is experiencing positive or negative emotions. The low level of engagement was marked by tiredness, boredom, and relaxation. This level has a small amount of positive and negative impact. Thus, whether a child has low positive affect or low negative affect, engagement may be low. The diagram depicted in Fig.3.2 presents the emotional model developed by [1].

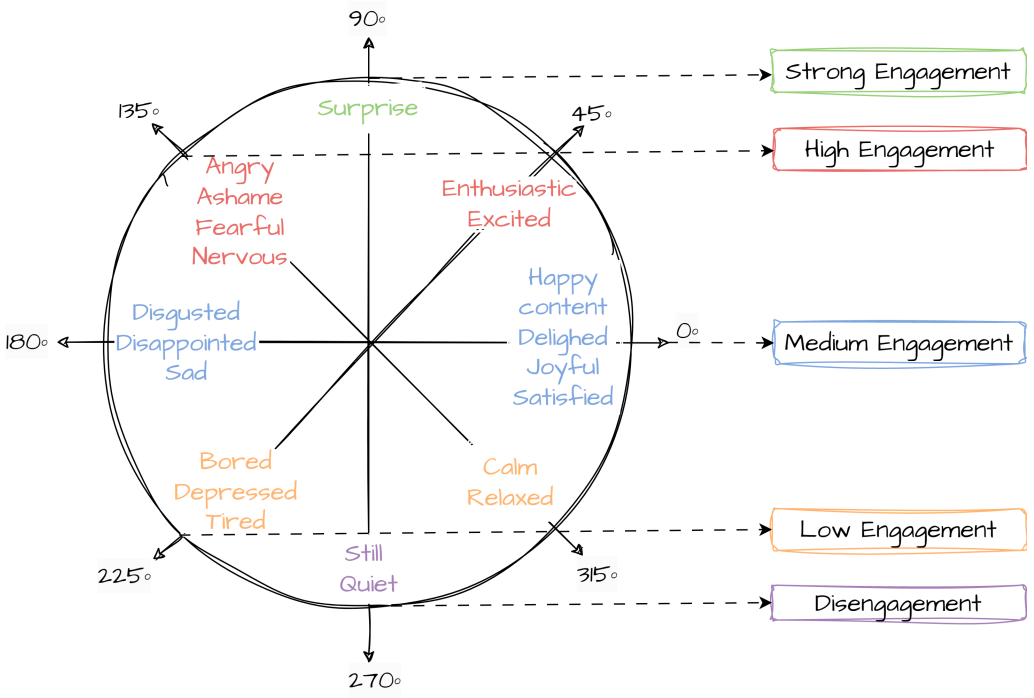

Figure 3.2: The affective model developed by [1], in which emotions are directly related to levels of engagement.

3.4 Emotion Recognition

A selection of representative studies on emotion recognition through facial expressions is highlighted below and summarised in the Table 3.4.

Witherow et al. [62] used a network to learn generic facial expression patterns from adult expressions, which were then fine-tuned using the transfer learning technique to capture representative features of kid facial emotions. They construct and train a CNN model for classifying adult and child facial expressions. Their model architecture comprises three convolutional layers with the ReLU activation function, batch normalisation, and dropout. Maximum pooling was used following each convolutional layer. A ReLU-activated fully connected layer and a softmax classification layer follow the convolutional layers. Their efforts yielded 76,03% on CAFE dataset.

Zheng et al. [63]) classify children and adult datasets using a Support Vector Machine (SVM) classification-based facial expressions recognition algorithm, achieving an accuracy of 77.40% on the CAFE dataset. They also study the differences in facial expressions between children and adults.

Liu et al. [64] model was tested on the FER2013 dataset and consists of three different structured CNN subnets trained separately. The three subnets contain eight to ten layers, respectively. The features extracted by these subnets are concatenated together by adding a fully connected layer at the end, and a softmax layer is used as the output layer of the whole network. The whole network is structured by assembling these subnets together. They achieved an accuracy of 62,44% classifiying the seven classes of emotions provided in the FER2013 dataset.

Khaireddin et al. [65] works used a VGGNet architecture to classify emotions through feeding facial expression images into the model. The CNN had four convolutional blocks that extract high-level features, and the fully connected layers classify the emotion of the image, obtaining 73,28% on the FER dataset.

Lopez-Ricon et al. [66] in order to categorise facial expressions in children, they compare the AFFDEX SDK and a convolutional neural network (CNN) with Viola-Jones trained using the AffectNet dataset and tuned with the NIMHChEF dataset using transfer learning. Then, they compared the CNN and the AFFDEX SDK for classification on the CAFE dataset, achieving the best performance with the CNN-AFFDEX Viola Jones retrained model at 44.88%.

Pramerdorfer et al. [67] identify existing bottlenecks, and by forming an ensemble of modern deep CNNs, they were able to obtain a test accuracy of 75,20% on the FER2013 dataset, outperforming previous works without requiring auxiliary training data or face registration.

Table 3.3: Representative studies on Emotion Recognition through facial expression.

Article	Dataset	Methods	Results
Witherow et al. [68]	CAFE dataset	CNN	Acc=76,03%
Zheng et al. [63]	CAFE dataset	Shape features + SVM	Acc=77,40%
Liu et al. [64]	FER2013 dataset	CNN	Acc=62,44%
Khaireddin et al. [65]	FER2013 dataset	VGG	Acc=73,28%
Pramerdorfer et al. [67]	CAFE dataset	Ensemble of modern deep CNNs	Acc=75,2%
Lopez-Ricon et al. [66]	CAFE dataset	CNN-AFFDEX Viola-Jones Re-Trained	Acc=44,88%

3.5 Engagement Recognition

This dissertation focus primarily on the automatic recognition of children's engagement while interacting with an humanoid robot, in this case the NAO robot. In recent years, automatic engagement recognition became an hot research topic, mostly motivated by the massification of distance learning due to the COVID19 pandemic. A set of representative studies, is highlighted below and summarised in the Table 3.4.

Lin Geng et al. [69] investigated student participation in online courses. Using the DAiSEE dataset and a 3D convolutional network consisting of eight convolutional layers, five maximum pooling layers, two fully connected layers, and a softmax classifier. In [69] the problem of unbalanced data distribution was approached by employing focal loss. The proposed methodology was able to extract temporal and spatial information from videos and classify engagement in four levels with 56.2% of accuracy.

Woo Han Yun et al. [19] investigated an automated approach for detecting children's participation in an educational setting using convolutional neural networks (CNNs). They produce their own video dataset. The CNN, a VGG FaceNet, extracts the low-level features. They use a temporal dynamics module to generate high-level features, which are then passed through a fully connected layer and a softmax layer to classify engagement into two levels: engaged and disengaged. They attain an accuracy rate of 81.44%.

The work developed by Omid et al. [70] provides a deep learning model that addresses the data sparsity barrier, by performing pre-training on the widely accessible FER 2013

dataset before training on specialised engagement data developed by the authors. They proposed a method composed of two stages: the first uses a CNN adapted to VGG-B to train facial emotion recognition model to offer rich face representation; the second stage uses the model's weights to initialise their deep learning-based model to identify engagement, which consists of four blocks of two convolutional layers followed by a max polling layer, three fully connected layers, and a softmax classifier used to classify two classes: engaged and disengaged with an accuracy of 72.38%.

Ognjen Rudovic et al.[41] have created a method to automatically measure the child's emotional states and engagement, which can subsequently be utilised to optimise child-robot interaction and track therapeutic success. In their approach they adopted a very complete input dataset that includes: NAO robot position, the child's gaze direction, facial expression, body posture, tone of voice, heart rate, skin conductance (SCT), body temperature, and accelerometer data, [41] proposes a personalised deep learning framework: the Personalized Perception of Affect Network (PPA-net).

Jack Hadfield et al. [40], estimate the level of a children engagement achieving 77,11%, during a child robot joint attention task, based on the position of the robot, the position of the child, the angle between the child's gaze and the robot, the angle between the child's body facing and the robot, and the distance of the hands from the respective shoulders. Their approach is a multi-view deep-based estimation of the child's pose, using cameras strategically placed in a room to extract the relevant keypoints and using a deep neural network to classify the engagement levels based on these features.

Table 3.4: Representative works regarding Engagement Recognition.

Article	Objective	Sensors	Dataset	Features	Methods	Results
Lin Gen et al. [69]	Study the engagement of students in online courses	Computer camera	DAiSEE	Spatiotemporal feature	3D Convolutional Networks	Acc=56,2%
Woo Han et al. [19]	Automatic children engagement recognition method based on CNN focused in an educational environment	RGB video camera	Own Dataset	Low-level features/High-level features extracted by CNN	VGG facial pretrained network; Temporal dynamics module	Acc=81,44%
Omid et al. [70]	Automatic Recognition of Student Engagement using CNNs to detect facial expressions and fine-tune classification of engagement		FER 2013; Own Datatset	Facial Expressions	Adapted VGG-B	Acc=72,38%
Rudovic et al. [41]	Automatically estimate levels of the child's affective states and engagement then be used to optimize the child robot interaction and monitor the therapy progress	Sensors on the child's wrist; Camera; Microfone	Own dataset	Robo NAO position, child's gaze direction; facial expression; body posture; tone of voice; heart rate; skin condutance (SCT), body temperature and accelerometer data	Personalized Perception of Affective network (PPA-net)	ICC=59%
Hadfield et al. [40]	Estimate the level of engagement between the child and the robot	4 Kinect Cameras	Own dataset	Robot's position; Child's position; Angle between child's gaze and the robot; Angle between the child's body facing and the robot; Distance of the hands from the respective shoulders	DNN	Acc=77,11%

4 Materials and Methods

This chapter summarises the examined datasets, hardware materials, and software tools used in the work described in this dissertation.

4.1 Datasets

In a classification or detection task, a representative dataset is essential. Finding a suitable public dataset to train and test the proposed engagement classification methodologies was tough because datasets with children are just a few, and those that exist are not openly accessible because of children's privacy and image rights and concerns. The PInSoRo dataset [71], the Dartmouth Database of Children's Faces [72], CAFE dataset [73], FER2013 [74] and the EmoReact dataset [75] are five usable datasets.

The PInSoRo dataset is a large open dataset containing recordings of 120 children who recorded child-child and child-robot interactions during free play. This dataset seemed to have lots of potential in terms of available data, however the access to this dataset is extremely complicated and time-consuming.

The Dartmouth Database includes images of sixty male and sixty female models between the ages of six and sixteen. Five different camera angles and eight distinct facial expressions are used to photograph the models. The models are taken against a dark background while wearing black helmets and bibs to hide their hair and ears. Although there is an access request methodology that is apparently less complex than the one required to the PInSoRo dataset, so far, no response was obtained to the requests performed officially by the ISR-UC.

The EmoReact dataset is comprised of 1102 annotated videos representing the multimodal emotions of children aged four to fourteen. A prompt affirmative response was obtained after formal request on EmoReact's website, and for that reason this dataset was used for a preliminary attempt.

The Child Affective facial Expression (CAFE) dataset consists of 1192 photographs of children ages 2 to 8 exhibiting six basic emotions. A formal request has been made by the ISR-UC to use the CAFE dataset, which was accepted and enabled its use in the described work.

The publicly available dataset Facial Emotion Recognition 2013 (FER2013), which consists of 48x48-pixel grayscale images of adult faces, was also used to train, test and validate the proposed emotion and engagement classification methodologies. Despite the fact that it is composed by adult images due to its extensive use in other projects, it allows the comparison of the proposed methods with others of the state-of-art.

4.1.1 EmoReact Dataset

The EmoReact Dataset consists of 1102 audiovisual clips annotated for seventeen emotional states, representing the multimodal emotions of children between the ages of four and fourteen, of various ethnicities and genders. EmoReact was developed in 2016 using videos downloaded from the now-defunct YouTube channel "React Channel," which included children's reactions to food, technology, YouTube videos, and gaming devices. Each of the downloaded original videos had multiple children reacting to an issue. Using ELAN [76], they manually split the videos into five-second clips, so that each clip contained only one child's response to a certain topic. In this segmentation, only videos longer than three seconds were evaluated, resulting in a total of 1254 clips. To produce their labels, they utilised Amazon's Mechanical Turk (MTurk), an online crowdsourcing platform [77]. After making a determination based on their preliminary research, they selected a group of six workers, consisting of three males and three females. Three independent workers annotated each video for seventeen labels. They acquired the labels from MTurk and then evaluated the level of worker agreement using Kripendorff's alpha [78]. In this phase, 152 videos were removed since it appeared the annotators were imprecise. The final batch of 1102 videos was acquired following this processing. They selected the eight emotion categories with the highest levels of coder agreement for their initial dataset analysis and research experiments. However, it was allowed that each video could contain multiple emotion labels. The length of the videos in the dataset runs from 2 to 10 seconds and were recorded at 24 FPS (Frames Per Second). The dataset was divided into three subsets, each containing 432, 303, and 367 videos: the training set, the validation set, and the test set. In order to increase the generalisation of models and results, these sets are defined in a manner that is person-independent (people from training data are not included in validation or test data).

The authors of the dataset provided three folders containing the aforementioned subsets as well as three text files with the corresponding names. Additionally, they provided text files containing the names of videos in the training, validation, and test sets in the exact order they appear in the labels and features. The selected emotions were curiosity, uncertainty, excitement, happiness, surprise, disgust, fear, and frustration, and their presence in the text files was determined in this precise order. In other words, there is a text file containing the same number of lines as the number of videos in each folder. The eighth

emotion are represented by each column, while valence (mean value of attractiveness or repulsion of emotions; a negative value implies repulsive emotions and a positive value represents attractive emotions) is defined by the final column. Therefore, a "1" will be entered in the column if the emotion was present in the video. For instance, the expression [1,0,0,0,1,0,0,0,3.2222] indicates that the video contains facial expressions of curiosity and surprise with a valence of 3.2222.

Table 4.1 shows the EmoReact distribution of videos for training, validation and testing.

Table 4.1: Distribution of videos for training, validation and testing.

	Train	**Validation**	**Test**
Number of videos	432	303	367

4.1.2 Dataset CAFE (The Child Affective Facial Expression)

The CAFE dataset was designed to offer fresh data to the field of emotion development research on the emotional facial expressions of children aged two to eight. The collection consists of 1192 images of 154 children, 90 girls and 64 boys (27 African Americans, 16 Asians, 77 Caucasian/European Americans, 23 Latinos, and 11 South Asians), posing with seven distinct facial expressions (happy, surprise, angry, disgusted, fearful, sad, and neutral). All images were shot against a white background, and the photographer, an assistant researcher with laboratory expertise in developmental work with children, encouraged the kids to express their feelings naturally. Not all children were able to replicate the seven expressions, and they were therefore eliminated from the dataset. The image labels were created by 100 undergraduate students, half of whom were female and half of whom were male, from a variety of racial and ethnic backgrounds, using two methods for validation: the Facial Affective Coding System (The Facial Affective Coding System (FACS))[79] to identify each facial expression and a second method in which untrained research participants identified each facial expression, followed by establishing agreement between raters. The research employed a mix of the two methodologies, with the FACS-trained photographer photographing all the children and the untrained volunteers being asked to identify each shot in the set on two separate occasions. Thus, while a FACS-trained researcher photographed the children, untrained individuals were asked to interpret the children's facial expressions.

After submitting a formal request via the Databrary platform (www.databrary.org), a free open data repository for development research, the dataset was made available to ISR-UC research team. The dataset is organised into 45 sessions; each session corresponds to an individual child who pose one of the seven emotions. With the exception of surprise, angry and disgust, CAFE characterises emotions with the mouth open and closed or with

the tongue; hence, each session can contain between seven and thirteen representative photos.

4.1.3 Dataset FER2013

The FER2013 dataset contains 48x48 grayscale images of adult faces. The faces are indicative of seven facial expressions and have been mechanically registered such that the face occupies roughly the same amount of area and is roughly centred in each image. The training set has 28709 examples, whereas the public test set contains 3589 examples, representing 24% of the training set. The dataset is organised by training and testing, with seven subfolders for each emotion. Thus, the dataset had the required templates for the work described in this dissertation, without the need for further processing.

Table 4.2 shows the distribution of the quantity of facial expression images by dataset FER2013 training and testing.

Table 4.2: Distribution of the quantity of facial expression images by dataset FER2013 training and testing.

Emotions	**No. of training images**	**No. of testing images**
angry	3995	958
disgust	436	111
fear	4097	1024
happy	7215	1774
neutral	4965	1233
sad	4830	1247
surprise	3171	831

Table 4.3 shows the used datasets and their applications in this dissertation.

Table 4.3: A listing of the datasets examined.

Datasets	**Type**	**Ages**	**Utility**
EmoReact	Children	4-14	Initial tests
FER2013	Adults	-	Validation of our network's
CAFE	Children	2-8	Selected one

4.2 CNN Networks

The primary CNN architectures that were utilised in this research were VGG16[33], ResNet18 [34], ResNet50 [34], ResNeXt50 [35], and ConvNeXt [35]. These CNNs were

chosen because of their ability to perform well in previous research. The selected network will be detailed in further depth in Section 4.2.1. Figure 4.1 illustrates the structure of each of the six different types of architecture.

The parts of each of these six networks include convolutional layers, pooling layers, activation functions, and FC layers. This is something that all of these networks have in common. The number of convolutional layers, the type and position of pooling layers, the type of activation functions, and the number of FC layers vary amongst models.

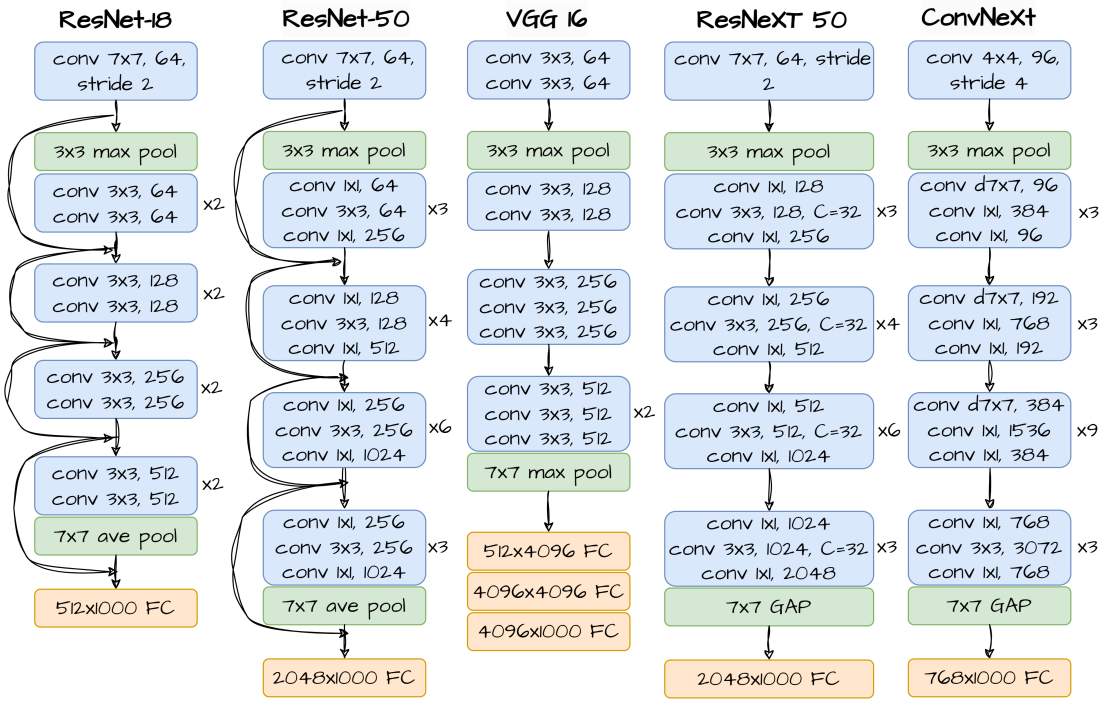

Figure 4.1: Architecture comparison of VGG16, Resnet18, ResNet50, ResNext50, and ConvNeXt.

4.2.1 ConvNeXt Network

The ConvNeXt network is rather new, having been launched only last year. Liu et al. [35] discovered the network while attempting to answer the question, "How do design decisions in Transformers impact ConvNets' performance?" To answer this question, they conducted a set of tests on the ResNet50 network, matching the features of a conventional CNN to those of a transformer by employing tiny tricks employed in transformers in conjunction with advancements in regularisation, augmentation, and optimisation.

The following tests were performed as part of this strategy: increasing the number of training epochs from 90 to 300; using AdamW Optimizer; Using several data augmentation techniques, such as Mixup, CutMix, RandAugment, and Random Erasing; Using regularisation schemes, such as stochastic depth and label smoothing; Changing the stage

compute ratio; Changing stem to "Patchify"; Using depthwise convolution; Using inverted bottlenecks; Using large kernel sizes; Replacing ReLU with GELU; Using fewer activation functions; Replacing BN with LN and separate downsampling layers.

The figure 4.2 below provides an overview of the network.

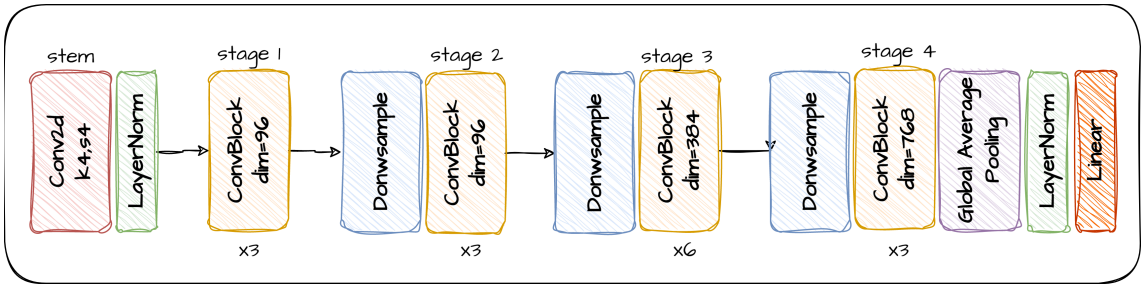

Figure 4.2: ConvNeXt architecture.

4.3 Performance Metrics

The classification accuracy of each model was determined to assess the performance of the frameworks. Classification accuracy is defined as the percentage of correct predictions relative to the total number of predictions:

$$\text{Accuracy}(\%) = \frac{\text{No. of correct predictions}}{\text{Total no. of predictions}} \times 100 \tag{4.1}$$

In its most basic form, the confusion matrix is a table that presents the results of classification efforts in a graphical format. This matrix is of the format N by N, where N is the total number of classes. The rows in this table reflect the predicted instance of the class, while the columns represent the true label of the class. The cells that indicate the incorrect predictions are the ones that are not on the main diagonal, which contains the accurate predictions.

The picture below illustrates an example of a confusion matrix for a binary classification consisting of only two classes. Where TP, or true positives, were the successful positive classifications, TN, or true negatives, were the right negative classifications, FP, or false positives, were the erroneous positive classifications, and FN, or false negatives, were the inaccurate negative classifications.

4.4 Hardware Materials and Software Tools

4.4.1 NAO Robot

The NAO robot was founded in the French firm Aldebaran Robotics from Bruno Maisonnier's boyhood ambition.

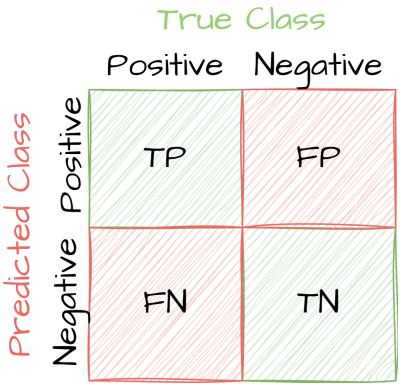

Figure 4.3: Confusion Matrix for a binary classification

NAO is a completely programmable humanoid robot that has proven beneficial in education, healthcare, research, and customer-facing contexts. Highlighting the field of education, it may be a fantastic learning tool to accompany students of all levels and help them improve, among other skills, social, literacy, programming, and research abilities.

The 57cm tall NAO is equipped with a variety of sensors, including: seven touch sensors on the hands, feet, and head; pressure and acceleration sensors; sonar; an inertial unit; four microphones; a speaker; a gyroscope; two infrared emitters and receivers, two ultrasonic sensors, and two 2D cameras. It is also capable of recognising twenty distinct languages and has twenty-five degrees of freedom, allowing for fluid movement. Finally, it may be programmed using its own platform, Choregraphe, and the programming languages C#, Java, or Python, enabling its usage in a variety of subjects.

The robot is termed humanoid because it is capable of human-like behaviours and can be programmed to execute a variety of acts, including dancing, sitting, standing, walking, communicating, responding to stimuli, and detecting objects, among others.

Figure 4.4: NAO Robot.

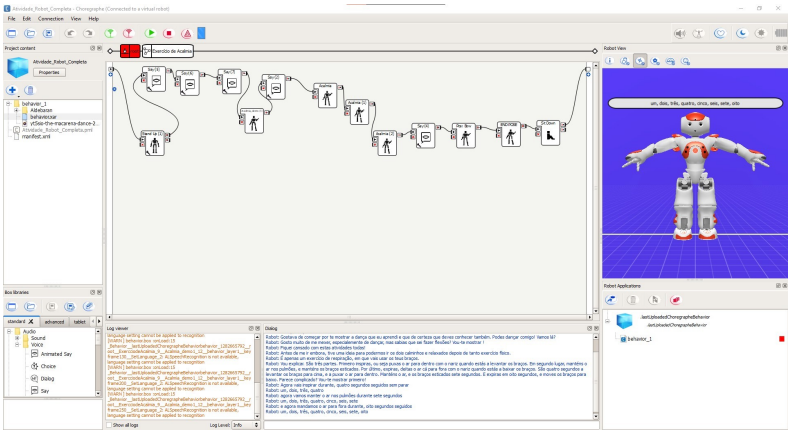

Figure 4.5: Choregraphe Software.

4.4.2 Choregraphe and NAOqi Framework

Choregraphe [80] is a cross-platform desktop application for Robot NAO that allows the creation of complex behaviors, dances, and human interactions.

The NAOqi Framework is the framework for programming and controlling NAO. It manages common robotics needs like parallelism, resources, synchronisation, and events. This system offers homogeneous information exchange, homogeneous module-to-module communication, and homogeneous programming (motion, audio, and video). The framework is cross-platform, therefore it may be used on Windows, Linux, and Mac for development. C++ and Python use the same API, making it cross-language. Also permits introspection, meaning the framework is aware of which modules provide access to which functions.

The created behaviours of Choregraphe are written in its own graphical language, which NAOqi understands and executes. Choregraphe interfaces with NAOqi to provide useful tools like the video monitor panel, the behaviour management panel, the toolbar, the robot view, and the timeline editor. This software includes, among other preprogrammed behaviours, stand up, sit down, speech, and recognise speech. This approach, termed as "Animation Mode," simplifies the programming of complex behaviours in Choregraphe. In addition, we can test it on a virtual robot before sending it to a real robot.

4.4.3 Python, PyTorch and NVIDIA GeForce RTX 3060

Python [81], which is frequently used in the development of ML algorithms, was chosen as the programming language for this project. A variety of Python library packages presented in Table 4.4, were employed in the course of the work described in this dissertation. Pytorch [82] is a free and open-source framework for ML that provides users with the instruments required to construct, parameterise, train, and optimise artificial neural networks.

Visual Studio Code (VSCode) [83] was selected as the IDE for developing the code as it allows us to use Jupyter Notebooks, which helps the writing, debugging, and execution of code.

Training Artificial Neural Networks (ANN) with many layers takes a long time and is hard to do. But this problem can be fixed by letting a Graphics Processing Unit (GPU) handle the work of updating the network. For ML libraries to run on a GPU, the graphics card must have access to the latest drivers and support Compute Unified Device Architecture (CUDA) [84] and its libraries. CUDA is an NVIDIA framework for parallel computing that lets developers use the full power of the GPU's graphics processors. The NVIDIA CUDA Deep Neural Network library (cuDNN) [85] is a library of Deep ANN primitives that is sped up the GPU. A NVIDIA GeForce RTX 3060 [86] was used as the GPU for this project. Table 4.5 shows the most important details about it.

Table 4.4: Python packages.

Package	**Version**
Python	3.10.5
Torch	1.11.0
Numpy	1.22.4
Matplotlib	3.5.2

Table 4.5: NVIDIA GeForce RTX 3060 specifications.

Cuda Cores	3584
Video Memory	12 GB GDDR6
Memory Bus	192-bit
Engine Clock	Base:1320 MHz, Boost:1777 MHz
Memory Clock	15 GHz
Power Consumption	170W
Supported OS	Windows, Linux

5 Developed Work

This chapter describes the dissertation's developed work and all steps leading up to its completion, in particular: pre-processing of datasets; design and development of the engagement classification system; research and development of two different methods for automatic engagement classification; and design and implementation of child-robot interaction activity using a NAO humanoid robot.

5.1 Datasets pre-processing

In this dissertation, three public datasets were used for training, test and validation of the proposed automatic engagement classification methodologies, namely: EmoReact, FER2013, and CAFE.

The EmoReact dataset is divided into three sets: training, validation and testing sets, containing 432, 303 and 367 videos, respectively. The EmoReact dataset also includes three text files providing the labels for each video and another group of text files giving the order of the video labels. However, both the folder structure and the labelling provided in each recorded video are not appropriate to the engagement classification system proposed in this dissertation. Due to that reason, a pre-processing of the dataset was required to reorganise the data and label them in a way they can be used as inputs to the proposed engagement classification system.

The FER2013 is a dataset of adults that has been extensively studied by the scientific community. It was used to validate several results, with the advantage of not requiring any further processing, and it can be applied directly to the proposed engagement classification system.

The CAFE is a dataset composed of images with faces of children. To be used in the proposed engagement classification system, a simple reorganisation of image sessions was required.

5.1.1 Organisation of EmoReact dataset folders based on emotions

The EmoReact dataset is originally separated in three folders: training, validation, and test, which is not suitable to the proposed engagement framework. The reorganisation of the dataset consisted in subdividing each folder (corresponding to training, validation, and test sets) in eight sub-folders, each referring to an emotion (curiosity, disgust, excitement, fear, frustration, happiness, surprise). To organise the videos in subfolders of emotions it was necessary to carefully analyse the text files containing the labels present in each video (emotions based on facial expressions) and the file with the names of the videos, to figure out which labels of emotions correspond to each video, once the labels are in the corresponding order of the video names. It is worth noting that each video contains more than one facial expression, and this fact makes reorganising the dataset very challenging. A diagram summarising the reorganisation process is shown in Fig. 5.1.

The folders reorganisation process was performed automatically with a Python code specifically developed for that purpose. First, the video's name was picked in iteration x, and then the correct video was selected in the folder corresponding to videos, i.e., the video with the name specified by the name file in iteration x. On line x, line with the number corresponding to iteration x, of the labels file, the columns are examined to determine the presence or absence of the emotion whose facial expression was displayed in the video. The same video is subsequently moved to the folders corresponding to the emotions based on facial expressions associated with it in the labels file. This means that if a video contains three distinct emotions, it will be copied to the three folders corresponding to those emotions.

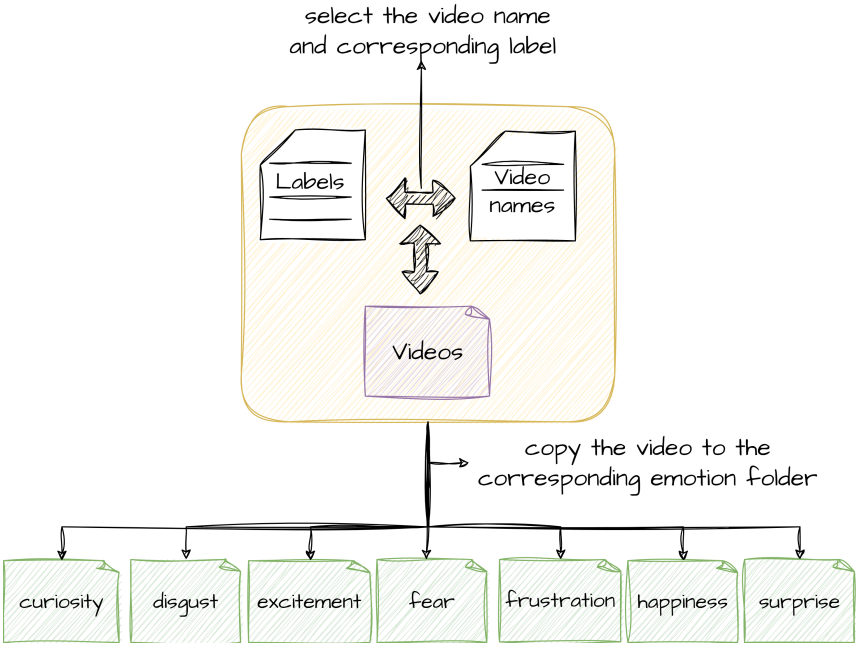

Figure 5.1: EmoReact reorganisation process.

5.1.2 Organisation of CAFE dataset folders based on emotions

The CAFE dataset is arranged in sessions of seven to thirteen images, with each session corresponding to a particular child. The difference in the number of images per session depends on whether it comprises (or not) facial expressions with open mouth. In the original organisation of the dataset, each image of a particular child is labelled by the emotion based on his/her facial expression.

In order to prepare the CAFE dataset to be applied as input to the proposed engagement classification system, it was necessary to reorganize all the images into seven folders, each corresponding to a different emotion (angry, disgust, fearful, happy, neutral, sad, surprise). Table 5.1 shows the result of distributing images per emotions in the CAFE dataset.

A Python script was created to perform this operation, which consists in reading the filename of each image and move it to the appropriate emotion folder. Since both "angry" and "angryopen" represent the primary emotion, both were moved to the same folder, in this case "angry".

Table 5.1: Distribution of the quantity of facial expression images by dataset CAFE training and testing.

Emotions	**No. of training images**	**No. of testing images**
angry	158	47
disgust	148	43
fearful	108	32
happy	166	49
neutral	178	52
sad	84	24
surprise	80	23
Total	**922**	**270**

5.1.3 Extraction and image cropping using MTCNN

The emotion recognition method, used in one of the approaches to classify engagement 5.2.1, uses images to extract the relevant features for emotion based on expression recognition. For the datasets composed by videos, (e.g. EmoReact, and our own) it is necessary to extract frames to be used as static images. The videos contained 24 Frames Per Second (FPS). Due to the different duration of the videos, it was important to make sure that a constant number of frames per video was chosen. In this case a constant number of 20 frames were extracted from each video.

The step corresponds to the number of frames that must be skipped between extracted

frames, and is calculated according to:

$$Step = \frac{\mathrm{T} \times 24(\mathrm{FPS})}{\mathrm{TNF}} \tag{5.1}$$

where T corresponds to the total time of the video in seconds, and TNF corresponds to the total number of frames. Figure 5.2 shows how the frame extraction is performed in each video.

Considering that the purpose is to detect facial emotions, after obtaining the frames from each video, the image was cropped to only include the children's faces. Everything else in the image is irrelevant to emotion detection and can be discarded as noise. The images were cropped using the Multi-Task Cascaded Convolution Neural Networks (MTCNN) [87] method, which are able to detect faces in an image using three stages of CNNs; therefore, it is only necessary to select the bounding boxes of the various faces and save them to a folder

In order to save resources, memory, and time, these two steps were performed sequentially, i.e., the frames were extracted from the video and the crop function was called prior to saving the images. After that, the images of the children's faces were saved in the folder for the eight different emotions, as depicted in Fig. 5.3.

This method was used on both EmoReact and the data collected during the recorded sessions of the child-robot. The only difference is that for the data from the test, 500 frames were chosen from each of the three videos.

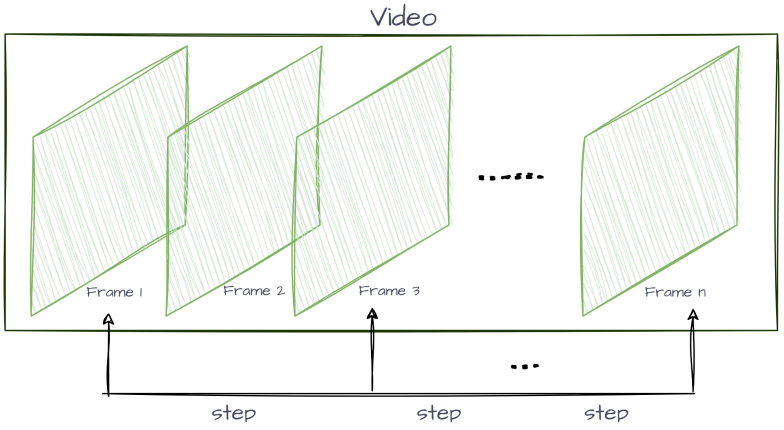

Figure 5.2: Diagram illustrating the process of extracting frames from a video.

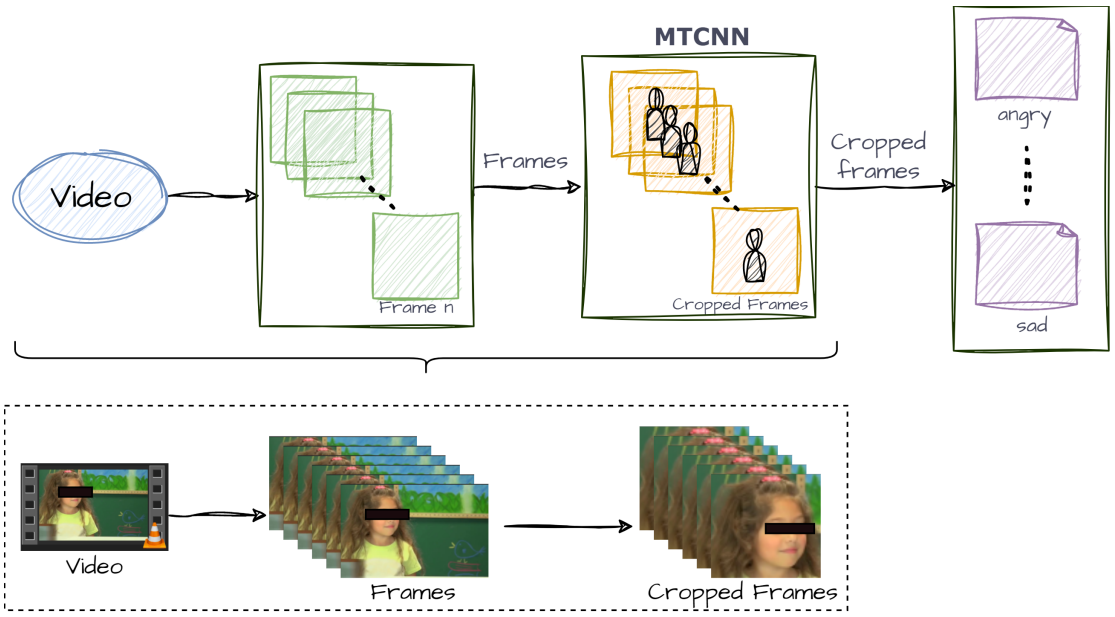

Figure 5.3: Diagram illustrating the process of cropping frames from a video.

5.1.4 Selection of images

After reorganising the children's images in folders corresponding to emotions, it was necessary to manually select and analyse the images one by one to eliminate incorrect detection of children's faces, and to correct poorly made crops. There were a total of 300.000 frames in the folders, making the selection a time-consuming task.

Another problem that occurred during the dataset pre-processing was that since each video may contain more than one emotion there were frames that did not represent any emotion or that expressed a different emotion than the one indicated by their folder. This fact occurs because facial expressions are not labelled for each frame but rather for each video. Consequently, it was necessary to re-select manually the representative images for each emotion, which was a very arduous and time-consuming task and very error-prone. Table 5.2 presents the information contained in the pre-processed EmoReact dataset.

5.2 Engagement Classification System

This section describes the emotion recognition approach and the two proposed engagement classification methods.

Table 5.2: Distribution of the quantity of facial expression images by dataset EmoReact training and testing.

Emotions	No. of training images	No. of testing images
curiosity	16262	4554
disgust	4763	835
excitement	14067	2569
fear	1120	174
frustration	5818	1632
happiness	27894	7712
surprise	31174	1666
Total	**73098**	**19142**

5.2.1 Emotions Recognition

This section describes the work performed to recognise emotions based on facial expressions. The model for emotion recognition based on facial expressions was built using the theoretical and practical CNN architectures described in Sections 2.1.1 and 4.2 and applying the techniques and concepts in Section 2.1.2 and 2.1.3. Algorithm 1 presents the general procedure for emotion recognition based on facial expressions.

The main objective was to design a network that was capable of accurately classifying seven distinct emotions. Each network was fed with images of children's faces from the dataset, each of which was labelled with the corresponding facial expression. A resize can be used to change the size of an image; the larger the image size, the more information the network has to train and learn, but it also takes longer to train the network. Additionally, a data augmentation technique was used to apply a horizontal flip, a vertical flip, and a rotation of a specific angle to the images and add them to the original dataset. These three techniques were evaluated simultaneously, in pairs, and separately (e.g., lines 2-3).

The network needs to be initialised; in this work, the ResNet18, VGG16, ResNet50, ResNeXt50, and ConvNeXt networks were experimented with, and during the initialisation, it was decided whether to use the pre-trained network or not. The transfer learning technique, specifically the fine-tuning sub-technique, was used by selecting a network pre-trained with ImageNet data. ImageNet is a visual dataset with more than 14 million pictures that are meant to be used for research on visual object recognition, (e.g., lines 4-5).

In this work, two types of optimisers (learning rate) were used: an invariant learning rate with a fixed value and the ADAM (Adaptive Moment Estimation) learning rate, which adapts its value during the network training phase. ADAM combines two main concepts to

improve the efficiency of the algorithm: adaptive first- and second-order moment estimates and adaptive learning rate updates. The first-order adaptive moment estimates, known as the Exponential Weighted Moving Average (EWMA), are used to compute an estimate of the mean gradient, which is then used to update the model parameters. Adaptive second-order moment estimates, known as the Squared Exponential Weighted Moving Average (EWMA Squared), are used to calculate an estimate of the second moment of the gradient and are then used to adjust the learning rate, which can be adjusted differently for each model parameter. Momentum is calculated as a weighted average of previous gradients and is added to the gradient update to help the algorithm follow a more consistent direction on irregular loss surfaces or ones with many local minima. Momentum has a practical effect in that it lets the optimisation algorithm gain speed in the right direction and "cross" flat or low-gradient areas faster without making sudden direction changes. The momentum value is a hyper-parameter that can be adjusted for different tasks and models. Normally, values between 0.9 and 0.99 are used (e.g., line 6). The training process starts by running the batch training data through the network. Next, the network output is compared to the expected and known outputs of the loss function. The gradient descent algorithm, in combination with the backpropagation method, allows for the calculation of the gradients of the loss function in relation to the network weights, which are then used to update the network weights in order to minimise the loss function. Backpropagation is repeated for each training epoch until the maximum number of epochs is reached, (e.g., lines 7-14).

To avoid overfitting, it is common to test the network on a separate validation dataset. When training is complete, the network weights are stored for future use. The network is now ready to be used to make predictions on new data (e.g., lines 15-16).

On the testing data, the model predicts the classes while being evaluated by comparing the predicted labels to the true labels and calculating the model's performance metrics (accuracy and confusion matrix), (e.g., lines 18-20).

In order to identify the optimal combination of variables, it is necessary to test a range of network configurations and adjust the network's parameters, such as the learning rate, the batch size, and the number of epochs.

The methodology applied to the EmoReact dataset was equally applied to the FER2013 and CAFE: i) Transfer learning - by using a pre-trained network with ImageNet; ii) Data Augmentation - applying a horizontal flip, a vertical flip, and a rotation, to the dataset images. Fig. 5.4 and Fig. 5.5 shows the global pipeline.

Algorithm 1: CNN - Image Classification

```
1: procedure ImagesClassification
2:     Load the dataset
3:     Process input images (e.g. resize, data augmentation)
4:     Initialize the CNN model with the chosen architecture.
5:     Initialise weights of the CNN network
6:     Define a optimizer and a learning rate scheduling
7:     for each − epoch, ... do
8:         Shuffle training data
9:         for each − batch, ... do
10:            Perform forward propagation through the network
11:            Calculate error (e.g. using softmax loss function)
12:            Perform backpropagation to calculate gradients
13:            Update weights of the network (e.g. using gradient descent)
14:        end for
15:        Calculate model performance on validation set
16:        Save model weights with best validation performance
17:    end for
18:    Test the model with testing data
19:    Calculate classification accuracy
20:    return Predicted class labels of test data
21: end procedure
```

The best result, in terms of accuracy, was obtained with the CAFE dataset and ConvNeXt network, with a maximum accuracy of 85.92%. Table 5.3 shows the parameters corresponding to that accuracy result. Additionally, the training was conducted three times to get maximum and average results (accuracy).

All of the results and variations of the referred parameters (Batch Size, Learning Rate, Image Size, No. Epochs and Data Augmentation) may be seen in the Appendix 7.1.

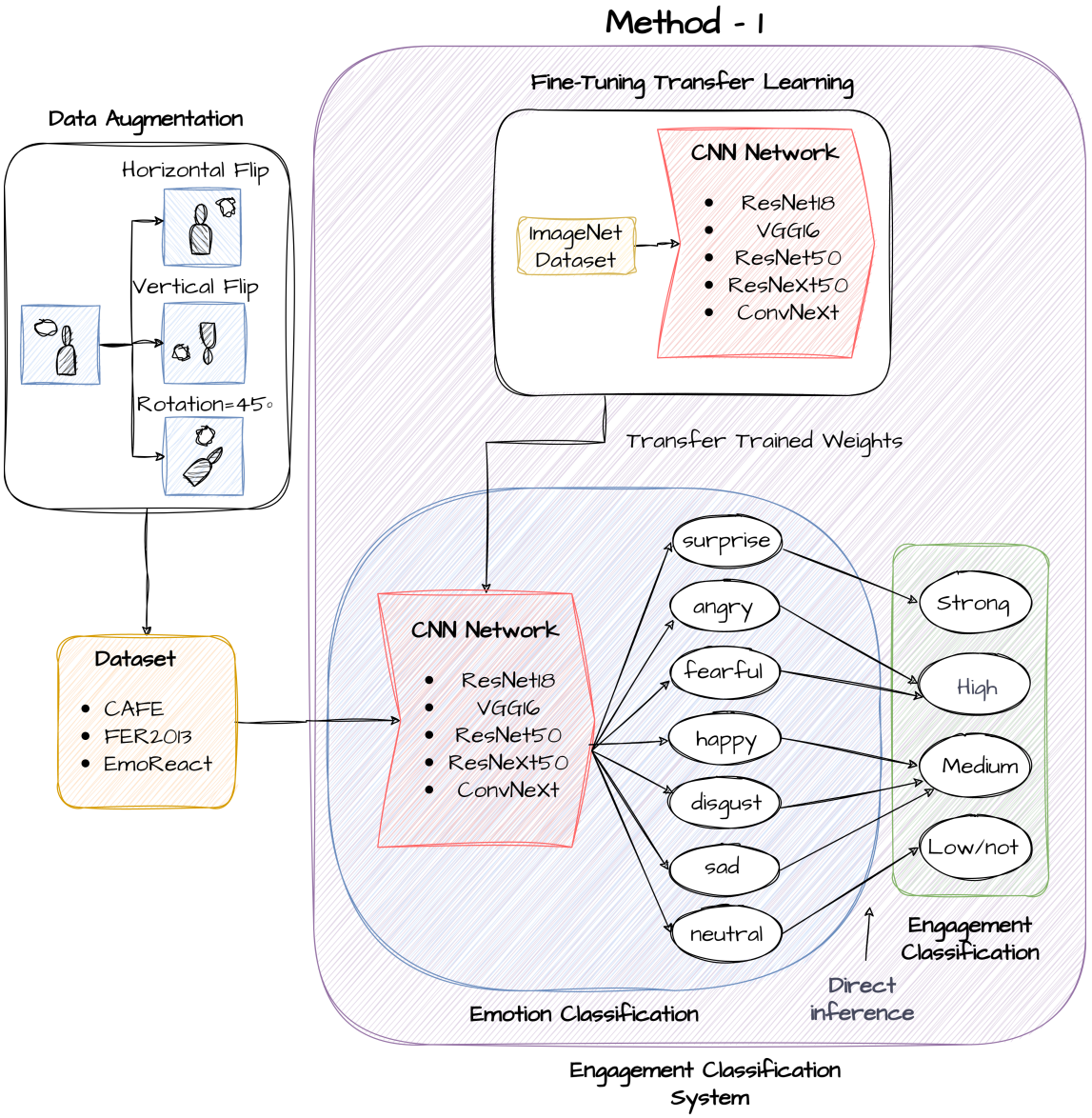

Figure 5.4: Engagement Classification: Method-1

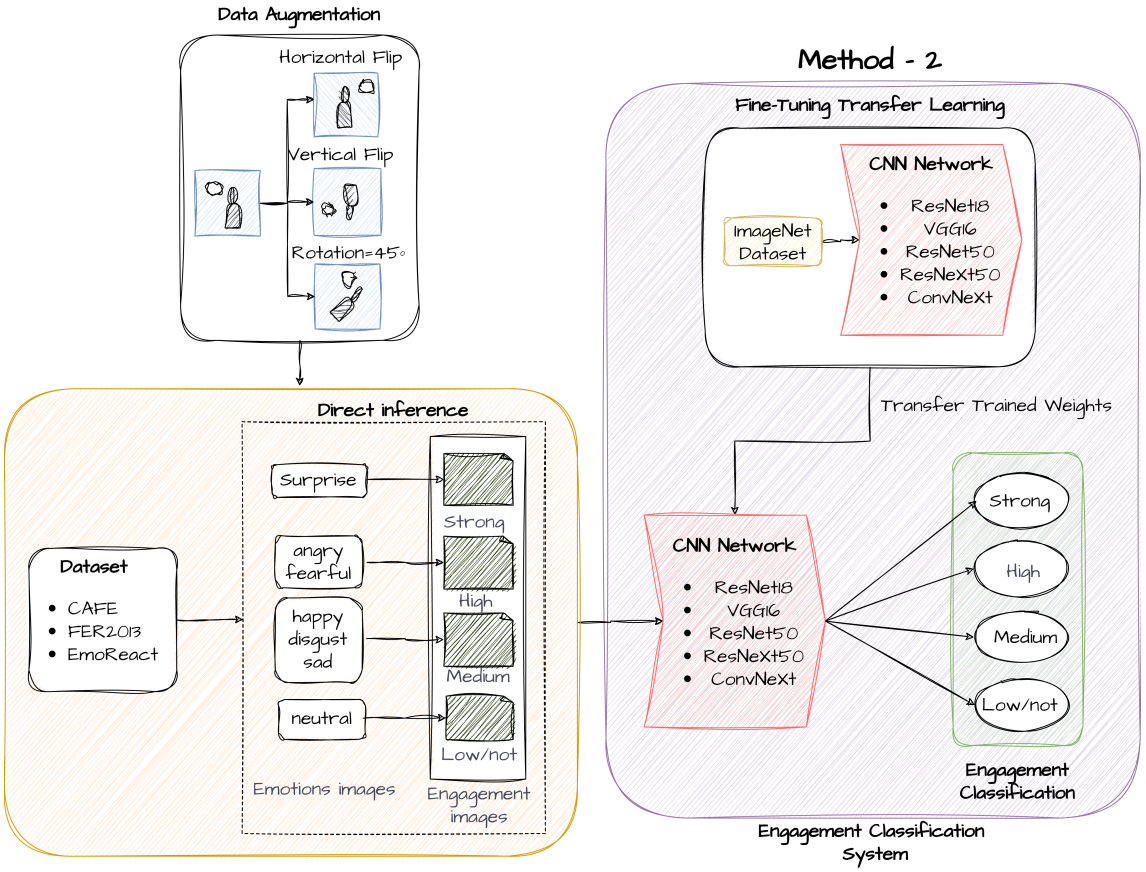

Figure 5.5: Engagement Classification: Method-2

Table 5.3: Parameters for our best network.

Network	ConvNeXt
Dataset	CAFE dataset
Batch Size	64
Learning Rate	0,0001 Adam
Image size	128×128
No. Epochs	50
Pre-trained	True
Data Augmentation	Horizontal Flip; Vertical Flip; Rotation=45^{o}

5.3 Engagement Classification Methods

This section describes two proposed approaches for engagement classification based on emotion recognition. Emotions can be used to determine the level of engagement. This affective model serves as the basis for our classification of engagement levels, and Table 5.4 shows which level of engagement correspond to each of the seven emotions that our

model classifies.

It is possible to observe in Table 5.4 that the set of 7 emotions do not encompass the entire spectrum of emotions covered by the affective model (see. Fig. 3.2), nor do we have emotions for engagement levels that are low or not engaged. In this instance, we had to modify the model to accommodate the neutral emotion, not only because there was no other level available but also because the model does not predict neutral as an emotion.

Table 5.4: Relationship between engagement level and emotions based on affective model.

Level of Engagement	**Emotions**
strong	surprise
high	angry; fearful
medium	happy; disgust; sad
low/not	neutral

5.3.1 Method-1: CNN returns the emotion classes that were followed by the direct inference of engagement

In this first approach, the emotion classification best-achieving network described in Section 5.2.1 was employed, as shown in Fig. 5.4. This method accepts of children's faces as images, as well as labels corresponding to each emotion and returns emotion classes. A direct inference of engagement is then performed by considering the correspondence between emotions and engagement levels as presented in Table 5.4.

The direct inference of engagement based on emotion recognition is detailed in Algoriyhm 2. This direct inference of engagement based on emotions yielded satisfactory results as it will be discussed later in this dissertation.

5.3.2 Method-2: CNN returns the engagement classes

In this second method, before training the network, the affective model (see Table 3.2) was applied to determine the direct relationship that exists between the levels of engagement and the emotion, as shown in Fig. 5.5.

This means that the input from the network consists of images of children's faces with engagement level labels. The network used is the same as in Method-1 and has the same parameters as Table 5.3, but it only returns four engagement classes rather than seven emotions. This did not achieve such satisfactory results compared to Method-1, as it will be discussed later in this dissertation.

Algorithm 2: Direct Inference of engagement

```
for x, in predicted ... do
    if x == angry or x == fearful then
        x = high
        continue
    end if
    if x == happy or x == disgust or x == sad then
        x = medium
        continue
    end if
    if x == neutral then
        x = low/not
        continue
    end if
    if x == surprise then
        x = strong
        continue
    end if
```

5.4 NAO Robot Activity

The designed robot-child activity has two main goals: to engage the child and to promote a calming and relaxing experience. Thus, the activity was separated into these two primary sections, which are described in more detail below and illustrated in Fig. 5.6. It was programmed using Choreographe, NAO proprietary software, as described in Chapter 4.

In order to attract the child's attention, the activity began with the NAO introducing itself and speaking in short phrases. Using the software's block programming made it possible to accomplish this; the blocks that allow the robot to talk, recognise speech, stand up, and sit down are available (pre-programmed), so we simply made use of them. Next, we introduce a piece of the "Macarena" dance, an open-source [88] recruit that is likewise programmed using blocks. To conclude the engagement part, the NAO demonstrates the ability to perform push-ups, another open-source block programme [88].

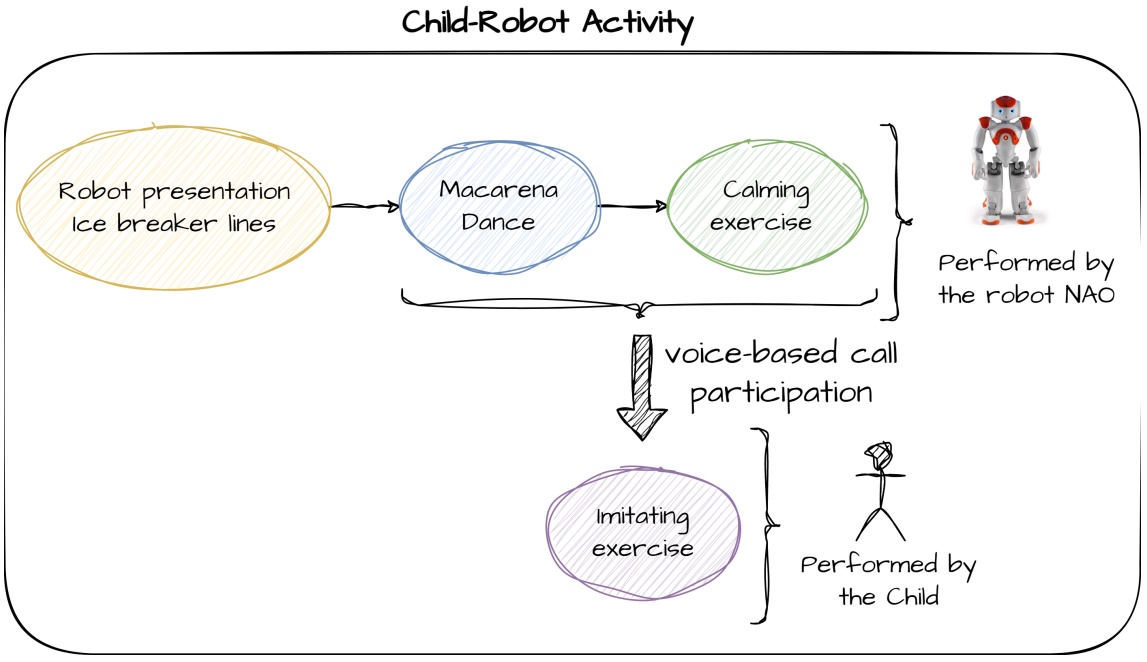

Figure 5.6: Diagram of Complete Child-Robot Activity.

At the end of the activity, the breathing control exercise was introduced (discussed in 3) to provide some relaxation and tranquillity. The breath control exercise consists of three phases: in the first phase, lift the arms for four seconds while inhaling, and in the second phase, maintain the arms in a stretched position for seven seconds while holding the breath. Finally, lowering the arms, exhale for eight seconds. These steps are repeated three times in succession, then the workout ends. Fig. 5.7 better illustrates the described phases.

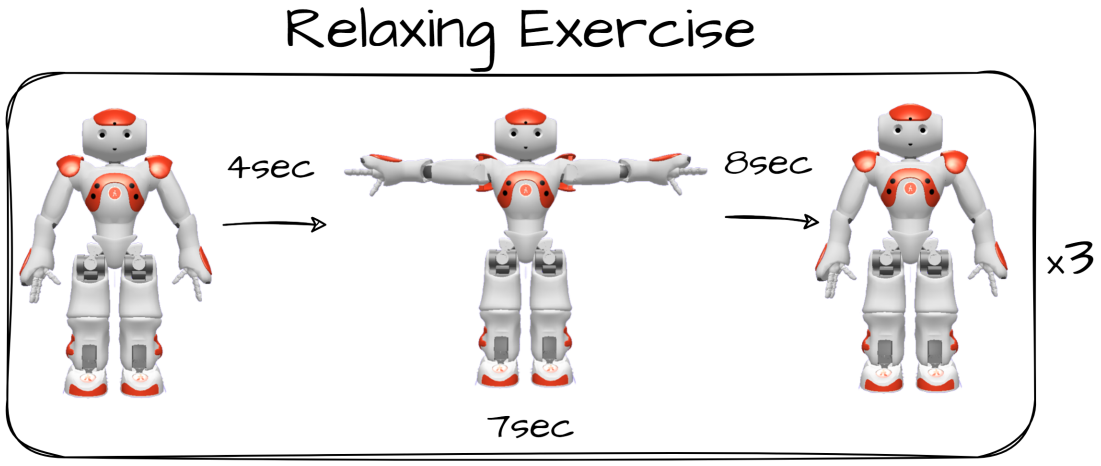

Figure 5.7: Diagram depicting the three main phases of the relaxing exercise.

It was necessary to programme the robot's limbs and joints so that it could perform the outlined movements, making the activity more dynamic while teaching the child. This programming was made easy by the software, which enabled us to physically move the robot's limbs and save the joint positions throughout the movements. This enabled the building of a timeline in which the movements that were caused in the real NAO robot were gradually saved and could be duplicated.

Primary School Demonstration

A brief demonstration of the NAO robot activity was held at the "Escola Básica no1 de Condeixa-a-Nova" during the development process. This demonstration displayed the initial part of the activity, which consisted of attracting the children's attention with ice-breaking lines and encouraging them to dance with the robot. The demonstration was conducted five times with five separate pre-school through fourth-grade classes, totalling around 200 school-aged children.

Students were particularly attentive to the robot and desired to replicate its behaviour by touching it and observing its actions. The image in Fig. 5.8 below was captured during one of these sessions.

Figure 5.8: Image captured during a demo session.

Test with Children

One of the main goals of this work was to combine the two major components created throughout the project: the robot-child interaction and the classification of engagement.

To accomplish this goal three children, two female and one male, ages 4, 7, and 6, respectively, participated in the robot activity testing. For this activity to be carried out with the children, the parents signed an "Informed Consent" form in which the entire process of data collection and use is explained.

Each testing session lasted five minutes and was performed in the presence of the parents. For the tests, the robot-child activity described in the last section was done in separate sessions while a smartphone camera on a tripod recorded them.

Each video corresponding to each session was then processed, i.e., it was converted into 500 frames per video, and a crop was applied to each frame using the approach described in Section 5.1.3. The images from each of the three videos are then applied to the two classification engagement methods outlined in Section 5.3.1 and Section 5.3.2. The obtained results are displayed in Chapter 6, where each frame was classified into one of four engagement classes according to Method-1 and Method-2.

Figure 5.9: Image captured during the testing session with child4, child6 and child7, respectively.

As shown in Fig. 5.10, two different methods were applied to classify engagement levels from videos recorded during the child-robot interaction sessions.

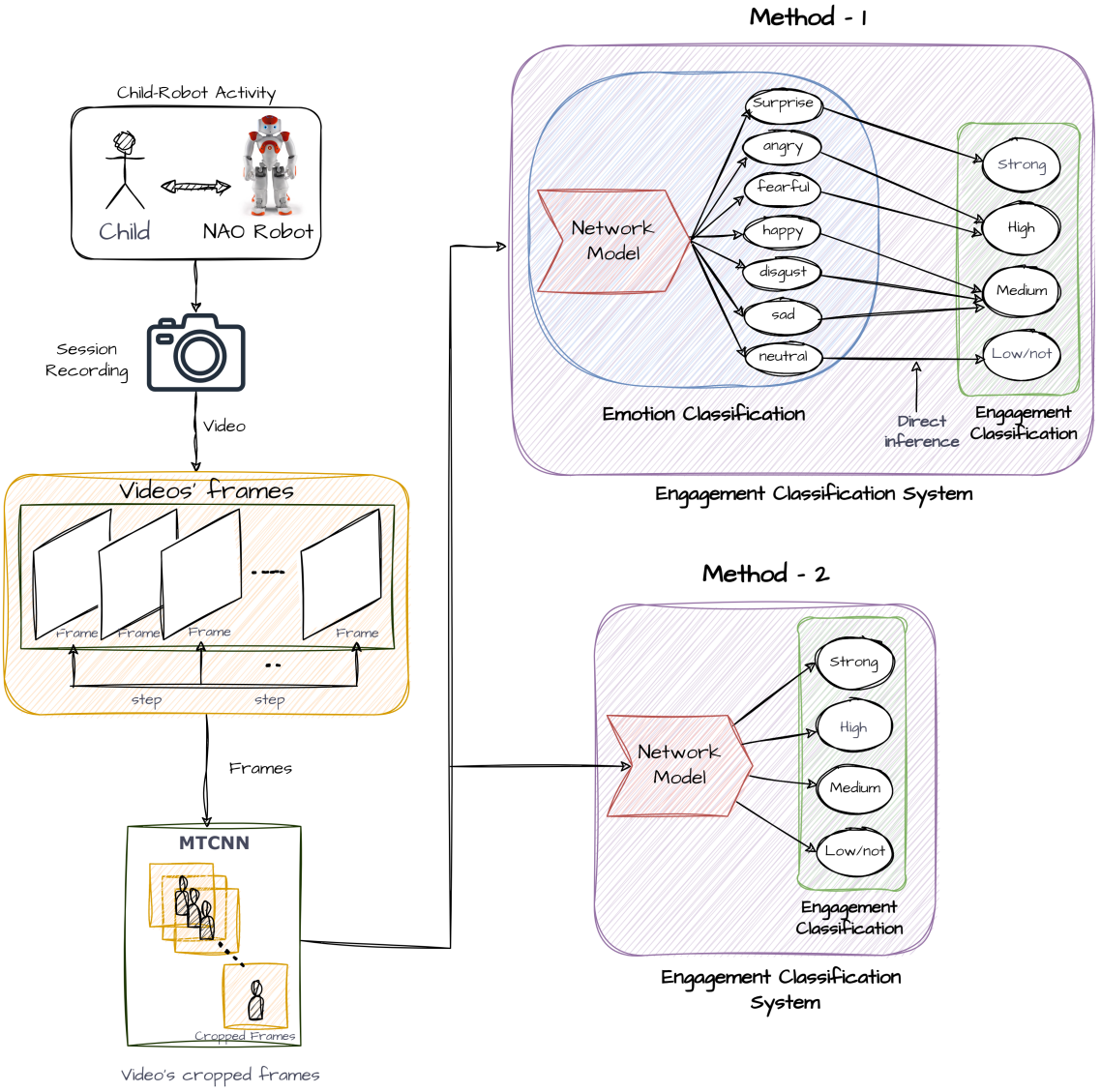

Figure 5.10: Engagement Classification Pipeline.

6 Results and Discussion

This chapter describes the experimental results that were achieved for the Emotion Recognition System with the ResNet18, VGG16, ResNet50, ResNeXt50 and ConvNeXt networks, as well as the Method-1 and Method-2 for the Engagement Classification System. The validation of the proposed child-robot activities with target participants (children from 5 to 7 years old). A discussion of results is also presented in this chapter.

6.1 Results of the Engagement Classification System

This section first presents the results obtained for the proposed Emotion Classification approach and and then presents the results for the Engagement Classification methods.

6.1.1 Emotion Classification Results

This section presents a summary of the best results obtained for emotion classification with the ResNet18, VGG16, ResNet50, ResNeXt50 and ConvNeXt networks. To achieve these results, it was necessary to adjust a number of hyper-parameters that directly affect the training performance of the network, such as batch size, learning rate, number of epochs, and image size as input to the network.

The **batch size** is the number of training examples used in one iteration and affects the training time per iteration. The larger the batch size, the longer it takes to execute a single iteration, but it also enables a more precise update of network weights. This is because the gradient is calculated from more training samples, resulting in a more accurate estimate of the direction to adjust the weights. But the gradient calculated from a large batch may be too smooth, making it hard to know how to change the weights for the optimal solution.

The **learning rate** is the number of weights updated during the training phase, also known as the "step size." The learning rate determines how quickly the network converges to an ideal solution and adapts to the problem. Higher learning rates allow for quick convergence, oscillating and diverging around the optimal solution. Lower learning rates allows slower convergence, which can get accurate results but is a time-consuming task and requires a greater number of training epochs, because the fact the algorithms take

very small steps when updating the weights.

An **epoch** is a complete pass of the entire training dataset through the algorithm. The number of epochs affects the convergence speed of the network to the optimal solution. In general, the larger the number of epochs, the closer the network approaches a good solution. However, there is a risk of overfitting, which is when the network adjusts too much to the training data and becomes less general.

The **size of the images** influences the amount of memory required to train a network as well as the duration of the training. For instance, larger images may provide more information about the objects within the image, but smaller images may not capture all relevant data. This can influence the network's accuracy and performance during classification or object detection tasks.

This variation of the parameters is presented in detail in the appendix tables 7.1 where the tests with best results are highlighted. It is possible to see that the best results were always obtained using the pre-trained network with ImageNet data, implying that the transfer learning technique was used as well as the use of data augmentation, in particular the combination of the three techniques: vertical flip, horizontal flip, and 45-degree rotation. The results show that using a larger image size improves model performance, but depending on the hyper-parameter combinations, using a larger image size is not always possible due to computational capacity. This observation holds true for batch size as well. The best results were obtained with an ADAM learning rate of 0,001, which is a very commonly used value. In each model's training, 30 or 50 epochs were used so that the network would have enough iterations to learn and adjust the weights.

The results obtained with the ResNet18, VGG16, ResNet50, ResNeXt50 and ConvNeXt networks for the FER2013, EmoReact, and CAFE datasets are compared to state-of-the-art results, as shown in the Tables 6.1- 6.5.

It is difficult to compare EmoReact dataset-derived results to the current state-of-art. This is because, the EmoReact dataset was modified to suit the emotion and engagement classification algorithms by converting videos to frames and manually selecting the most relevant frames, making it a dataset just tested in this dissertation. Additionally, and to the best of our knowledge, there is a limited amount of research that has used the EmoReact dataset, and the metrics used for these few works, such as ROC and F1-score [89], are different from those presented in this dissertation due to the fact that EmoReact is composed by videos with both visual and audio components. The results achieved with the EmoReact dataset are presented in Table 6.1.

Table 6.1: Accuracy achieved using different networks on EmoReact dataset.

Network	Acc(%)
ResNet18	43.29%
VGG16	**47.12%**
ResNet50	43.56%
ConvNeXt	46.68%

By comparing Table 6.2 and Table 6.3 it is possible to observe that the results obtained by the tested networks VGG16, ResNet50, ResNeXt50, and ConvNext, in the ranking of the seven emotions on dataset FER2013, were slightly lower than those achieved in other state-of-the-art works. However, the accuracy results obtained with the ConvNeXt network are very close to state-of-the-art results. This difference can be explained by the fact that the batch size used by us is smaller than the batch size used by [67] due to memory capacity.

Table 6.2: Accuracy achieved using different networks on FER2013 dataset.

Network	Acc(%)
VGG16	68.18%
ResNet50	66.87%
ResNeXt	68.57%
ConvNeXt	**71.81%**

Table 6.3: Acc(%) comparison with state-of-art methods using different networks on FER2013 dataset.

Work	Methods	Acc(%)
Pramerdorfer et al. [67]	ResNet50	72.40%
Pramerdorfer et al. [67]	VGG16	72.70%
Khaireddin et al. [65]	VGG	**73.28%**
Liu et al. [64]	CNN	62.44%

For the CAFE dataset, the results obtained by the ResNeXt50 and ConvNeXt networks presented in Table 6.4 clearly surpass the results reported in the state of the art (see Table 6.5).

Table 6.4: Accuracy achieved using different networks on CAFE dataset.

Network	Acc(%)
ResNeXt	82.86%
ConvNeXt	**85.92%**

Table 6.5: Acc(%) comparison with state-of-art methods using different networks on CAFE dataset.

Work	Methods	Acc(%)
Zheng et al. [63]	Shape features + SVM	**77.40%**
Witherow et al. [62]	CNN	76.03%
Lopez-Ricon [66]	CNN-AFFDEX Viola-Jones Re-Trained	44.88%
Nagpal et al. [90]	msDBM + RF	48.00%
Dias et al. [91]	CNN + Triple Loss	72.68%

Emotion Classification Best Results

The model designed for engagement classification was based in the one acquired by training the ConvNeXt network on the CAFE dataset. Table 6.6 presents the mean, maximum, and standard deviation of the training performed three times with the parameters described in Table 5.3, where a maximum of 85.92% and a mean of 84.19%. Figure 6.1 shows the training loss curve and the validation loss curve. Due to the fact that both curves declined, suggesting that the model was adapting itself to the training data, it was concluded that neither overfitting nor underfitting happened.

Table 6.6: Results with ConvNeXt Network to classify emotions.

Results	Acc(%)
max	85.92%
mean	84.19%
standard deviation	1.22%

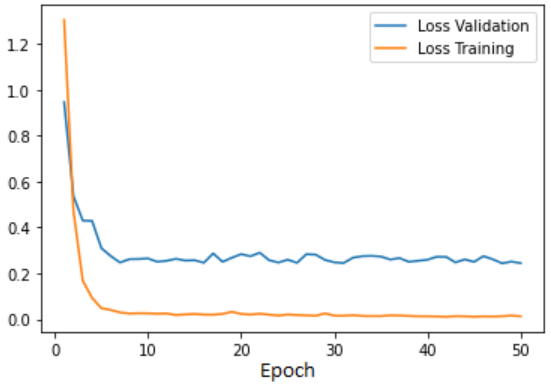

Figure 6.1: Training and validation losses curves.

Table 6.7 presents the accuracy per class and the total number of images of each class; for example, for class 0, "angry," the model predicted correctly 85% of 47 images, resulting in 40 images well predicted. Figure 6.2 presents the confusion matrix, where the main diagonal shows how many images were well classified and how many were misclassified in the rest of the cells.

Table 6.7: Classification results for 7-emotions classes.

Classes	Acc(%)	No. Images
0 (Angry)	85%	47
1 (Disgust)	81%	43
2 (Fearful)	75%	32
3 (Happy)	91%	49
4 (Neutral)	90%	52
5 (Sad)	87%	24
6 (Surprise)	86%	23

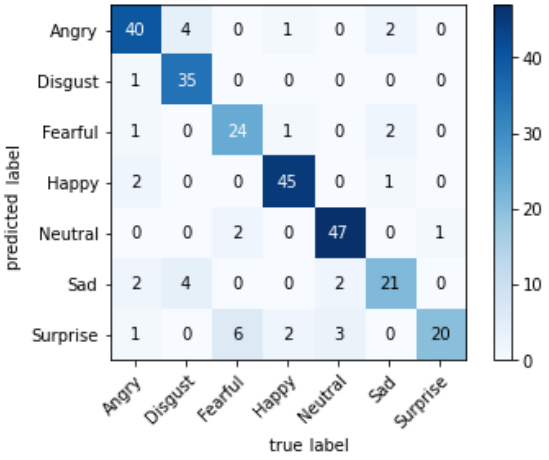

Figure 6.2: Confusion Matrix for 7-emotions classes.

6.1.2 Engagement Classification Results

This section presents the best testing results obtained for Method-1 and Method-2 for engagement classification. A fair comparison of the results presented in this dissertation with those presented in the state of the art (see Table 6.8) is not possible, because they used different proprietary datasets, and different classes of engagement (e.g. some works use only two classes: engaged and not engaged).

Table 6.8: Acc(%) comparison with state-of-art methods to engagement recognition.

Work	Dataset	Results (%)
Lin Gen et al. [69]	DAiSEE	Acc=56.20%
Woo Han et al. [19]	Own Dataset	Acc=81.44%
Omid et al. [70]	Own Dataset	Acc=72.28%
Rudovic et al. [41]	Own Dataset	ICC=59%
Hadfield et al. [40]	Own Dataset	Acc=77.11%

Method-1: CNN returns the emotion classes that were followed by the direct inference of engagement

In Method-1 the best network achieved (ConvNeXt network with CAFE datatset) for the Emotion Classification system was employed, and because no networks were trained, there are no loss curve graphs. Table 6.9 shows that Method-1 has a maximum accuracy

of 88.14% in classifying the level of engagement. Table 6.10 and the Fig. 6.3 presents the accuracy per class and the confusion matrix, respectively.

Table 6.9: Results with ConvNeXt Network to classify engagement.

Results	**Acc(%)**
max	88.14%
mean	87.89%
standard deviation	0.17

Table 6.10: Classification results for 4-engagement classes.

Classes	**Acc(%)**	**No. Images**
0 (Strong)	86%	23
1 (High)	82%	79
2 (Medium)	91%	116
3 (Low-Not)	90%	52

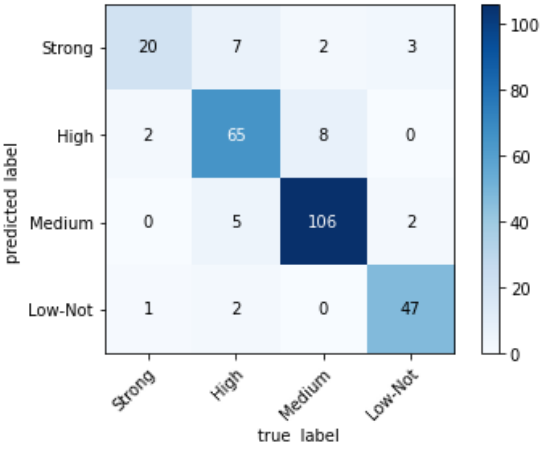

Figure 6.3: Confusion Matrix for 4-engagement classes.

Method-2: CNN returns the engagement classes

Table 6.11 shows that the Method-2 has a maximum accuracy of 85.92% in classifying the level of engagement. 6.12 and the Fig. 6.5 presents the accuracy per class and the confusion matrix, respectively. It is possible to concluded that no overfitting nor under-fitting occurred based on the model curves depicted in Fig. 6.4, which indicate the model was adjusting to the training data.

Table 6.11: Results with ConvNeXt Network to classify engagement.

Results	Acc(%)
max	85.92%
medium	84.93%
standard deviation	0.69

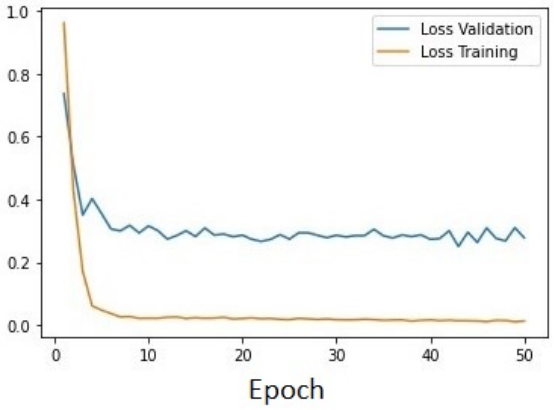

Figure 6.4: Training and validation losses curves.

Table 6.12: Classification results for 4-engagement classes.

Classes	Acc(%)	No. Images
0 (Strong)	73%	23
1 (High)	77%	79
2 (Medium)	93%	116
3 (Low-Not)	88%	52

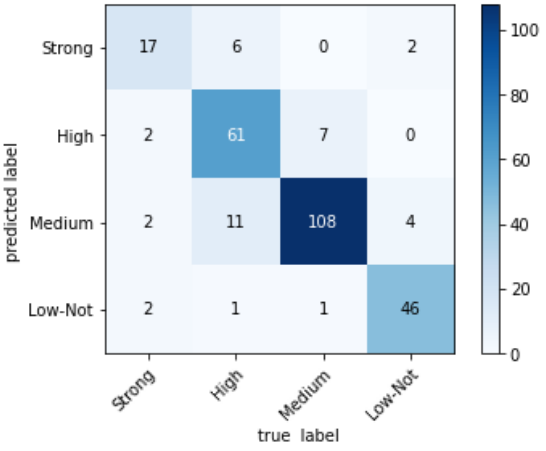

Figure 6.5: Confusion Matrix for 4-engagement classes.

Comparison between Method-1 and Method-2 Performance

Comparing the results obtained with Method-1 and Method-2 for the Engagement Classification, the percentage difference in accuracy is 2.22%, which, despite being a low value, can be significant in this type of classification problem. By analysing the two confusion matrices presented in Figures 6.2 6.3 it is possible to observe that the Method-1 model correctly predicted more images per class.

This discrepancy in accuracy, despite employing the same network with the same parameters in both Method-1 and Method-2, may be justified by the fact that the inference of engagement is made after the emotion classification, through the direct relation between emotions and engagement levels. The Method-1 model is required to be more discriminating, returning seven classifications, and to learn more specific features. In Method-2, the model returns four classes, making it easier to learn the characteristics of the face images by covering some features that correspond to more than one emotion. However, a more

discriminating method, Method-1, turns out to have positive effect on results. Another fact is that re-labeling the dataset before training the network makes the dataset unbalanced because the medium level of engagement class contains many more images than the other classes; this probably affected the performance of the model.

Results with Engagement Classification Method-1 and Method-2 on the FER2013 dataset

The methods developed for engagement classification were tested on the FER2013 dataset. The process was the same as with the CAFE dataset, except that the network used was ConvNeXt trained on the FER2013 dataset. The accuracy results in Table 6.13, reveal that although the results are not low, they do not surpass the results obtained with the CAFE dataset. Method-1 continues to perform better than Method-2, as it did for the CAFE dataset.

Table 6.13: Accuracy achieved using Method-1 and Method-2 on the FER2013 dataset.

Method	Acc(%)
Method-1	75.85%
Method-2	75.27%

6.2 Experimental Test with Children

This section presents the results obtained by applying Method-1 and Method-2 to the images extracted from the recorded videos during the child-robot activity. Table 6.14 and Table 6.15 presents the percentage of the number of occurrences of each class in each video.

Table 6.14: Results with Method-1.

Classes	Child4	Child6	Child7
Strong	0%	0.66%	0%
High	0%	10.57%	0%
Medium	98.46%	50.88%	100%
Low-Not	1.54%	37.89%	0%

Table 6.15: Results with Method-2.

Classes	Child4	Child6	Child7
Strong	0%	0.22%	0%
High	0%	0%	0%
Medium	100%	99.73%	100%
Low-Not	0%	0%	0%

Comparison between Method-1 and Method-2 Performance

Comparing the results of applying Method-1 and Method-2 to the test images extracted from the videos of the session recorded with the three children, it was anticipated that Method-1's model would perform better after the analysis of the performance with the

CAFE dataset. Due to the absence of labels in the data acquired from the children's tests, it is not possible to compare the model's predictions with the true labels.

The results are shown as a percentage of the number of occurrences of each class (levels of engagement) during the child-robot activity, on the video recorded with each child, e.g., the class "medium" occurred in 98.46% of the frames. Observing the data, it is possible to conclude that Method-1 was able to classify each video into a greater number of classes, being more discriminative, compared to Method-2. Furthermore, by examining the extracted frames from the videos, it can be seen that child4 and child7 maintain a consistent facial expression throughout the session, whereas child6 exhibits more facial expression variation, providing some validation for our model.

7 Conclusion

The main goals of this dissertation were to design and implement a child-robot activity that would engage, interact, and teach a calming exercise and to develop an engagement classification system based on the recognition of emotions through facial expressions that could be used to analyse child-robot activity during offline and online processes.

EmoReact, the FER2013, and the CAFE datasets were used to conduct an analysis to determine the accuracy performance of the emotion and engagement classifications measured in accuracy.

In terms of the accuracy of emotion classification, the results reveal that proposed model based on a ConvNext architecture classified emotions with an accuracy of 85,92%, achieving the best performance with the CAFE dataset and surpassing the results obtained for the EmoReact and FER2013 datasets.

The engagement classification system tested with the CAFE dataset performs effectively when compared to previously published works, with an accuracy of 88,14% and 85,92% for Method-1 and Method-2 respectively. Although, a fair comparison with other methods of the state of the art is debatable, because the number of classes and the datasets used in those works are different.

The application of the engagement classification system to videos recorded with children has also shown positive outcomes and promising results for application in offline analysis. Further work is required, in particular the evaluation of our results by professional therapists that can truly assess if the automatic recognition of engagement is being carried out in an effective manner.

It is also important to emphasise that one of the limitations of this work is that the datasets that were used do not include emotions that correspond to the levels of low engagement and not engaged. Additionally, the affective model that relates levels of engagement to emotions does not predict neutral as emotion. The association of neutral emotion to a low/not engaged level was idealised by us and for that reason it is an association that can affect the correct performance in the engagement classification.

The results presented, despite being promising, have great potential for improvement, for example by applying other computer vision techniques, testing adaptations to the

chosen network, or even testing other networks (e.g. temporal analysis), and exploring new datasets with different classes of emotions and features.

The main objective proposed in the introduction was successfully achieved. The emotion and engagement classifications were trained, tested, and applied in an offline real child-robot activity.

7.1 Future Work

There are a number of different strategies that, if implemented, might contribute to the improvement of the work developed in this dissertation.

Classification engagement based on other features can improve the results. The direction of the kid's gaze, the kid's body posture, and speech recognition are features that used with in combination with the emotion classification could improve the results.

There are additional machine learning approaches that can be researched and analysed, as well as other networks and adjustments to the one used. The performance of results might also be improved by exploring and building new datasets that fits the requirements of this work was explored with the help of psychology therapists.

Real-time engagement classification is also envisaged to allow a re-adaptation of the robot behaviour to the child during the interaction activity. The work can be directed towards the development of reinforcement learning algorithms so that the robot can adjust to the child by capturing his attention or maintaining his level of engagement through the feedback (reward) received by the engagement classification.

References

[1] Khawlah Altuwairqi, Salma Kammoun Jarraya, Arwa Allinjawi, and Mohamed Hammami. A new emotion–based affective model to detect student's engagement. *Journal of King Saud University - Computer and Information Sciences*, 33:99–109, 1 2021.

[2] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. 11 2015.

[3] Hugo Rafael Mendes Luís. Deep learning based human activity recognition: A real-time perspective. MSc Dissertation, October 2020.

[4] Mayank Mishra. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939, convolutional neural networks, explained - towards data science, 8 2020.

[5] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 12 2021.

[6] https://www.youtube.com/watch?v=3gyeDlZqWkot=546s. Transfer learning, 1 2023.

[7] Alexander Libin and Elena Libin. Robots who care: Robotic psychology and robotherapy approach. 2005.

[8] Alexander V. Libin and Elena V. Libin. Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robotherapy approach. *Proceedings of the IEEE*, 92:1789–1803, 2004.

[9] Luke J. Wood, Abolfazl Zaraki, Ben Robins, and Kerstin Dautenhahn. Developing kaspar: A humanoid robot for children with autism. *International Journal of Social Robotics*, 13:491–508, 6 2021.

[10] Anas Tahir and Uvais Qidwai. Humanoid robots for children with autism: Experiences with robosapien and nao view project humanoid robots for children with autism: Experiences with robosapien and nao. 2016.

[11] Melissa L. Finucane. Emotion, affect, and risk communication with older adults: Challenges and opportunities. *Journal of Risk Research*, 11:983–997, 2008.

[12] Mehdi Khamassi, George Velentzas, Theodore Tsitsimis, and Costas Tzafestas. Robot Fast Adaptation to Changes in Human Engagement During Simulated Dynamic Social Interaction With Active Exploration in Parameterized Reinforcement Learning. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):881–893, December 2018.

[13] A.V. Libin and E.V. Libin. Person-robot interactions from the robopsychologists' point of view: the robotic psychology and robotherapy approach. *Proceedings of the IEEE*, 92(11):1789–1803, 2004.

[14] Tony Belpaeme, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuayahuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayova, Georgios Athanasopoulos, V. Enescu, Rosemarijn Looije, Mark Neerincx, Yiannis Demiris, Raquel Ros, Aryel Beck, Lola Cañamero, Antoine Hiolle, Matthew Lewis, Ilaria Baroni, and Remi Humbert. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1:33–53, 12 2012.

[15] Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. 3 2013.

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks.

[17] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. 8 2017.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks.

[19] Woo Han Yun, Dongjin Lee, Chankyu Park, Jaehong Kim, and Junmo Kim. Automatic recognition of children engagement from facial video using convolutional neural networks. *IEEE Transactions on Affective Computing*, 11:696–707, 2 2020.

[20] Vincent Christlein, Lukas Spranger, Mathias Seuret, Anguelos Nicolaou, Pavel Král, and Andreas Maier. Deep generalized max pooling. 8 2019.

[21] Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Comparison of methods generalizing max- and average-pooling. 3 2021.

[22] Yin Xia, T Tony Cai, and Wenguang Sun. Gap: A general framework for information pooling in two-sample sparse inference.

[23] Gerard Jacques van Wyk and Anna Sergeevna Bosman. Evolutionary neural architecture search for image restoration. 12 2018.

[24] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Has¸im Has¸im Sak. Convolutional, long short-term memory, fully connected deep neural networks.

[25] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines.

[26] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. 9 2021.

[27] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). 6 2016.

[28] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[30] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth, 2016.

[31] Zhifei Zhang. Derivation of backpropagation in convolutional neural network (cnn). *University of Tennessee, Knoxville, TN*, 22:23, 2016.

[32] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2019.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[37] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[38] Chris Lytridis, Christos Bazinas, George A. Papakostas, and Vassilis Kaburlasos. On measuring engagement level during child-robot interaction in education. *Advances in Intelligent Systems and Computing*, 1023:3–13, 2020.

[39] A L David, Daniel David, Silviu-Andrei Matu, and Oana Alexandra David. Robot-based psychotherapy: Concepts development, state of the art, and new directions. *International Journal of Cognitive Therapy*, 7:192–210, 2014.

[40] Jack Hadfield, Georgia Chalvatzaki, Petros Koutras, Mehdi Khamassi, Costas S Tzafestas, and Petros Maragos. A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task. 2 2018.

[41] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. 2018.

[42] Hifza Javed, Won Hyong Lee, and Chung Hyuk Park. Toward an automated mea-

sure of social engagement for children with autism spectrum disorder—a personalized computational modeling approach. *Frontiers in Robotics and AI*, 7, 2 2020.

[43] Yongli Feng, Qingxuan Jia, Ming Chu, and Wei Wei. Engagement evaluation for autism intervention by robots based on dynamic bayesian network and expert elicitation. *IEEE Access*, 5:19494–19504, 2017.

[44] Subin Solomen and Pravin Aaron. 2(2): 237-241 breathing techniques-a review. *237 International Journal of Physical Education, Sports and Health*, 2:237–241, 2015.

[45] Ravinder Jerath, John W. Edry, Vernon A. Barnes, and Vandna Jerath. Physiology of long pranayamic breathing: Neural respiratory elements may provide a mechanism that explains how slow deep breathing shifts the autonomic nervous system. *Medical Hypotheses*, 67:566–571, 2006.

[46] Valentin Magnon, Frédéric Dutheil, and Guillaume T. Vallet. Benefits from one session of deep and slow breathing on vagal tone and anxiety in young and older adults. *Scientific Reports*, 11, 12 2021.

[47] Marc A. Russo, Danielle M. Santarelli, and Dean O'Rourke. The physiological effects of slow breathing in the healthy human. *Breathe*, 13:298–309, 12 2017.

[48] Heather Mason, Matteo Vandoni, Giacomo Debarbieri, Erwan Codrons, Veena Ugargol, and Luciano Bernardi. Cardiovascular and respiratory effect of yogic slow breathing in the yoga beginner: What is the best approach? *Evidence-based Complementary and Alternative Medicine*, 2013, 2013.

[49] Patrick R. Steffen, Derek Bartlett, Rachel Marie Channell, Katelyn Jackman, Mikel Cressman, John Bills, and Meredith Pescatello. Integrating breathing techniques into psychotherapy to improve hrv: Which approach is best? *Frontiers in Psychology*, 12, 2 2021.

[50] Jaruwan Vierra, Orachorn Boonla, and Piyapong Prasertsri. Effects of sleep deprivation and 4-7-8 breathing control on heart rate variability, blood pressure, blood glucose, and endothelial function in healthy young adults. *Physiological Reports*, 10, 7 2022.

[51] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166:140–164, 8 2005.

[52] Heather L. O'Brien and Elaine G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59:938–955, 4 2008.

[53] Dafni Anagnostopoulou, Niki Efthymiou, Christina Papailiou, and Petros Maragos. Engagement estimation during child robot interaction using deep convolutional networks focusing on asd children. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3641–3647, 2021.

[54] Francesca D'Errico, Marinella Paciello, and Luca Cerniglia. When emotions enhance students' engagement in e-learning processes. *Journal of e-Learning and Knowledge Society*, 12(4), September 2016.

[55] David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219 – 235, 1985. Cited by: 3315.

[56] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86 – 98, 2014. Cited by: 326.

[57] Michelle S. M. Yik, James A. Russell, and Lisa Feldman Barrett. Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology*, 77(3):600 – 619, 1999. Cited by: 313.

[58] Minsu Jang, Cheonshu Park, Hyun-Seung Yang, Jae-Hong Kim, Young-Jo Cho, Dong-Wook Lee, Hye-Kyung Cho, Young-Ae Kim, Kyoungwha Chae, and Byeong-Kyu Ahn. Building an automated engagement recognizer based on video analysis. page 182–183, 2014.

[59] Nancy A. Remington, Leandre R. Fabrigar, and Penny S. Visser. Reexamining the circumplex model of affect. *Journal of Personality and Social Psychology*, 79(2):286 – 300, 2000. Cited by: 208.

[60] Phillip C Schlechty. *Engaging students: The next level of working on the work*. John Wiley & Sons, 2011.

[61] James A Russell. Emotion concepts. *Handbook of emotions*, pages 491–503, 2000.

[62] Megan Witherow, Manar D. Samad, and Khan M. Iftekharuddin. Transfer learning approach to multiclass classification of child facial expressions. page 34, 9 2019.

[63] Zhi Zheng, Xingliang Li, Jaclyn Barnes, Chung-Hyuk Park, and Myounghoon Jeon. Facial expression recognition for children: Can existing methods tuned for adults be adopted for children? In Masaaki Kurosu, editor, *Human-Computer Interaction. Recognition and Interaction Technologies*, pages 201–211, Cham, 2019. Springer International Publishing.

[64] Kuang Liu, Mingmin Zhang, and Zhigeng Pan. Facial expression recognition with cnn ensemble. pages 163–166, 2016.

[65] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013.

[66] Alejandro Lopez-Rincon. Emotion recognition using facial expressions in children using the nao robot. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 146–153, 2019.

[67] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art. *CoRR*, abs/1612.02903, 2016.

[68] Zhenjie Song. Facial expression emotion recognition model integrating philosophy and machine learning theory. *Frontiers in Psychology*, 12, 2021.

[69] Lin Geng, Min Xu, Zeqiang Wei, and Xiuzhuang Zhou. Learning deep spatiotemporal feature for engagement recognition of online courses.

[70] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cecile Paris. Automatic recognition of student engagement using deep learning and facial expression. 2 2018.

[71] Charlotte E R A N D Senft Emmanuel A N D Belpaeme Tony Lemaignan Séverin and Edmunds. The pinsoro dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PLOS ONE*, 13:1–19, 2 2018.

[72] Kirsten A. Dalrymple, Jesse Gomez, and Brad Duchaine. The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set. *PLoS ONE*, 8, 11 2013.

[73] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, 5, 2014.

[74] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013.

[75] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis Philippe Morency. Emo react: A multimodal approach and dataset for recognizing emotional responses in children. *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 137–144, 10 2016.

[76] Hennie Brugman and Albert Russel. Annotating multimedia / multi-modal resources with elan annotating multi-media / multi-modal resources with elan. 2009.

[77] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6:3–5, 1 2011.

[78] K. Krippendor. *Content analysis: An introduction to its methodology.* 2004.

[79] Kerstin Limbrecht-Ecklundt, Holger Hoffmann, Andreas Scheck, Steffen Walter, Sascha Gruss, David Hrabal, and Harald Traue. Pictures of facial affect-ulm (pfa-u): A new facs-based set of pictures for basic emotions. *Research on Emotion in Organizations*, 9:153–168, 07 2013.

[80] http://doc.aldebaran.com/1 14/software/choregraphe/choregraphe_overview.html. Choregraphe, 1 2023.

[81] https://www.python.org/. Python, 1 2023.

[82] https://pytorch.org/. Pytorch, 1 2023.

[83] https://code.visualstudio.com/. Vscode, 1 2023.

[84] https://developer.nvidia.com/cuda zone. Cuda, 1 2023.

[85] https://developer.nvidia.com/cudnn. cudnn, 1 2023.

[86] https://www.nvidia.com/en-eu/geforce/graphics-cards/30-series/rtx 3060-3060ti/. Nvidia geforce rtx 3060, 1 2023.

[87] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 10 2016.

[88] https://funlab.nd.edu/the-nao-base/special movements/. Macarena dance, 1 2023.

[89] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. volume Vol. 4304, pages 1015–1021, 01 2006.

[90] S. Nagpal, M. Singh, Mayank Vatsa, Richa Singh, and A. Noore. Expression classification in children using mean supervised deep boltzmann machine. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June:236 – 245, 2019.

[91] William Dias, Fernanda Andaló, Rafael Padilha, Gabriel Bertocco, Waldir Almeida, Paula Costa, and Anderson Rocha. Cross-dataset emotion recognition from facial expressions through convolutional neural networks. *Journal of Visual Communication and Image Representation*, 82, 1 2022.

A Results Tables

A.1 Results with Resnet18

Table A.1: Results with ResNet18 - EmoReact dataset and batch size 64.

Classes [Acc(%)]	**#1**	**#2**	**#3**	**#4**	**#5**
0	43%	47%	39%	44%	37%
1	29%	27%	17%	23%	26%
2	25%	16%	16%	22%	17%
3	1%	1%	8%	1%	0%
4	6%	12%	7%	21%	2%
5	61%	61%	65%	59%	73%
6	10%	9%	17%	9%	13%
Img Size	32x32	32x32	48x48	48x48	48x48
LRate	0,01	0,01	0,0001	0,01	0,0001
Pre-Trained	True	True	False	False	True
Epochs	10	15	10	15	15
TrainAcc	**61,96%**	**60,92%**	**61,75%**	**60,61%**	**62,40%**
TestAcc	**41,45%**	**41,42%**	**41,17%**	**41,31%**	**43,29%**

A.2 Results with VGG16

Table A.2: Results with VGG16 - EmoReact dataset.

Classes [Acc(%)]	**#1**	**#2**	**#3**	**#4**
0	0%	41%	58%	57%
1	0%	17%	59%	60%
2	0%	9%	52%	54%
3	0%	27%	87%	87%
4	0%	17%	64%	64%
5	100%	79%	58%	56%
6	0%	12%	79%	78%
Img Size	32x32	64x64	64x64	64x64
Pre-Trained	False	True	True	True
LRate	0,01	0,0001 Adam	0,0001 Adam	0,0001 Adam
Epochs	15	50	50	50
DA	-	HF/R=45º	HF/R=45º	HF/VF/R=45º
Batch Size	64	32	64	32
TrainAcc	**37,94%**	**65,00%**	**74,00%**	**76,00%**
TestAcc	**40,29%**	**46,80%**	**46,82%**	**47,12%**

Table A.3: Results with VGG16 - FER2013 dataset, pre-trained and no. of epochs 50.

Classes [Acc(%)]	**#1**	**#2**	**#3**	**#4**
0	59%	66%	61%	59%
1	69%	68%	61%	60%
2	51%	50%	52%	50%
3	87%	85%	87%	86%
4	59%	58%	61%	61%
5	57%	55%	57%	54%
6	81%	82%	81%	80%
Img Size	64x64	64x64	64x64	64x64
LRate	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,001 SGD
DA	HF/R=45º	HF/R=45º	HF/VF/R=45º	HF/VF/R=45º
Batch Size	32	64	32	32
TrainAcc	**81,05%**	**82,00%**	**74,00%**	**70,00%**
TestAcc	**67,75%**	**67,40%**	**68,18%**	**66,77%**

A.3 Results with ResNet50

Table A.4: Results with ResNet50 - EmoReact dataset, pre-trained with no. of epochs 50.

Classes [Acc(%)]	**#1**	**#2**	**#3**
0	40%	42%	
1	31%	36%	
2	18%	16%	
3	5%	28%	
4	4%	25%	
5	69%	54%	
6	14%	10%	
Img Size	128x128	64x64	
LRate	0,0001 Adam	0,0001 Adam	
DA	-	HF/VF/R=45º	
Batch Size	32	32	
TrainAcc	**63,80%**	**64,00%**	
TestAcc	**43,56%**	**39,44%**	

Table A.5: Results with ResNet50 - FER2013 dataset and pre-trained.

Classes [Acc(%)]	**#1**	**#2**	**#3**	**#4**
0	54%	51%	51%	50%
1	62%	54%	53%	54%
2	49%	45%	44%	46%
3	86%	86%	85%	78%
4	50%	57%	60%	55%
5	55%	54%	48%	53%
6	78%	75%	70%	80%
Img Size	128x128	64x64	64x64	64x64
LRate	0,0001	0,0001 Adam	0,0001 Adam	0,0001 Adam
Epochs	50	50	30	30
Batch Size	64	32	32	16
TrainAcc	**63,96%**	**63,33%**	**62,39%**	**62.60%**
TestAcc	**64,05%**	**63,58%**	**62,17%**	**61,66%**

Table A.6: Results with ResNet50 - FER2013 - continuation.

Classes [**Acc(%)**]	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**
0	58%	56%	54%	54%	60%	58%
1	61%	61%	66%	61%	65%	63%
2	42%	51%	50%	46%	47%	47%
3	83%	85%	85%	86%	85%	86%
4	65%	56%	66%	65%	61%	59%
5	50%	51%	52%	54%	53%	56%
6	79%	77%	84%	83%	82%	80%
Img Size	64x64	64x64	64x64	64x64	64x64	64x64
LRate	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam
Epochs	30	30	30	50	50	50
DA	HF	HF/VF	HF/VF/R=45º	HF/R=45º	HF/R=45º	HF/VF/R=45º
Batch Size	32	32	32	32	64	64
TrainAcc	**70,90%**	**68,48%**	**88,03%**	**83,95%**	**88,44%**	**78,00%**
TestAcc	**64,72%**	**64,33%**	**66,87%**	**66,67%**	**63,31%**	**66,28%**

A.4 Results with ResNeXt50

Table A.7: Results with ResNeXt50 - FER2013 dataset.

Classes (Acc(%))	**#1**	**#2**
0	60%	59%
1	64%	60%
2	50%	50%
3	87%	86%
4	66%	61%
5	57%	54%
6	80%	80%
Img Size	64x64	64x64
LRate	0,0001 Adam	0,0001 SGD
Epochs	50	50
DA	HF/VF/R=45º	HF/VF/R=45º
Batch Size	64	32
TrainAcc	**80,58%**	**76,85%**
TestAcc	**68,57%**	**66,77%**

Table A.8: Results with ResNeXt50 - CAFE dataset.

Classes (Acc(%))	**#1**
0	82%
1	69%
2	68%
3	95%
4	92%
5	79%
6	82%
Img Size	64x64
LRate	0,0001 Adam
Epochs	50
DA	HF/VF/R=45º
Batch Size	32
TrainAcc	**91,99%**
TestAcc	**82,86%**

A.5 Results with ConvNeXt

Table A.9: Results with ConvNeXt - FER2013 dataset, pre-trained, no. of epochs 50, batch size of 64.

Classes [Acc(%)]	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
0	63%	57%	61%	56%	30%	59%	63%	65%
1	66%	57%	60%	54%	65%	70%	64%	67%
2	54%	48%	50%	49%	51%	52%	52%	55%
3	88%	85%	86%	85%	86%	88%	87%	88%
4	66%	65%	63%	65%	68%	66%	66%	69%
5	61%	58%	57%	54%	60%	60%	59%	59%
6	84%	78%	82%	78%	82%	83%	83%	84%
Img Size	64x64	64x64	64x64	64x64	64x64	64x64	64x64	128x128
LRate	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam	0,0001 Adam
DA	HF/VF/ R=45º		HF	VF	R=45º	HF/VF/ R=20º	HF/VF/ R=40º	HF/VF/ R=45º
TrainAcc	**83,45%**	**67,07%**	**82,09%**	**71,65%**	**82,65%**	**88,20%**	**87,86%**	**91,08%**
TestAcc	**71,13%**	**67,21%**	**68,07%**	**66,39%**	**69,61%**	**70,05%**	**70,00%**	**71,81%**

Table A.10: Results with ConvNeXt - CAFE dataset, pre-trained, image size of 128x128, learning rate of 0,0001 Adam, data augmentation (HF/VF/R=45º) and no. of epochs 50.

Classes [Acc(%)]	**#1**	**#2**	**#3**
0	85%	87%	85%
1	69%	74%	81%
2	68%	75%	75%
3	95%	97%	91%
4	94%	92%	90%
5	79%	66%	87%
6	78%	73%	86%
Batch Size	64	64	64
TrainAcc	**93,35%**	**94,02%**	**92,53%**
TestAcc	**83,33%**	**83,70%**	**85,92%**