Tampere University

Fareeda Madeen Gouher Mohammad

# PREDICTABILITY OF MATERNAL SENSORY SIGNALS USING MACHINE LEARNING MODELS

Master's Thesis
Faculty of Information Technology and Communication Sciences
Examiners: Dr. Sari Peltonen
Dr. Mervi Vänskä
December 2024

# ABSTRACT

Fareeda Madeen Gouher Mohammad: Predictability of Maternal Sensory Signals using Machine Learning Models
Master's Thesis
Tampere University
Master of Science (Technology) in Signal Processing and Machine Learning
December 2024

---

The study of mother-infant interactions is fundamental to understanding early childhood development, as these interactions influence emotional, social, and cognitive development. Mother-infant interactions help form secure attachment bonds, advance language skills, and nurture social competencies. Traditionally, the analysis of these interactions has relied on video observations and manual coding by trained psychologists, a method that is labor-intensive and costly. Therefore, there is an urgent need for cost-efficient and dynamic approaches. This study introduces an advanced approach using Machine Learning (ML) and Deep Learning (DL) technologies to automate the analysis of mother-infant interactions, with a focus on video action recognition. We employ a novel hybrid architecture integrating Pose-Guided Motion History Analysis (PGMHA) and Transformer-Based Sequence Modeling (TBSM) to enhance action recognition in mother-infant interaction videos. A transformer-based model is used to predict tactile stimuli, Voice Activity Detection (VAD) for auditory stimuli, and Computer Vision techniques to analyze visual stimuli, focusing on gaze direction detection. The proposed model achieves 74.8% accuracy in identifying and categorizing maternal behaviors, demonstrating its efficacy in recognizing interaction patterns. This approach provides new possibilities for psychologists to efficiently analyze large volumes of mother-infant interaction data with sufficient accuracy, reducing the time and resources needed for manual coding. By automating the recognition of key interaction behaviors, it supports more precise and scalable research, potentially leading to earlier identification of parenting and dyadic interaction problems and more targeted interventions.

Keywords: Machine Learning, Motion History Images, Transformers, Predictability of Maternal Sensory Signals, Action Recognition

# USE OF AI IN THESIS

I have utilised AI tools in my thesis:

☐ No
☒ Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools:

1. SciSpace
2. QuillBot
3. Grammarly
4. Claude.ai

Purpose of using AI tools:

1. SciSpace: Utilized for collecting academic papers and generating summaries. This tool helped streamline the literature review process, enabling me to efficiently gather relevant studies and distill their findings into concise overviews. This was particularly useful in identifying gaps in the literature.
2. QuillBot: Employed for paraphrasing and enhancing the clarity of my writing. It helped refine complex sentences and ensure that my arguments were articulated clearly and effectively. The tool was essential in improving the overall readability of this thesis.
3. Grammarly: Used for grammar checking and style improvements. It assisted in identifying and correcting grammatical errors, awkward phrasing, and stylistic inconsistencies. This ensured that the thesis met high standards of academic writing.
4. Claude.ai : Used for style improvements and literature review.

Sections where AI tools were used:

1. SciSpace, Claude.ai is used for literature review in Section 3.
2. Quillbot and Grammarly are used to check for grammar, punctuation, and style issues throughout the entire document.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

# PREFACE

First and foremost, I would like to express my deepest gratitude to the Almighty for the myriad of remarkable opportunities and blessings bestowed upon me throughout this journey. None of this would have been possible without His divine will and guidance.

I am immensely grateful to my beloved parents for their unwavering support, encouragement, and love. Your steadfast faith in me has been the driving force behind my endeavors. To my brother, thank you for graciously allowing me to be our parents' favorite child. Your light-heartedness and wit have kept me grounded.

I wish to convey my heartfelt appreciation to my supervisors, Dr. Sari Peltonen and Dr. Mervi Vänskä, for granting me the invaluable opportunity to work as a research assistant on the Mother-Infant Interaction Project at the 3D Media Group. I am deeply grateful for your invaluable assistance, compassion, and understanding. Your guidance and feedback have been instrumental throughout this process.

To my cherished friends, thank you for your constant encouragement, for making these two years feel like a breeze, and for preventing homesickness from settling in. Your companionship has transformed this journey from merely bearable to truly enjoyable. Together, we've navigated the unique charms of Finland, from the mesmerizing Northern Lights to the serene beauty of the countless lakes. You've helped me embrace the Finnish way of life, teaching me the art of 'sisu' and the joy of a proper sauna session.

Finally, I would like to commend myself for persevering, even when the path seemed daunting. It takes immense courage to uproot a comfortable life, move to a new country, and start anew. I extend my gratitude to my 20-year-old self, Fareeda, for refusing to relinquish her dreams and her pursuit of knowledge. To my future self, I offer this heartfelt advice: never stop learning, and always hold fast to your dreams. As I take the next steps in my career, I carry with me not just a degree, but a treasure trove of experiences, relationships, and insights that will continue to shape my path forward.

Tampere, 5 December 2024

Fareeda Madeen Gouher Mohammad

# CONTENTS

1. INTRODUCTION ..... 1
2. THEORETICAL FUNDAMENTALS ..... 4
2.1 Video Processing ..... 4
2.1.1 Video Representation ..... 4
2.1.2 Frame Extraction ..... 5
2.1.3 Optical Flow ..... 5
2.2 Machine Learning ..... 6
2.2.1 Supervised Learning ..... 7
2.2.2 Artificial Neural Networks ..... 7
2.2.3 Convolutional Neural Networks ..... 9
2.2.4 Activation Functions: ..... 10
2.2.5 Recurrent Neural Networks ..... 11
2.2.6 Transformers ..... 12
2.2.7 Training Neural Networks ..... 15
2.3 Video Action Recognition ..... 17
2.3.1 Traditional Approaches ..... 17
2.3.2 Deep Learning Approaches ..... 18
2.3.3 3D CNNs and Beyond ..... 20
2.3.4 Transformer-Based Approaches ..... 22
2.4 Mother-Infant Interactions ..... 24
2.4.1 Importance in Child Development ..... 24
2.4.2 Key Components of Interactions ..... 25
2.4.3 Traditional Assessment Methods ..... 26
2.5 Predictability of Maternal Sensory Signals ..... 26
2.5.1 Concept and Importance ..... 26
2.5.2 Quantifying Predictability ..... 27
2.5.3 Neural Correlates of Predictability ..... 28
2.6 Domain Adaptation in Machine Learning ..... 29
2.6.1 Transfer Learning ..... 29
2.6.2 Fine-Tuning Strategies ..... 30
2.6.3 Layer-wise Fine-Tuning ..... 30
2.7 Voice Activity Detection (VAD) ..... 31
2.7.1 Feature Extraction ..... 31
2.7.2 Decision Making ..... 32

2.7.3 Challenges in VAD ..... 32
2.8 Computer Vision for Gaze Direction Detection ..... 33
2.8.1 Eye Detection ..... 33
2.8.2 Pupil Localization ..... 34
2.8.3 Gaze Estimation ..... 34
2.8.4 Challenges in Gaze Detection ..... 34
3. REVIEW OF MOTHER-INFANT INTERACTION ANALYSIS ..... 36
3.1 Video Action Recognition: Foundations and Advances ..... 36
3.2 Predictability of Maternal Sensory Signals and Its Role in Infant Development ..... 39
3.3 Machine Learning models for Analyzing Mother-Infant Interactions 40
3.4 Signal processing for mother-infant interactions ..... 42
4. METHODOLOGY ..... 45
4.1 Dataset Preparation ..... 45
4.1.1 Data Types Collected ..... 46
4.1.2 Data Acquisition Equipment and Collection Techniques ..... 46
4.1.3 Data Preprocessing ..... 48
4.1.4 Data Augmentation ..... 48
4.2 Hybrid Deep Learning Framework ..... 49
4.2.1 Pose-Guided Motion History Analysis (PGMHA) ..... 49
4.2.2 Transformer-Based Sequence Modeling (TBSM) ..... 51
4.3. Voice Activity Detection (VAD) for Auditory Stimuli Analysis ..... 52
4.3.1 Audio Preprocessing and Feature Extraction ..... 53
4.3.2 Unsupervised Clustering using Gaussian Mixture Models ..... 53
4.3.3 Cluster Analysis and Interpretation ..... 54
4.3.4 Temporal Analysis of Vocal Patterns ..... 54
4.3.5 Integration with Multimodal Analysis ..... 54
4.4 Computer Vision Techniques for Visual Stimuli Analysis ..... 54
4.4.1 Object Detection and Tracking ..... 55
4.4.2 Maternal Touch Detection ..... 55
4.4.3 Infant Gaze Estimation ..... 55
4.4.4 Joint Attention Analysis ..... 56
4.4.5 Temporal Analysis ..... 56
5. RESULTS ..... 58
5.1 Analysis of Tactile Stimuli in Mother-Infant Interactions ..... 58
5.2 Preliminary Results for VAD and Visual Stimuli Detection ..... 60
5.3 Current Limitations and Future Work ..... 62

5.4 Implications and Potential Applications..... 62
5.5 Ethical Considerations..... 62
6. CONCLUSIONS..... 63
REFERENCES..... 64

# LIST OF FIGURES

***Figure 2.1*** *Optical flow of a test video frame. ..... 6*
***Figure 2.2*** *Basic Artificial Neural Network. ..... 8*
***Figure 2.3*** *Basic Layers of a CNN. ..... 10*
***Figure 2.4*** *Different activation functions. ..... 11*
***Figure 2.5*** *Transformer Architecture sourced from [16]. ..... 13*
***Figure 2.6*** *Architecture of ViViT Transformer sourced from [19]. ..... 23*
***Figure 4.1*** *Methodological framework for maternal predictability analysis. ..... 45*
***Figure 4.2*** *Experimental Setup: Mother-Infant Interaction Recording Environment. ..... 47*
***Figure 4.3*** *Mother-Infant Dyad During Free-Play Session ..... 48*
***Figure 5.1*** *Interaction Classification Examples. ..... 59*
***Figure 5.2*** *Voice Activity Detection results showing audio waveform (top) and corresponding classification of audio segments (bottom) ..... 60*
***Figure 5.3*** *Key Interaction Points Detection in Mother-Infant Interaction using OpenPose Model. ..... 61*
***Figure 5.4*** *Skeletal Pose Tracking using Movenet Multipose Model. ..... 61*

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural Networks |
| BoVW | Bag of Visual Words |
| CNN | Convolutional Neural Network |
| DL | Deep Learning |
| DTW | Dynamic Time Warping |
| ECG | Electroencephalography |
| fMRI | Functional Magnetic Resonance Imaging |
| GMM | Gaussian Mixture Model |
| GRU | Gated recurrent Unit |
| HOF | Histograms of Optical Flow |
| HOG | Histograms of Oriented Gradients |
| I3D | Inflated 3D CNN |
| KNN | K-Nearest Neighbor |
| LPC | Linear Predictive Coding |
| LSTM | Long Short-Term Memory Networks |
| MFCCs | Mel Frequency Cepstral Coefficients |
| MHI | Motion History Images |
| ML | Machine Learning |
| MP4 | MPEG-4 Video Format |
| NLP | Natural Language Processing |
| PGMHA | Pose Guided Motion History Analysis |
| ReLU | Rectified Linear Unit |
| RL | Reinforcement Learning |
| RNN | Recurrent Neural Network |
| SVM | Support Vector Machine |
| STIP | Spatial Temporal Interest Points |
| TBSM | Transformer Based Sequence Modelling |
| VAD | Voice Activity Detection |
| VidTr | Video Transformer |
| ViViT | Video Vision Transformer |
| YOLO | You Only Look Once |
| ZCR | Zero Crossing Rate |

# 1. INTRODUCTION

Mother-infant relationship is an important cornerstone of early childhood development, influencing a variety of developmental outcomes from emotional regulation to cognitive development. These interactions have a significant impact on the child's cognitive, linguistic, motor, and socio-emotional skills [1], as well as brain structure and functioning [2]. For instance, responsive and compassionate caring during early encounters can help to develop stable attachment ties. This is critical for enhancing a child's feeling of security and emotional well-being [3]. The quality of these interactions influences the child's language learning [4], social skills [5], and general psychological health [6]. Furthermore, this relationship serves as the foundation for the child's subsequent growth, influencing a variety of areas of functioning [1, 7]. This process, defined by reciprocal interactions between the mother and infant, establishes a dynamic environment that influences the child's developmental trajectory. The cumulative evidence highlights the critical nature of early interaction in establishing the foundation for lifelong health, learning, and behavior [1,7]. The systematic review of Rocha et al. [1] emphasizes the impact of mother-infant interaction on many developmental domains during the first year of life, whereas Bale et al. [7] investigated the relevance of these early experiences in programming neurodevelopmental outcomes and possible illnesses. These findings underscore the broad importance of early mother-infant interactions for both short- and long-term developmental trajectories.

The predictability of maternal sensory signals—consistent patterns in a mother's auditory, visual, and tactile behavior, is particularly important in fostering a child's sense of security and consistency in their environment. Predictability refers to the degree to which a mother's behaviors are consistent and repetitive, allowing the infant to form expectations about their surroundings [8]. This predictability encompasses three fundamental components: tactile stimuli (physical interactions and movements), auditory stimuli (vocal exchanges and acoustic patterns), and visual stimuli (gaze direction and object manipulation). Each component contributes distinctly to the overall measure of maternal behavioral predictability. For example, a mother who regularly responds to her infant's cues with soothing vocalizations and gentle touch creates a predictable setting in which the

infant can learn emotional regulation skills [9]. Consistent and predictable maternal behavior helps infants in developing expectations about their environment, which are important for neurological, cognitive, and emotional regulation [10].

Traditionally, the assessment of mother-infant interaction and maternal predictability has relied on video observations. These observations are followed by meticulous manual coding by trained psychologists. This process is not only labor-intensive, and time-consuming but also prone to subjectivity and inconsistencies due to human error and varying levels of observer training [11]. While this method can provide detailed insights, it is limited in its scalability and objectivity, making it difficult to analyze large datasets and draw reliable conclusions.

Machine learning presents a promising solution to these limitations by providing objective, quantifiable measures of interaction patterns through automated multimodal data analysis. Action recognition models, designed to detect and categorize specific actions within video sequences using techniques such as convolutional neural networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based architectures [12], are particularly well-suited for this task. By accurately identifying and classifying various actions and responses, researchers can better understand the predictability of maternal sensory signals and their impact on infant and child development.

Recent advances in machine learning (ML) and deep learning (DL) have revolutionized the analysis of mother-infant interactions through automated processing of large-scale video data. These technologies enable not only sophisticated research applications but also practical clinical tools for early detection of developmental concerns and real-time feedback systems for caregivers. Such capabilities support the development of evidence-based interventions and policies for early childhood development.

Despite the promising prospects, several challenges remain in implementing machine learning models for this purpose. High-quality annotated datasets are crucial for training effective models, yet collecting and labeling such data is resource-intensive, often requiring manual annotation by domain experts [13]. Additionally, the interpretability of deep learning models is an ongoing area of research; understanding why a model makes certain predictions is vital for gaining trust and acceptance among practitioners [14]. Addressing these challenges requires interdisciplinary collaboration between computer scientists, psychologists, and healthcare professionals.

This thesis aims to leverage advanced machine learning techniques to elucidate the complex dynamics of mother-infant interactions and their impact on early childhood de-

velopment. Our research objectives are designed to address both the technical challenges of processing multifaceted interaction data and the broader goal of enhancing our understanding of crucial developmental processes.

The specific objectives of this thesis are:

1. To develop and design methods for quantifying maternal sensory predictability in mother-infant interactions using advanced machine learning techniques.

2. To create and evaluate a novel Hybrid Deep Learning Framework for recognizing and analyzing mother-infant interactions in video data focusing particularly on tactile stimuli recognition as a crucial first step toward comprehensive predictability analysis.

3. To establish groundwork for calculating the entropy rate of maternal sensory signals, thereby contributing to the understanding of mother-infant relationship.

To address these objectives, a novel approach is presented to enhance the analysis of mother-infant interactions using advanced machine learning techniques. A Hybrid Deep Learning Framework that integrates Pose-Guided Motion History Analysis (PGMHA) and Transformer-Based Sequence Modeling (TBSM) to accurately identify and classify actions within mother-infant interaction videos is proposed. While our primary focus is on tactile stimuli recognition, our methodology also incorporates preliminary implementations of Voice Activity Detection (VAD) for analyzing auditory stimuli and computer vision techniques for visual stimuli analysis, including gaze direction detection. These components, though at different stages of development, collectively lay the groundwork for a comprehensive understanding of the predictability of maternal sensory signals. This work aims to automate the analysis process, reducing the reliance on manual coding and enabling the study of larger datasets with greater accuracy and efficiency.

The thesis is structured as follows: Chapter 2 presents the comprehensive theoretical fundamentals used in this study. Chapter 3 explores existing research on mother-infant interactions and the application of machine learning in this context. Chapter 4 describes the methodology, including the proposed Hybrid Deep Learning Framework and its components. Chapter 5 presents the results of the experiments and evaluates the performance of the proposed framework and discusses the implications of our findings, potential applications, and future research directions. This section concludes the thesis by summarizing the key contributions and insights.

# 2. THEORETICAL FUNDAMENTALS

This chapter provides a comprehensive review of key concepts in video processing, machine learning, and mother-infant interaction analysis, focusing on their application to video action recognition in early childhood development. It covers fundamental principles of video processing and machine learning, with emphasis on deep learning architectures relevant to video action recognition. The chapter explores both traditional and state-of-the-art approaches in this field. It then examines the importance of mother-infant interactions in child development, traditional assessment methods, and the concept of predictability in maternal sensory signals. Finally, it explores domain adaptation in machine learning, highlighting its relevance in transferring knowledge to the specific context of mother-infant interaction analysis. This multifaceted approach provides a robust theoretical foundation for understanding the intersection of video analysis techniques and developmental psychology.

## 2.1 Video Processing

Video processing is a fundamental component in video action recognition systems. It involves techniques to represent, analyze, and manipulate video data efficiently. This section covers three key aspects of video processing: video representation, frame extraction, and optical flow.

### 2.1.1 Video Representation

A video is essentially a sequence of images (frames) displayed in rapid succession to create the illusion of motion. In digital form, a video can be represented as a three-dimensional array of pixel values:$\mathrm{V}\ (x, y, t)$, where:

- $x$ and $y$ represent spatial coordinates
- $t$ represents the temporal dimension (frame number).

Each element of this array typically contains color information, often in RGB (Red, Green, Blue) format. For a color video with resolution $W \times H$ and $N$ frames, the full representation would be: $V(x,y,t,c) \in R^{W \times H \times N \times 3}$ , where $c$ represents the color channel (R, G, or B). The resolution of each frame (Width × Height) is $W \times H$, the number of frames is $N$ and the number "3" corresponds to the three-color channels (Red, Green, and Blue).

In practice, videos are often compressed to reduce storage requirements. Common compression formats include MPEG-4, H.264, and H.265. These formats use various techniques such as motion compensation and discrete cosine transforms to achieve high compression ratios while maintaining visual quality.

### 2.1.2 Frame Extraction

Frame extraction is the process of obtaining individual images from a video sequence. This is crucial for many video analysis tasks, including action recognition. The frame rate of a video, measured in frames per second (fps), determines the temporal resolution. Common frame rates include 24 fps (cinematic), 30 fps (standard video), and 60 fps (high-speed video).

For a video with frame rate *f*, the time *t* of the *n*th frame can be calculated as: $t = \frac{n}{f}$.

Frame extraction can be uniform (extracting frames at regular intervals) or adaptive (extracting key frames based on content change). Uniform extraction at a rate of *r* frames per second can be represented as:

$F = V\left(x, y, \left\lfloor t * \frac{f}{r} \right\rfloor\right)\{t = 0,1,2, \dots\}$, where $\lfloor\rfloor$ is the floor function.

### 2.1.3 Optical Flow

Optical flow is an important concept in video processing that represents the apparent motion of objects between consecutive frames. It is particularly important for action recognition as it captures motion information explicitly. The optical flow constraint equation is given by: $I_x \cdot u + I_y \cdot v + I_t = 0$, where:

- $I_x, I_y$ are spatial derivatives of the image intensity and $I_t$ is the temporal derivative
- $u$ and $v$ are the $x$ and $y$ components of the optical flow

This equation is based on the brightness constancy assumption, which states that the intensity of a pixel remains constant as it moves from frame to frame. One popular method for computing optical flow is the Lucas-Kanade method [15]. It assumes that the flow is constant in a local neighborhood around the pixel under consideration. The flow vector $\begin{bmatrix} u \\ v \end{bmatrix}$ is then computed by solving: $\begin{bmatrix} u \\ v \end{bmatrix} = (A^T A)^{-1} A^T b$, where:

$$A = \begin{pmatrix} I_{x1} & I_{y1} \\ I_{x2} & I_{y2} \\ \vdots & \vdots \\ I_{xn} & I_{yn} \end{pmatrix}, \; b = -\begin{pmatrix} I_{t1} \\ I_{t2} \\ \vdots \\ I_{tn} \end{pmatrix}.$$

Optical flow can be visualized using color coding, where hue represents the direction of motion and intensity represents the magnitude. In Figure 2.1, the motion of objects is represented by colors, with brighter colors indicating faster motion of an example video [77].

***Figure 2.1*** *Optical flow of a test video frame.*

## 2.2 Machine Learning

Machine Learning (ML) is a subfield of artificial intelligence that focuses on developing algorithms and statistical models that enable computer systems to improve their performance on a specific task through experience. In the context of video action recognition for mother-infant interactions, machine learning techniques, particularly deep learning approaches, play a crucial role in automatically identifying and classifying complex behaviors and interactions from video data.

This section provides an overview of key machine learning concepts and architectures relevant to our research. We begin with supervised learning, the primary paradigm used in action recognition tasks. We then explore various neural network architectures, including artificial neural networks, convolutional neural networks, recurrent neural networks, and transformers. Finally, we discuss the process of training these networks and the importance of activation functions.

### 2.2.1 Supervised Learning

Supervised learning is a machine learning approach where the algorithm learns to map input data to output labels using a labeled training dataset. In the context of video action recognition, the input data are video clips, and the output labels are action categories.

Mathematically, let D denote a dataset comprising input-output pairs defined as:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\},$$

where $x_i$ represents input features and $y_i$ represents corresponding labels. The goal is to learn a function $f$ such that: $f(x) \approx y$. The learning process involves minimizing a loss function $\mathcal{L}$ that measures the discrepancy between predicted and true labels: $\mathcal{L}(\mathrm{f(x)}, \mathrm{y})$. For classification tasks, such as action recognition, a commonly used loss function is the cross-entropy loss:

$$\mathcal{L}(f(x), y) = -\sum_{i=1}^{C} y_i \log(f(x)_i)$$

where $\mathrm{C}$ is the number of classes, $y_i$ is the true probability of class $i$ (usually 0 or 1 for hard labels), and $f(x)_i$ is the predicted probability for class $i$. The objective is to find the optimal parameters $\theta^*$ of the function $f$ that minimize the expected loss over the entire dataset:

$$\theta^* = arg\,\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i; \theta), y_i)$$

where $N$ is the number of samples in the dataset.

In practice, this optimization is often performed using gradient-based methods, which is further discussed Section 2.2.7 on training neural networks.

### 2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the human brain's architecture and functioning. They consist of interconnected nodes or neurons organized in layers. Each connection between neurons has an associated weight, which is adjusted during training to minimize prediction error.

An ANN typically comprises three types of layers, as shown in Figure 2.2:

1. Input Layer: Takes in the input features.

2. Hidden Layers: Intermediate layers where computations are performed. A network can have one or more hidden layers.

3. Output Layer: Produces the final prediction or output.

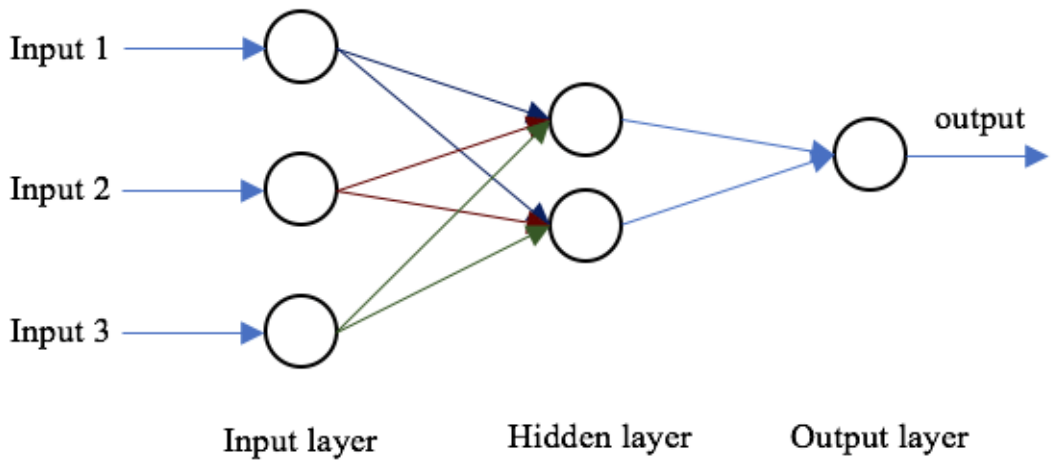


***Figure 2.2*** *Basic Artificial Neural Network.*

Consider a neural network with one hidden layer. The input to each neuron in the hidden layer is a weighted sum of the inputs from the previous layers with the addition of a bias term:

$$z_j = \sum_i w_{ij} x_i + b_j$$

where $z_j$ is the input to the j-th neuron in the hidden layer, $w_{ij}$ is the weight from the i-th input to the j-th hidden neuron, $x_i$ is the i-th input feature and $b_j$ is the bias term for the j-th neuron.

This input is passed through an activation function σ to introduce non-linearity. It is given as:

$$a_j = \sigma(z_j)$$

where σ could be a sigmoid, ReLU, or another activation function. Some of the most common activation functions are discussed in Section 2.2.4. In the output layer, the process is similar, but the activation function depends on the specific task (e.g., softmax for classification):

$$\hat{y}_k = \sigma\left(\sum_j w_{jk} a_j + b_k\right)$$

where $\hat{y}_k$ is the predicted output for class k, and $w_{jk}$ and $b_k$ are the weights and bias for the output layer.

The network is trained using backpropagation, which involves computing the gradient of the loss function $\mathcal{L}$ with respect to each weight and bias. The gradient descent algorithm updates the weights:

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - \eta \frac{\partial \mathcal{L}}{\partial w_{ij}}$$

where $\eta$ is the learning rate, $\frac{\partial \mathcal{L}}{\partial w_{ij}}$ is the partial derivative of the loss function with respect to the weight $w_{ij}$.

### 2.2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are specialized neural networks designed to process grid-like data, such as images or video frames. They use convolutional layers to automatically and adaptively learn spatial hierarchies of features. A basic CNN architecture is represented in Figure 2.3.

The key operation in CNNs is the convolution, defined in the continuous domain as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau.$$

In the discrete domain, for 2D data such as images, the convolution operation is:

$$(I * K)(i, j) = \sum_{m} \sum_{n} I(m, n) K(i - m, j - n)$$

where *I* is the input (e.g., an image) and *K* is the kernel (or filter).

CNNs typically consist of several types of layers:

1. Convolutional layers: Apply a set of learnable filters to the input.

2. Activation layers: Apply a non-linear activation function elementwise.

3. Pooling layers: Reduce the spatial dimensions of the feature maps.

4. Fully connected layers: Typically used in the final stages for classification.

The output of a convolutional layer can be computed as:

$$a_{i,j,k}^{l} = \phi\left(\sum_{m=0}^{F_h-1} \sum_{n=0}^{F_w-1} \sum_{c=0}^{C_{l-1}-1} w_{m,n,c,k}^{l} \cdot a_{i+m,j+n,c}^{l-1} + b_{k}^{l}\right)$$

where, $a_{i,j,k}^{l}$ is the activation of the neuron at position $(i, j)$ in the k-th feature map of layer *l*, $\phi$ is the activation function, $F_h$ and $F_w$ are the height and width of the filter, $C_{l-1}$ is the number of channels in the previous layer, $w_{m,n,c,k}^{l}$ is the weight at position (*m,n*) connecting the c-th channel of the previous layer to the k-th feature map of the current layer and $b_{k}^{l}$ is the bias for the k-th feature map of the current layer.

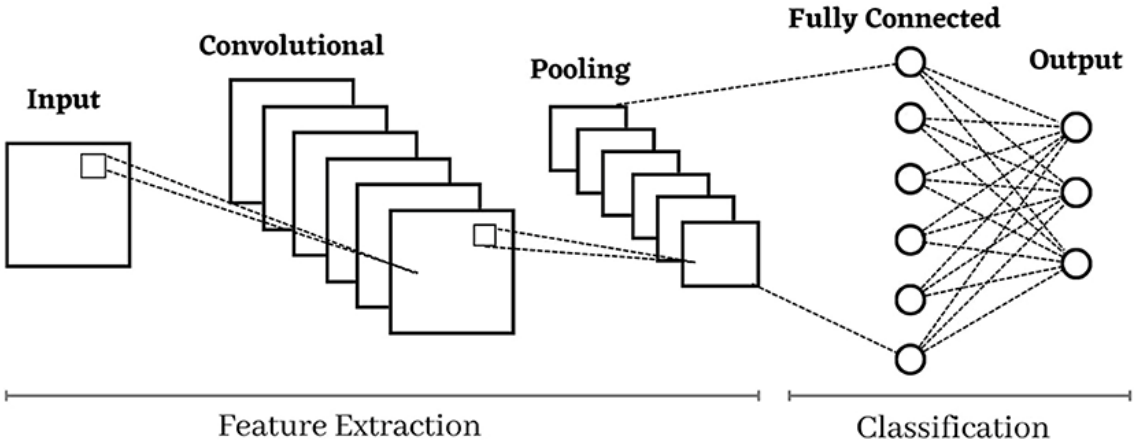


***Figure 2.3*** *Basic Layers of a CNN.*

CNNs have been highly successful in various computer vision tasks, including image classification, object detection, and action recognition in videos. Their ability to learn hierarchical representations of spatial features makes them particularly well-suited for analyzing visual data.

## 2.2.4 Activation Functions:

Activation functions introduce non-linearity into neural networks, allowing them to learn complex patterns. Different activation functions can influence the training dynamics and performance of neural networks.

Some common activation functions are:

1. Step Function: The step function is one of the simplest activation functions and is defined as: $H(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}$. The step function has limited practical applications in modern neural networks due to its binary output and non-differentiability.

2. Sigmoid: The sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, where $e$ is the base of the natural logarithm. The sigmoid function maps input values to the range (0, 1), which is useful for binary classification.

3. Hyperbolic Tangent: The hyperbolic tangent function can be defined as: $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$. This function maps input values to the range (-1, 1), providing zero-centered outputs, which can help in faster convergence during training.

4. Rectified Linear Unit (ReLU): ReLU introduces sparsity in the network by activating only positive values. It is widely used in hidden layers of deep networks. It is defined as: $\text{ReLU}(x) = \max(0, x)$.

5. Leaky ReLU: To address the "dying ReLU" problem (where neurons output zero for all inputs), the Leaky ReLU function allows a small, non-zero gradient when $x < 0$. This defined as: $LeakyReLU(x) = \begin{cases} x & if\ x > 0 \\ \alpha x & if\ x \leq 0 \end{cases}$.
6. Softmax: The softmax function is used in the output layer of multi-class classification problems. The function is defined as: $Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ ,where $x_i$ is the score for class *i*, and the denominator sums over all class scores.

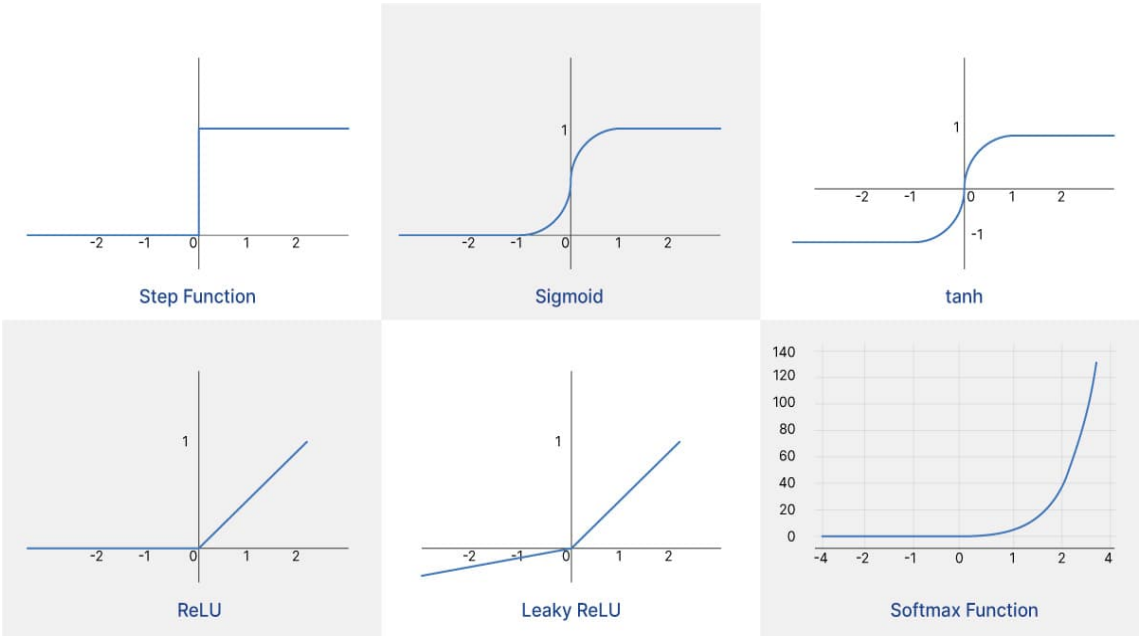


***Figure 2.4*** *Different activation functions.*

These activation functions as illustrated in Figure 2.4, serve different purposes and are chosen based on the specific requirements of the layer and the overall network architecture.

## 2.2.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed to process sequential data by maintaining an internal state (memory). This makes them particularly suitable for tasks involving time series or sequential data, such as video analysis.

The basic RNN formulation is:

$$h_t = \phi(W_{hx}x_t + W_{hh}h_{t-1} + b_h)$$

$$y_t = W_{yh}h_t + b_y$$

where, $h_t$ is the hidden state at time *t*, $x_t$ is the input at time *t*, $y_t$ is the output at time *t*, $W$ and $b$ are weight matrices and bias vectors and $\phi$ is an activation function

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are popular variants of RNNs designed to address the vanishing gradient problem in standard RNNs.

### 2.2.6 Transformers

Transformers are neural network architectures that rely entirely on an attention mechanism to determine global dependencies between input and output. They were originally proposed for Natural Language Processing (NLP) tasks but have now been successfully adapted for video analysis.

Transformer architecture offers novel prospects for analyzing mother-infant interactions. To understand their application in video analysis, in this section we explore the standard Transformer architecture and then discuss how it is modified for video data in later sections. The standard Transformer, introduced by Vaswani et al. in "Attention Is All You Need" (2017) [16], is based entirely on attention mechanisms, dispensing with recurrence and convolutions. Transformer architecture is illustrated in Figure 2.5. Its key components include:

**1. Multi-Head Attention**: The core of the Transformer is the multi-head attention mechanism. For a single attention head, the attention function is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where $Q, K$, and $V$ are query, key, and value matrices, respectively and $d_k$ is the dimension of the key vectors. Query matrix $Q$ represents the current item of interest in the sequence, used to compute attention scores with respect to all other positions. Key matrix $K$ represents all positions in the sequence, used in conjunction with the query to determine the importance or relevance of each position. Value matrix $V$ contains the actual content or information at each position, used to compute the weighted sum that forms the output of the attention mechanism.

Multi-head attention extends this by applying multiple attention functions in parallel:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

Where Concat represents concatenation of the outputs from *h* different attention heads, and $W^O$ is a learned output projection matrix. Each head is computed as:

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Here, $W_i^Q, W_i^K, and\ W_i^V$ are learned parameter matrices that project the input matrices into different subspaces for each attention head. This allows the model to jointly attend to information from different representation subspaces at different positions.

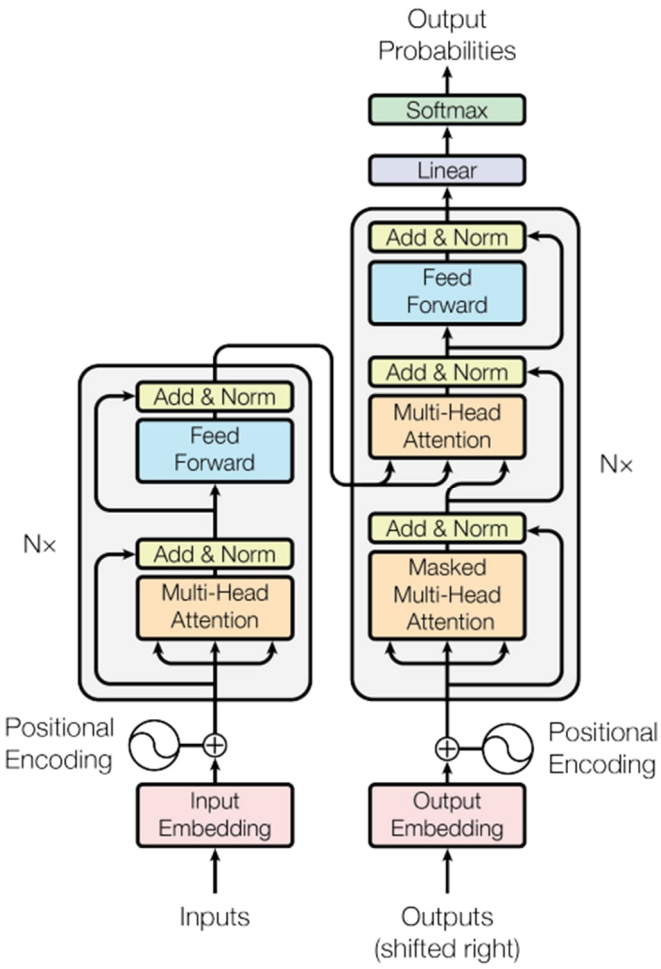


***Figure 2.5*** *Transformer Architecture sourced from [16].*

**2. Positional Encoding**: Since Transformers do not have recurrence or convolution, they need positional encoding to make use of the order of the sequence. The original publication used sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right)$$

where *pos* is the position, *i* is the dimension, and $d_{model}$ is the dimensionality of the model's internal representations (typically 512 in the original publication). This $d_{model}$ dimension is maintained throughout the model, including in embeddings and layer outputs.

**3.Feed-Forward Networks**: Each layer in the Transformer contains a fully connected feed-forward network, applied to each position separately and identically:

$$FFN(x) = \max(0, xW_1 + b_1)\, W_2 + b_2.$$

**4.Layer Normalization**: Layer normalization is applied after each sub-layer in the encoder and decoder:

$$\text{LayerNorm}(x) = \gamma \odot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

where μ and σ are the mean and standard deviation of the inputs, γ and $\beta$ are learned parameters, $\odot$ denotes element-wise multiplication and $\epsilon$ is a small constant added for numerical stability.

**Video Transformers**

Video Transformers represent an adaptation of the Transformer architecture, to the domain of video understanding. These models are specifically engineered to capture the spatiotemporal nature of video data, enabling the analysis of complex temporal dependencies and spatial relationships across frames.

Key Differences from Standard Transformers:

1. Input Representation: Instead of word embeddings, video transformers use patch embeddings from video frames.
2. Spatiotemporal Attention: Attention mechanisms are modified to capture both spatial and temporal relationships.
3. 3D Positional Encoding: Positional encodings are extended to include temporal information.

Video Transformers are primarily used for various video understanding tasks, including action recognition, temporal localization, and video captioning.

Examples of Video Transformer Architectures:

1. TimeSformer (Time-Space Transformer): Introduced by Bertasius et al. [17], TimeSformer applies divided space-time attention, processing spatial and temporal dimensions separately to reduce computational complexity while maintaining performance.
2. VidTr (Video Transformer): Proposed by Y. Zhang et al. [18], VidTr employs a hierarchical structure with multiple Transformer layers operating at different temporal scales, enabling efficient modeling of both short-term and long-range dependencies.
3. ViViT (Video Vision Transformer): Developed by Arnab et al. [19], ViViT explores various factorization strategies for spatial and temporal dimensions, including the

use of tubelet (three-dimensional spatiotemporal patches) embeddings and factorized encoder designs.

These Video Transformer architectures demonstrate the versatility and potential of applying Transformer-based models to video understanding tasks, offering powerful alternatives to traditional convolutional and recurrent approaches in computer vision.

For this project, ViViT has been selected due to its state-of-the-art performance in video action recognition tasks and its flexible architecture that effectively captures both spatial and temporal information in mother-infant interaction videos. The ViViT architecture, including its tubelet embedding approach and various factorization strategies, will be discussed in detail in Section 2.3.4, providing a comprehensive understanding of its application to our specific research context.

### 2.2.7 Training Neural Networks

Training neural networks involves adjusting the network parameters (weights and biases) to minimize a loss function that quantifies the difference between the network's predictions and the actual target values. This process is important for making the neural network learn from data and make accurate predictions. The key techniques used for training neural networks are gradient descent and backpropagation.

**1. Gradient Descent**

Gradient descent is an optimization algorithm used to minimize the loss function by iteratively updating the model parameters in the direction of the negative gradient. The update rule for gradient descent is given by:

$$\theta(n+1) = \theta(n) - \eta \nabla_\theta L(\theta(n))$$

where:

- $\theta(n)$ represents the model parameters (weights and biases) at iteration *n*,
- $\eta$ is the learning rate, which controls the size of the step taken towards the minimum of the loss function,
- $\nabla_\theta L(\theta(n))$ denotes the gradient of the loss function $L(\theta(n))$ with respect to the parameters $\theta$ at iteration *n*.

The gradient $\nabla_\theta L(\theta(n))$ is computed to indicate the direction and rate of change of the loss function concerning each parameter. The learning rate $\eta$ helps to balance the speed of convergence versus the risk of overshooting the minimum.

**2. Backpropagation**

Backpropagation is an algorithm used to compute the gradients of the loss function with respect to each parameter efficiently. It involves two main steps: the forward pass and the backward pass.

1. Forward Pass: During the forward pass, the input data is passed through the network to compute the predicted output. The activation of each neuron is calculated using the same notation as introduced in Section 2.2.2:

$$z_j = \sum_i w_{ij} x_i + b_j$$

$$a_j = \sigma(z_j)$$

$$\hat{y}_k = \sigma\left(\sum_j w_{jk} a_j + b_k\right)$$

where σ is the activation function (such as ReLU, sigmoid, etc.), and $\hat{y}_k$ is the predicted output.

2. Backward Pass: In the backward pass, the algorithm computes the gradient of the loss function with respect to each weight and bias using the chain rule of calculus. For each weight $w_{ij}$, the gradient is computed as:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}}.$$

This involves propagating the error backward through the network, from the output layer to the input layer.

**3. Optimization Algorithms:**

Advanced optimization techniques enhance the basic gradient descent approach. These include:

Stochastic Gradient Descent (SGD): Updates weights using a small batch of data, which can lead to faster convergence and better generalization.

Adam Optimizer: An adaptive method that combines momentum and adaptive learning rates:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla \mathcal{L}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)(\nabla \mathcal{L})^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where $m_t$ and $v_t$ are the first and second moment estimates, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, $\nabla\mathcal{L}$ is the gradient of the loss function, $w_t$ is the weight at time step *t,* $\eta$ is the learning rate, $\beta_1$ and $\beta_2$ are decay rates, and $\epsilon$ is a small constant to prevent division by zero.

These approaches help achieve faster convergence and more stability, leading to more effective neural network training. The Adam optimizer adapts the learning rate for each parameter, which is very useful for problems with sparse gradients or noisy data.

## 2.3 Video Action Recognition

Video action recognition is an important area of computer vision that involves identifying and classifying actions depicted in video sequences. The task has received significant attention due to its wide range of applications, including surveillance, human-computer interaction, video indexing, and autonomous systems. The goal of video action recognition is to automatically recognize and categorize the events occurring within a video. Unlike image recognition, which only deals with spatial information, video action recognition involves both spatial and temporal dynamics, making it a more complex and challenging problem.

### 2.3.1 Traditional Approaches

Traditional approaches to video action recognition primarily rely on handcrafted features combined with classical machine learning techniques. These methods involve several key steps: feature extraction, feature representation, and classification.

1. Feature Extraction: Feature extraction is the first step, where important patterns from video frames are identified. Some widely used features include:

**Spatio-Temporal Interest Points (STIP):** Extends the concept of spatial interest points in images to the temporal domain. STIPs are points in the video that exhibit significant changes over both space and time. Mathematically, these are detected using a 3D Harris corner detector, which is an extension of the 2D Harris corner detector applied to the spatial domain:

$$H = R(x, y, t) = \det(M) - k \cdot \text{trace}(M)^2$$

where $M$ is a 3D second moment matrix, and $k$ is a constant.

**Histogram of Oriented Gradients in 3D (HOG3D):** Generalizes HOG from 2D images to 3D video sequences. It captures the gradients of pixel intensities across both space and time. The gradient magnitude and orientation are computed, and histograms are formed to represent the frequency of various gradient directions.

**Optical Flow**: Captures the apparent motion of objects in consecutive frames by calculating the motion field.

2. Feature Representation: Once features are extracted, they are represented in a way that can be used for classification. Common representations include:

- Bag of Visual Words (BoVW): Quantizes local features into a finite number of clusters, forming a histogram that indicates the frequency of each "visual word" in the video. The visual vocabulary is learned using K-means clustering on the extracted features.
- Fisher Vectors: Enhance BoVW by encoding not only the frequency but also the distribution of features, using Gaussian Mixture Models (GMMs) to model feature distributions.

3. Classification: Finally, the features are classified using traditional machine learning algorithms:

- Support Vector Machines (SVM): Often used due to their effectiveness in high-dimensional spaces, SVMs find the optimal hyperplane that separates the data into different action classes.
- K-Nearest Neighbors (KNN): Classifies actions by voting among the k closest feature vectors in the training set.

Traditional methods struggle with generalization due to reliance on handcrafted features, and they lack the ability to capture complex, long-term temporal dependencies. These limitations have led to the development of deep learning-based methods that can learn features directly from raw video data.

### 2.3.2 Deep Learning Approaches

Deep learning has transformed video action recognition by enabling models to automatically learn relevant features from raw video data, eliminating the need for handcrafted

features. This section explores key deep learning methods like CNNs, RNNs, and Two-Stream Networks that have contributed significantly to this field.

**Convolutional Neural Networks (CNNs)**

CNNs are effective in capturing spatial information from individual video frames. For action recognition, CNNs typically process each frame independently, extracting spatial features such as edges, textures, and shapes. The key operations in CNNs include convolution and pooling:

1. Convolutional Layer: The convolutional layer performs a fundamental operation in CNNs, where the output feature map is computed through a convolution operation between the input and learnable filters. For a given filter k at spatial position (*i,j),* the output activation $y_{i,j}^{k}$ is computed as:

$$y_{i,j}^{k} = \sum_{m}\sum_{n} x_{i+m,j+n} \cdot w_{m,n}^{k} + b^{k}$$

where, $y_{i,j}^{k}$ is the output of the *k*-th filter at position (*i,j*), *x* is the input frame, $w_{m,n}^{k}$ represents the filter weights, and $b^{k}$ is the bias.

2. Pooling Layer: Reduces the spatial dimensions of the feature map while retaining important information, usually via max pooling.

Although CNNs excel at spatial feature extraction, they lack the ability to model temporal dependencies, which is critical for video analysis.

**Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks**

To capture temporal dependencies across video frames, RNNs, particularly LSTMs, are employed. LSTMs are a type of RNNs designed to handle long-term dependencies using memory cells and gating mechanisms:

1. Forget Gate: The forget gate $f_t$ determines the extent to which previous cell state information should be retained: $f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$, where $\sigma$ represents the sigmoid activation function, $W_f$ denotes the forget gate weights, $h_{t-1}$ is the previous hidden state, $x_t$ is the current input, and $b_f$ is the forget gate bias. It decides what information to discard from the cell state. It outputs a value between 0 and 1 for each number in the cell state, where 1 represents "keep this" and 0 represents "forget this."
2. Input Gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$.

3. Candidate Cell State: $\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

   Input gate determines which new information will be stored in the cell state. It works in conjunction with the candidate cell state ($\widetilde{C_t}$) to update the cell state.

4. Cell State Update: $C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t}$

   The cell state is updated by forgetting aspects of the previous state (via $f_t$) and adding new candidate values (via $i_t$ and $\widetilde{C_t}$).

5. Output Gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

   It controls what information from the cell state will be output as the hidden state.

6. Hidden State: $h_t = o_t \cdot \tanh(C_t)$

   Hidden State represents the output of the LSTM unit for the current time step, which can be passed to the next time step or used for predictions. These equations illustrate how LSTMs manage information flow over time, making them well-suited for video data.

**Two-Stream Networks**

Two-Stream Networks address the challenge of capturing both spatial and temporal features by employing two separate CNNs: one for RGB frames (spatial stream) and another for optical flow (temporal stream). The final prediction is obtained by fusing the outputs of these streams:

$$f_{\text{final}} = \text{f}_{\text{spatial}}(\text{x}_{\text{RGB}}) + \text{f}_{\text{temporal}}(\text{x}_{\text{optical\_flow}})$$

This approach has shown to significantly improve the performance of action recognition models by capturing complementary information from both appearance and motion.

## 2.3.3 3D CNNs and Beyond

As an extension of 2D CNNs, 3D Convolutional Neural Networks (3D CNNs) are designed to capture spatiotemporal features by applying convolutions over three-dimensional volumes, integrating both spatial and temporal information in one operation. This approach is particularly effective for modeling motion and changes over time in video data.

**3D Convolutional Neural Networks (3D CNNs)**

3D CNNs extend the standard 2D convolution operation into the temporal dimension. Instead of using 2D kernels that slide over height and width, 3D CNNs use 3D kernels that slide over height, width, and time:

1. 3D Convolution Operation: For a given position (*i,j,k*) in the spatiotemporal volume, the output activation $y_{i,j,k}^{l}$ of the l-th filter is computed as:

$$y_{i,j,k}^{l} = \sum_m \sum_n \sum_p x_{i+m,j+n,k+p} \cdot w_{m,n,p}^{l} + b^{l}$$

Here, $y_{i,j,k}^{l}$ represents the output of the *l*-th filter at position (*i,j,k*), with *x* being the input volume (video clip), $w_{m,n,p}^{l}$ the 3D filter, and $b^{l}$ the bias.

By incorporating the temporal dimension into the convolution process, 3D CNNs can simultaneously capture the spatial structure within each frame and the temporal dynamics across frames. This makes them particularly suited for action recognition tasks where understanding motion is crucial.

2. Inflated 3D CNNs (I3D)

Inflated 3D CNNs (I3D) are an advancement over standard 3D CNNs, proposed to benefit from the strengths of both 2D CNNs and 3D CNNs. I3D models are initialized by inflating pre-trained 2D CNN models into 3D. The idea is to inflate the filters and pooling kernels of 2D networks to 3D by repeating the weights across the temporal dimension:

**I3D Initialization:** I3D initialization process leverages pre-trained 2D CNN weights to initialize 3D convolutional kernels. The inflation of 2D kernels to 3D is performed through a principled transformation that can be expressed as:

$$\mathrm{w}_{\mathrm{i,j,k}}^{\mathrm{3D}} = \mathrm{w}_{\mathrm{i,j}}^{\mathrm{2D}} \times \frac{1}{\sqrt{\mathrm{k}}}$$

where $\mathrm{w}_{\mathrm{i,j,k}}^{\mathrm{3D}}$ is the inflated 3D weight, $\mathrm{w}_{\mathrm{i,j}}^{\mathrm{2D}}$ is the pre-trained 2D weight, and *k* is the temporal kernel size.

This approach leverages the knowledge captured by large 2D CNNs, such as Inception networks, while extending their capability to handle spatiotemporal data. The result is a significant boost in performance with less training time compared to training a 3D CNN from scratch.

**Beyond 3D CNNs: Hybrid Models and Multi-stream Architectures**

While 3D CNNs have proven effective, they are computationally expensive and require large amounts of labeled video data. To address these challenges, hybrid models and multi-stream architectures have been explored:

- Hybrid Models: Combine 2D CNNs for spatial feature extraction with RNNs or LSTMs for temporal modeling, aiming to reduce computational complexity while maintaining performance.

- Multi-stream Architectures: Extend the two-stream networks to incorporate additional streams, such as depth information, optical flow, or even pose estimation, capturing richer information from the video data.

3D CNNs and their variants have set a new standard in video action recognition, offering a powerful framework for understanding complex spatiotemporal patterns. However, the field continues to evolve with the exploration of more efficient and effective architectures, including hybrid and multi-stream models.

### 2.3.4 Transformer-Based Approaches

Transformer-based approaches have recently emerged as a powerful tool for video action recognition, building on the success of transformers in natural language processing and other domains. These models excel in capturing long-range dependencies and relationships within sequential data, making them well-suited for video analysis.

Transformers rely on a self-attention mechanism, which allows them to weigh the importance of different elements within a sequence dynamically. The fundamental operation in a transformer is the self-attention mechanism. This mechanism allows the model to focus on different parts of the input sequence when making predictions, capturing dependencies regardless of their distance in the sequence.

**Vision Transformers (ViT)**

The Vision Transformer (ViT) is an adaptation of the transformer architecture for image and video data. In ViTs, an image (or video frame) is divided into patches, and each patch is treated as a token like words in NLP tasks:

1. Patch Embedding: For each patch, the embedding is computed through a learned linear projection, given as:

$$\text{Patch Embedding} = W_p \cdot \text{Patch} + b_p$$

where $W_p$ is a learned projection matrix, and $b_p$ is the bias term. Each patch embedding is then processed as a sequence element.

2. Positional Encoding: Since transformers do not inherently capture the order of sequence elements, positional encodings are added to the patch embeddings to retain spatial information:

$$\text{Positional Encoding}(pos, i) = \sin\left(\frac{pos}{10000^{2i/d}}\right).$$

These encodings allow the model to understand the relative positions of patches in the image or video frame.

Transformers for Video Action Recognition

For video action recognition, transformers extend the self-attention mechanism to both spatial and temporal dimensions. This dual attention allows the model to capture both the appearance of objects in individual frames and their motion across time.

The architecture typically includes:

- **Spatiotemporal Attention Layers:** Capture dependencies across spatial and temporal dimensions.
- **Feed-Forward Networks:** Further process the attended features.
- **Layer Normalization:** Stabilizes training and improves performance.

The final output is typically aggregated to produce a prediction about the action occurring in the video.

**Video Vision Transformer**

ViViT, represents a significant advancement in transformer-based architectures for video action recognition. It builds upon the success of Vision Transformers (ViT) by adapting the architecture to effectively process spatiotemporal information in videos. Figure 2.6, represents the architecture of ViViT Transformer and how attention mechanism is specialized to handle video data.

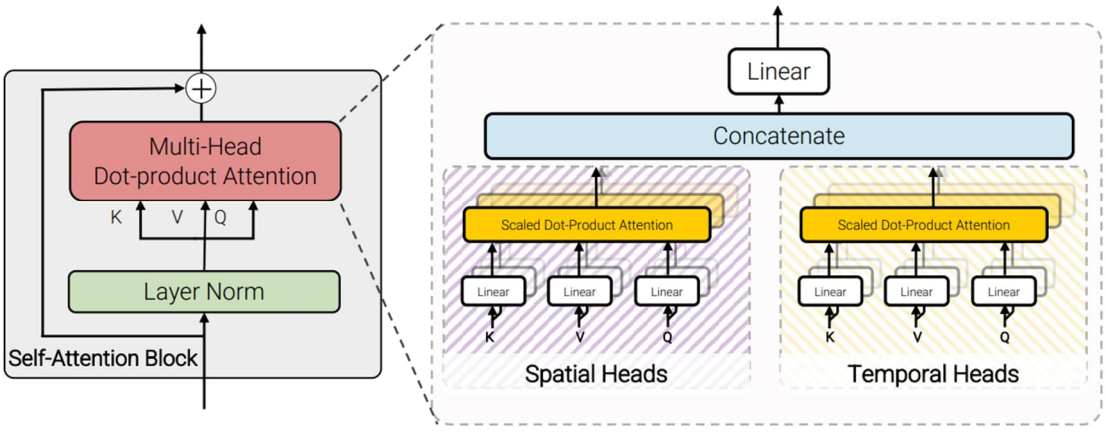


***Figure 2.6*** *Architecture of ViViT Transformer sourced from [19].*

Key Features of ViViT Architecture:

1. Tubelet Embedding: The video is divided into 3D "tubelets" (spatiotemporal patches), which are then linearly projected to create embeddings.

2. Spatiotemporal Attention: The attention mechanism is extended to operate over both spatial and temporal dimensions simultaneously. The attention equation is same as the one discussed in Section 2.2.6.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$$

where *M* is a mask that can be used to implement different attention patterns (e.g., spatial-only, temporal-only, or full spatiotemporal).

3. Factorized Encoder: Some variants of ViViT use a factorized encoder that separates spatial and temporal attention, reducing computational complexity.

ViViT has demonstrated superior performance on benchmark datasets for action recognition, outperforming previous CNN-based and hybrid approaches. It exhibits impressive scalability, benefits from transfer learning, and offers a degree of interpretability through its attention mechanisms.

In conclusion, transformer-based approaches like ViViT represent a significant advancement in video action recognition, offering powerful tools for modeling complex spatiotemporal relationships, albeit with challenges in terms of computational resources and data needs.

## 2.4 Mother-Infant Interactions

Mother-infant interaction plays a critical role in the early development of a child, influencing cognitive, emotional, and social growth. Understanding these interactions is crucial for evaluating quality of the dyadic relationship and noticing any potential developmental problems in the children. This section explores the importance, key components, and traditional methods used to evaluate mother-infant interaction.

### 2.4.1 Importance in Child Development

Mother-infant interaction is foundational to a child's early development, impacting their cognitive, emotional, and social abilities. The quality of interaction is closely linked to attachment theory, which maintains that secure attachments are created through consistent, responsive caregiving that leads to healthier developmental outcomes [3][20]. Secure attachment is associated with better emotional regulation, social competence, and cognitive skills [21].

A core concept in mother-infant interaction is contingent responsiveness, where the mother responds to the infant's cues promptly and appropriately [3]. This back-and-forth

interaction fosters a sense of security and trust in the infant, laying the groundwork for future relationships. In contrast, research shows that disruptions in mother-infant interactions, such as lack of maternal sensitivity and contingency e.g. due to maternal mental health problems, can lead to child insecure attachment, cognitive delays, and emotional difficulties [22]. The brain's plasticity during early life makes infancy a critical developmental period; positive interactions can promote neural development, while negative experiences can have long-lasting impacts.

To quantify the importance, studies often use metrics like dyadic synchrony—a measure of the mutual, reciprocal interaction between mother and infant, which is correlated with healthy emotional and social development [23].

### 2.4.2 Key Components of Interactions

Mother-infant interaction consists of several key components, each contributing to the overall quality and effectiveness of the relationship. The primary components include joint attention, emotional attunement, and nonverbal communication.

Joint Attention involves both the mother and infant focusing on the same object or activity, creating a shared experience. For instance, when a mother points to an object and the infant follows with their gaze, it fosters an understanding of shared focus and intent [24]. This skill typically develops during the latter half of the first year and is crucial for language development and social cognition.

Emotional attunement refers to the mother's ability to perceive and respond to her infant's emotional states. This synchronization of emotional states helps the infant learn to regulate their emotions. The Still-Face Experiment by Tronick et. al [25] illustrates this, where a lack of emotional response from the mother leads to distress in the infant, highlighting the importance of attunement.

Nonverbal communication is heavily relied upon in early interactions, including touch, facial expressions, gestures, and vocalizations. These interactions lay the foundation for later social-emotional and language development. The turn-taking nature of these early communications mimics conversational patterns, helping infants learn the rhythm and structure of language [26][27][28].

Each of these components is crucial for fostering a secure attachment and promoting healthy development.

### 2.4.3 Traditional Assessment Methods

Traditional methods for assessing mother-infant interactions involve observational techniques and standardized assessment tools. These methods aim to evaluate the quality of interaction, attachment style, and potential developmental outcomes.

- Structured observations, such as the Strange Situation Procedure developed by Ainsworth et al. [3], are used to assess attachment styles by observing infants' reactions to their mother's presence and absence. This method classifies attachments into secure, avoidant, ambivalent, or disorganized attachments.
- Standardized Tools like the Maternal Behavior Q-Sort [29] and the CARE-Index [30] are used to quantify the quality of maternal behaviors and their impact on the infant. These assessments involve scoring behaviors based on criteria such as sensitivity, responsiveness, and emotional warmth.
- Behavioral Coding involves recording specific behaviors during interactions, such as the frequency of eye contact, vocalizations, and physical touch. These behaviors are then analyzed to assess the quality of the interaction [31].

These traditional evaluation methods are important for early identification of vulnerable mother-infant dyads and potential infant developmental problems. Identifying these issues in advance, allows for timely interventions. They are highly beneficial in identifying mothers who require support in their parenting [32]. However, these evaluation methods often require trained professionals and can be time-consuming. This has led to a rising interest in automated and technology-based evaluations.

## 2.5 Predictability of Maternal Sensory Signals

The predictability of maternal sensory signals is a critical aspect of mother-infant interaction, influencing the infant's cognitive, emotional, and neural development [9]. Predictable sensory experiences provided by the mother help infants develop a sense of security and learn about the world around them. This section explores the concept of predictability in maternal sensory signals, how it can be quantified, and its neural correlates in both the mother and the infant.

### 2.5.1 Concept and Importance

Predictability of maternal sensory signals refers to the consistent and regular stimuli (e.g., touch, voice, gaze) that a mother provides to her infant. These patterns are crucial for the infant's development. They help the infant build expectations about their surroundings, which is foundational for learning and cognitive development [33].

Predictable sensory experiences help the infant to develop temporal contingency, which allows the infant to anticipate the mother's actions and responses. This predictability promotes a secure attachment, as the infant feels more in control and secure in their environment. Research indicates that infants exposed to more predictable maternal interactions show better emotional regulation, cognitive development, and social engagement [9].

In contrast, unpredictable maternal behaviors can have adversarial effects on infant development. Davis et al. [8] found that exposure to unpredictable maternal sensory signals was related with poorer cognitive function in children. This emphasizes the importance of considering both predictability and unpredictability in maternal behavior.

For example, if a mother consistently smiles and talks to her baby while feeding, the infant starts associating feeding time with positive emotions and social interaction, reinforcing their attachment to the mother. This predictability also helps in the development of the infant's internal working model—a mental representation of the world and social relationships that guides future behavior and expectations [34].

### 2.5.2 Quantifying Predictability

To quantify the predictability of maternal sensory signals, researchers observe and code specific maternal behaviors from mother-infant interaction videos using established coding systems and computer-assisted programs such as the Noldus Observer XT. The following observable behaviors are used to characterize maternal sensory input to the infant. This is based on the coding system of *Maternal Sensory Behavior Coding Scheme (MSBCS)* [8].

In Table 1, auditory behaviour refers to Verbal utterances including speech and laughing, tactile behaviour refers to touching, holding, supporting or carrying the child and visual behaviour refers to manipulating object while child is observing. These categorized behaviors are used to construct a discrete time series and subsequent transition matrix for calculating the entropy rate as a measure of predictability in maternal sensory signals. Lower entropy indicates higher predictability, as the sequences of actions are more regular and consistent. The Entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \quad (1)$$

where $P(x_i)$ is the probability of the *i*-th event in the sequence.

Lower entropy values indicate higher predictability of the sequence. Another approach involves cross-recurrence quantification analysis (CRQA), which examines how patterns of maternal and infant behaviors recur over time.

This method provides insights into how synchronized or predictable the mother's behaviors are in response to the infant's actions.

**Table 1: Categorization of Observable Maternal Behaviors for Predictability Analysis**

| Category | Sensory Signals | Observable Behaviour |
|---|---|---|
| 1 | No Behaviour | No Auditory, Tactile, or Visual Stimulation |
| 2 | Single Behavior | Only Auditory Stimulation |
| 3 | Single Behavior | Only Tactile Stimulation |
| 4 | Single Behavior | Only Visual Stimulation |
| 5 | Combinations of Two Behaviors | Both Auditory & Tactile Stimulation |
| 6 | Combinations of Two Behaviors | Both Auditory & Visual Stimulation |
| 7 | Combinations of Two Behaviors | Both Visual & Tactile Stimulation |
| 8 | Combination of All Behaviors | All States: Auditory, Tactile, & Visual Stimulation |

Tools like dynamic time warping (DTW) can also be used to measure the alignment of maternal sensory signals over time. These quantitative measures help in understanding how consistent and predictable a mother's behavior is, which is crucial for the infant's development.

### 2.5.3 Neural Correlates of Predictability

The predictability of maternal sensory signals has major implications for the infants' neural development. Studies concerning predictable interactions have shown to influence the development of neural circuits involved in emotion regulation, learning, and memory.

Research using techniques like Functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG) clearly showed that predictable maternal interactions are associated with activation of prefrontal cortex and limbic system in the infant's brain [35]. These areas are important for emotional regulation, decision-making, and social cognition. For example, consistent maternal touch has been associated with better development of the somatosensory cortex, which is responsible for touch processing [36]. Addi-

tionally, predictable auditory signals, such as a mother's voice, have been shown to increase synchronization in the infant's auditory cortex, facilitating language development [37].

In mothers, predictable interactions are also reflected in brain activity. The reward system of the brain, particularly the ventral striatum, is often activated when a mother engages in predictable, responsive caregiving. This neural response reinforces the behavior, making it more likely that the mother will continue to provide consistent and predictable sensory experiences [38].

These neural correlates underline the importance of predictability in maternal behaviors, not just for the infant's brain development, but also for reinforcing the caregiving behaviors in mothers, creating a positive feedback loop that benefits both mother and child.

## 2.6 Domain Adaptation in Machine Learning

Domain adaptation is a crucial technique in machine learning aimed at improving the performance of models when applied to new, but related, domains. It addresses the challenge of transferring knowledge from a source domain (where a model is trained) to a target domain (where it will be applied) that may differ in distribution. This section explores three key concepts in domain adaptation: transfer learning, fine-tuning strategies, and layer-wise fine-tuning. These techniques help bridge the gap between source and target domains, enhancing model generalization and efficiency.

### 2.6.1 Transfer Learning

Transfer learning leverages knowledge from one domain (source) to improve learning in another, related domain (target). The core idea is to transfer learned features or representations from a pre-trained model to a new task where labeled data might be scarce.

Transfer learning relies on two fundamental components: pre-trained models and feature extraction. Pre-trained models, which have been trained on large datasets such as ImageNet, serve as starting points and capture useful features applicable to similar tasks. Through feature extraction, these pre-trained model layers provide input features for new models, eliminating the need for training from scratch. For example, features extracted from the convolutional layers of a pre-trained CNN can be effectively utilized for different image classification tasks, leveraging the learned representations from the source domain. Given a pre-trained model $f_{\text{source}}$ trained on source domain $\mathcal{D}_S$ and a target domain $\mathcal{D}_T$, the objective is to minimize the target loss function $L_T$:

$$L_T(\theta_T) = \frac{1}{N_T}\sum_{i=1}^{N_T} \text{loss}(f_T(x_i;\theta_T), y_i)$$

where $\theta_T$ are the parameters fine-tuned for the target domain, and $N_T$ is the number of target samples. The model $f_T$ is initialized as a copy of the pre-trained model $f_{\text{source}}$, inheriting its architecture and learned parameters. Through fine-tuning, $f_T$'s parameters $\theta_T$ are then gradually updated to minimize the loss function $L_T$ on the target domain $\mathcal{D}_T$, potentially diverging from $f_{\text{source}}$ to better fit the new task.

### 2.6.2 Fine-Tuning Strategies

Fine-tuning involves adapting a pre-trained model to a new task by continuing training on the target domain dataset. It adjusts the model's parameters to better fit the new domain.

Fine-tuning strategies encompass two primary approaches: full fine-tuning and partial fine-tuning. In full fine-tuning all layers of the pre-trained model are updated during training. This is useful when the target domain is significantly different from the source domain. This can be mathematically expressed as:

$$\theta_T = \theta_S - \eta\nabla_{\theta_S} L_T$$

where $\eta$ is the learning rate, and $\nabla_{\theta_S} L_T$ represents the gradient of the target loss with respect to the pre-trained parameters, $\theta_S$ are the initial parameters of the pre-trained (source) model, and $\theta_T$ are the updated parameters of the target.

Partial fine-tuning, on the other hand, involves updating only the last few layers or specific parts of the model, proving effective when the target domain shares similarities with the source domain.

### 2.6.3 Layer-wise Fine-Tuning

Layer-wise fine-tuning is a specific fine-tuning strategy where different layers of the network are trained at different stages. This method aims to retain the useful features learned from the source domain while gradually adapting the model to the target domain. A stagewise approach to Layer-wise fine-tuning is as discussed below:

Stage 1: Train the last layers of the network with a low learning rate while keeping earlier layers frozen. This helps the model learn the specific features of the target domain. It can be expressed as:

$$\theta_L = \theta_L^{\text{old}} - \eta_L\nabla_{\theta_L} L_T$$

where $\theta_L$ denotes the parameters of the later layers, $\eta_L$ is the learning rate for these layers, and $\theta_L^{\text{old}}$ represents a subset of the original $\theta_S$, specifically the parameters of the later layers before the current update step in layer-wise fine-tuning.

Stage 2: Here we slowly unfreeze and fine-tune the earlier layers. This step helps to adjust lower-level features that are relevant to the target domain. This is represented as:

$$\theta_E = \theta_E^{\text{old}} - \eta_E \nabla_{\theta_E} L_T$$

where $\theta_E$ denotes the parameters of the earlier layers, and $\eta_E$ is the learning rate for these layers.

## 2.7 Voice Activity Detection (VAD)

Voice Activity Detection (VAD) is a fundamental technology in audio processing that identifies speech and non-speech sections in an audio signal. Its applications include speech recognition, telecommunications, and noise reduction. VAD aims to identify speech-containing regions in audio signals for optimal processing and enhanced performance in subsequent tasks. This section explores the methodologies involved in VAD, such as feature extraction, decision-making processes, and the challenges faced in practical implementations.

### 2.7.1 Feature Extraction

Feature extraction in VAD involves the analysis of audio signal to derive characteristics that will help distinguish between speech and non-speech segments. Some common features include:

Energy: It measures the signal's intensity. Speech most often tends to have elevated energy in comparison to silence or background sounds. Energy $E(t)$ within a short time frame could be calculated as:

$$E(t) = \frac{1}{N} \sum_{n=0}^{N-1} x^2(n)$$

where *x(n)* is the audio signal, and *N* is the number of samples in the time frame.

Zero-Crossing Rate (ZCR): This feature assesses the frequency with which a signal crosses the X-axis. Increased ZCR is a pointer to the presence of speech. ZCR can be calculated as:

$$\text{ZCR}(t) = \frac{1}{2N} \sum_{n=1}^{N-1} \left| \text{sgn}\big(x(n)\big) - \text{sgn}\big(x(n-1)\big) \right|$$

where sgn(·) denotes the sign function.

Mel-Frequency Cepstral Coefficients (MFCCs): These features represent the short-term power spectrum of a sound signal. The MFCCs capture the phonetic content of speech and are extracted using transformations that include Fourier Transform and Mel filter banks. Feature extraction is usually carried out over a short window, usually 20 ms, which captures the dynamic nature of speech features.

### 2.7.2 Decision Making

Voice Activity Detection employs sophisticated algorithms to classify audio frames as speech or non-speech segments. The classification process integrates various features extracted from the audio signal through statistical models and machine learning approaches.

Thresholding is a technique used in Voice Activity Detection (VAD). It involves setting thresholds on features such, as energy or Zero Crossing Rate (ZCR). For instance, if the energy level $E(t)$ of a frame surpasses a threshold value the frame is identified as containing speech.

Advanced techniques use models like Gaussian Mixture Models (GMMs) as well, as Hidden Markov Models (HMMs) to analyze the probability distributions of speech and non-speech characteristics. These models calculate the chances of a frame being categorized into one of the groups to make the classification decision.

Recent advancements have introduced machine learning algorithms, like Support Vector Machines (SVMs) and deep neural networks. These are trained on labeled audio data to identify intricate patterns. These supervised learning approaches have shown superior performance in challenging real-world conditions.

The decision rules and thresholds are fine-tuned based on performance metrics like accuracy, precision, recall, and F1-score. This ensures a balance between false positives and false negatives.

### 2.7.3 Challenges in VAD

Despite significant progress, VAD systems face several technical challenges that affect their reliability and precision. One key challenge is dealing with background noise, which requires VAD systems to distinguish speech, from noises and background conversations. The presence of noise can change the distribution of features resulting in misidentification of speech segments.

Dealing with non-stationary noise, which varies temporally, can be quite tricky as it affects performance stability. To tackle these issues, advanced techniques like noise adaptation and robust feature extraction have been developed. Moreover, overlapping speech scenarios which are quite common in natural conversations VAD methods often struggle. To handle such cases, modern approaches utilize source separation techniques.

In environments with Low Signal-to-Noise Ratio (SNR), speech detection becomes particularly difficult due to background noise interference levels being higher than the signal itself. Advanced signal processing techniques like spectral subtraction and Wiener filtering have shown to improve performance under these conditions.

Dealing with these challenges involves a mix of feature extraction methods, robust decision-making algorithms, and adaptive strategies, for making decisions robustly in different acoustic conditions.

## 2.8 Computer Vision for Gaze Direction Detection

Computer vision for gaze direction detection is about analyzing and understanding eye movements and gaze to know where someone is looking. Gaze detection comprises several steps: eye detection, pupil localization, gaze estimation and handling various technical challenges. Each of these steps relies on the one before it to develop a framework, for interpreting gaze direction from visual data.

### 2.8.1 Eye Detection

Eye detection is the first step in detecting gaze direction. It involves finding the eyes in a face image or video stream. This is critical as it sets the foundation for the rest.

Eye detection can be done in several ways. The Haar Cascade algorithm is a basic approach that uses machine learning to train classifiers for face features. The Haar cascade algorithm detects eyes by applying a series of increasingly complex classifiers on different parts of the image. This is a computationally efficient approach that is commonly used in real-time applications.

Haar-like features for eye detection calculates the difference between the sum of pixel intensities in adjacent rectangular regions, usually the lighter ("White Area") and darker ("Black Area") parts of the image. This difference summed over multiple rectangle pairs forms the feature value F that helps to distinguish eye regions from non-eye regions in the image. The Haar-like feature used for eye detection is:

$$F = \sum_{i=1}^{m} (\text{White Area} - \text{Black Area})$$

where *F* is the Haar feature, and the areas are different parts of the detected region. CNNs have become a powerful alternative for eye detection, using large datasets to learn complex patterns for accurate eye localization.

### 2.8.2 Pupil Localization

Once the eyes are detected, the next step is pupil localizationAccurate localization of the pupil is critical for gaze direction. The process usually involves thresholding, converting eye images to grayscale and applying segmentation to separate the darker pupil from the surrounding iris and sclera. The Hough Transform is an alternative approach for precise pupil detection, mathematically defined as: $\rho = x\cos\theta + y\sin\theta$, where $\rho$ is the radius and $\theta$ is the angle.

### 2.8.3 Gaze Estimation

Gaze estimation involves mapping the pupil's position to know where someone is looking. This process combines information from eye detection and pupil localization to estimate gaze direction accurately. Geometric models use the relationships between pupil position, corneal reflections and camera parameters to estimate gaze direction. This approach often uses a model of the eye's anatomy to map pupil position to a point of regard in 3D space. The gaze direction can be estimated by:

$$\text{Gaze Vector} = \text{Eye Position} - \text{Pupil Center}.$$

Recent advancements in deep learning have proven effective when applied to gaze estimation. CNNs are trained to predict gaze direction directly from eye images, learning complex mappings from vast training datasets. These diverse approaches provide valuable strengths for robust gaze detection and tracking.

### 2.8.4 Challenges in Gaze Detection

Detecting gaze direction comes with several technical hurdles in real-world applications. Changes in lighting conditions can significantly influence the appearance of eyes and pupils making it hard to maintain consistent detection. This task is further complicated by head movements, which can alter the appearance of the eyes and pupils, thus affect the precision of gaze estimation. Additionally, glare caused by reflections, from glasses or light sources can obscure the pupil and interfere with accurate detection. The unique

variations, in peoples features like eye shape and color impact the effectiveness of algorithms for detecting eyes and pupils. To overcome these challenges, several strategies have been developed such as adaptive algorithms that can adjust to changing lighting conditions and head movements, robust deep learning models trained on diverse datasets to handle differences in eye appearance and glare, and multi-view systems that use multiple cameras or sensors to capture eye images from different angles. The integration of these advanced image processing techniques, strong machine learning models, and adaptive systems has been crucial for improving the reliability and accuracy of gaze direction detection systems. This holistic approach allows the effective management of real-world variability while maintaining consistent performance across various conditions and subjects.

# 3. REVIEW OF MOTHER-INFANT INTERACTION ANALYSIS

This chapter provides a comprehensive review of existing research on mother-infant interactions and the application of machine learning in this context. It explores four key areas: video action recognition, the predictability of maternal sensory signals, machine learning models for analyzing mother-infant interactions, and signal processing techniques specific to these interactions. By examining these interconnected fields, this chapter establishes the foundation for understanding the current state of the art in analyzing mother-infant interactions.

## 3.1 Video Action Recognition: Foundations and Advances

Over the last two decades, Video action recognition, has undergone significant changes. This field primarily focuses on automatically identifying and classifying human actions and activities in video data. Its applications span from surveillance and human-computer interaction to behavioral analysis, which makes it particularly useful for studying mother-infant interactions.

The foundations of video action recognition were established in the early 2000s by Laptev [39] who introduced of Space-Time Interest Points (STIPs). This method expanded the idea of spatial interest points to the temporal domain. It provided a framework for identifying salient spatiotemporal positions in videos. STIPs were often combined with descriptors such as Histograms of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) as proposed by Dalal and Triggs [40].

During this period, another significant contribution was the development of 3D SIFT descriptors by Scovanner et al. [41]. These descriptors enhanced the popular SIFT algorithm to capture both spatial and temporal information in videos.

The concept of dense trajectories proposed by Wang et al. [42], marked a significant improvement in handcrafted features. This approach involved using optical flow to track points and extract features along their paths. This is mathematically represented as:

$$T = \{(x_t, y_t): t = 1, \dots, L\}$$

where $(x_t, y_t)$ represents the trajectory point at time *t*, and *L* is the trajectory length.

The emergence of deep learning techniques in the 2010s brought about a significant change in the field of video action recognition. Karpathy et al. [43] were one of the first to apply CNNs to video classification, experimenting with different fusion strategies to combine information across frames.

A significant breakthrough came with the introduction of two-stream CNNs by Simonyan and Zisserman [44]. This architecture processed spatial and temporal information separately:

$$P(y|V) = F\left(C_s(I_t), C_t\left(F_{t,t+\Delta}\right)\right)$$

where $C_s$ and $C_t$ are spatial and temporal stream CNNs respectively, $I_t$ is a single frame at time $t$, and $F_{t,t+\Delta}$ , represents optical flow fields between frames *t* and *t+Δ*.

Tran et al. [45] proposed 3D CNNs (C3D) to learn spatiotemporal features directly from raw video frames:

$$f(x) = \sigma(W * x + b)$$

where $x$ is a 3D video volume, $W$ are 3D convolutional filters, $b$ is a bias term, and $\sigma$ is a non-linear activation function.

The I3D (Inflated 3D ConvNet) architecture, introduced by Carreira and Zisserman [46], built upon the success of 2D CNNs by inflating their filters and pooling kernels into 3D, enabling them to capture spatiotemporal features effectively. Wang et al. [47] proposed the Temporal Segment Network (TSN), which addressed the issue of modeling long-range temporal structures in videos. TSN operates on a sequence of short snippets sparsely sampled from the entire video, combining the outputs using a segmental consensus function.

Non-local Neural Networks, introduced by Wang et al. [48], brought the attention mechanism to video understanding:

$$y_i = \frac{1}{C(x)} \sum_j f(x_i, x_j) g(x_j)$$

where *f* computes pairwise interactions and *g* computes a representation of the input *x* at position *j*.

Transformer-based architectures have shown promising results in recent years. Bertasius et al. [17] introduced TimeSformer, which applies self-attention across both spatial and temporal dimensions:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where *Q, K*, and *V* are query, key, and value matrices derived from the input video sequence as discussed in Section 2.2.6.

The X3D architecture, proposed by Feichtenhofer [49], presented a family of efficient video networks that progressively expand a tiny 2D image classification architecture along multiple network axes.

In 2021, Arnab et al. [19] introduced ViViT (Video Vision Transformer), which explored different ways to factorize the spatial and temporal dimensions of video for efficient processing with transformers. This work demonstrated that pure-transformer models can achieve state-of-the-art results on major action recognition benchmarks.

A significant development was the MaskFeat approach by Wei et al. [50]. This self-supervised learning method for video recognition uses masked feature prediction, achieving state-of-the-art results on various video understanding tasks.

G. Bertasius et al. [17] proposed a model, which adapts the Transformer architecture for video understanding by applying self-attention separately in the spatial and temporal dimensions. This approach, referred to as “divided attention,” improves video classification accuracy and reduces computational complexityty. Chen et al. [51] introduced the Video Language Model (VideoLLM), a large language model for video understanding. This model can generate natural language descriptions of video content and answer questions about videos, representing a significant step towards more comprehensive video understanding.

Despite these significant advancements, there are several challenges that still exist in video action recognition. Effectively capturing long-range temporal dependencies in videos remains a challenge, especially for complex, multi-step actions. As models evolve, balancing performance with computational efficiency is crucial, especially in real-time applications. Developing models that can recognize new actions with minimal or zero training examples is a key focus of current research. Incorporating other modalities of data (e.g., audio, text) with video for more comprehensive action understanding. This is a promising direction in future research. Interpretability is crucial as models become more complex in ensuring their results are interpretable and explainable. Interpretability is especially important in sensitive applications like analyzing mother-infant interactions. Privacy and ethics are aspects to consider when these technologies are applied to analyze human behavior. It is essential to address privacy concerns and ensure ethical use of the technology.

The field of video action recognition has progressed significantly from its early days of handcrafted features to the current era of advanced deep learning models. As research progresses, we can expect more accurate and efficient models that can be applied to a wide range of applications.

## 3.2 Predictability of Maternal Sensory Signals and Its Role in Infant Development

The concept of predictability in maternal sensory signals has gained significant interest over the past few decades. This field of research has progressed significantly from early attachment theories and animal-based studies into an advanced, multidisciplinary field, offering insights into how infants develop and introduces new ways for intervention and assistance for mothers and infants.

Early foundations of this research can be linked back to Bowlby's attachment theory [20] and Ainsworth's research on maternal sensitivity [3]. Although these early works did not specifically discuss predictability, they emphasized the significance of consistent and responsive caregiving for healthy infant development. These works laid the groundwork for future research. Bowlby's theory highlighted the importance of a consistent and responsive caregiver for the development of secure attachment, whereas Ainsworth's research provided empirical evidence for these ideas through their observations of mother-infant interactions.

The explicit focus on predictability of maternal signals became more prominent in the early 2000s. Davis et al. [8] conducted a seminal study demonstrating that unpredictable maternal signals are associated with poorer cognitive function in children. They quantified maternal signal predictability using entropy rate as discussed in Eq. (1). This work was pivotal in establishing a quantitative measure for maternal behavior predictability and linking it directly to child outcomes. The study showed that higher entropy (indicating less predictability) in maternal behavior was associated with lower cognitive scores in children, even after controlling for other factors such as socioeconomic status.

Building on these foundations, Molet et al. [52] showed that early-life unpredictability can lead to anhedonia-like behavior in adolescence. They used similar entropy-based measures to quantify predictability in maternal care patterns but extended the research to examine long-term outcomes. Their findings suggested that the impact of unpredictable maternal care extends far beyond infancy, potentially influencing emotional regulation and reward processing in later life.

Baram et al. [9] proposed a unifying theory that connects early-life unpredictability to altered changes of cognitive and emotional brain circuits. They argue that predictable patterns of sensory inputs are essential for the development of synaptic connections and neural network formation. This theory offers a neurobiological perspective on how early experiences of unpredictability can alter brain structure and function, in lasting ways.

In the recent years there have been significant advancements in this field of research, as researchers are employing more advanced methodologies and exploring broader implications of their work. Glynn and Baram [53] carried out a study on how unpredictable sensory signals can influence the developing brain and put forward mechanisms by which unpredictable patterns could affect normal neurodevelopmental processes. They highlighted the influence of stress hormones and changes in neural plasticity as factors that could mediate these impacts.

Davis et al. [54] furthered the research to cross-cultural contexts, highlighting the universal importance of predictable maternal signals while also noting cultural variations in caregiving practices. This study was particularly important in demonstrating that while the specific behaviors constituting predictable care might vary across cultures, the overall impact of predictability on child development appears to be consistent.

Advances in neuroscience have illustrated the neural mechanisms through which predictable maternal behaviors influence infant development. As discussed by Reindl et al. [55] using functional near-infrared spectroscopy (fNIRS) to study mother-infant interactions revealed that infants experiencing predictable maternal behaviors exhibit enhanced neural processing efficiency, particularly in brain regions vital for social cognition and emotional development.

In conclusion, the study of predictability in maternal sensory signals has evolved significantly, bridging neuroscience, developmental psychology, and behavioral research. From attachment theory to modern neuroimaging studies, this field has demonstrated how early-life experiences shape neural architecture and behavioral outcomes. These insights offer practical implications for early intervention programs and parental support initiatives, promising to enhance our understanding of mother-infant relationships and inform future developmental interventions.

## 3.3 Machine Learning models for Analyzing Mother-Infant Interactions

The application of machine learning to analyze mother-infant interactions has evolved significantly over the past few decades. It developed from traditional statistical approaches to advanced deep learning models. This section provides a comprehensive overview of the field's historical development, highlighting key contributions and recent advancements.

Initial efforts to analyze mother-infant interactions heavily relied on statistical methods and conventional machine learning techniques. These early approaches focused on quantifying the dynamic and often subtle exchanges between mothers and their infants.

Messinger et al. [56] applied dynamic time warping (DTW) and hidden Markov models (HMMs) to characterize face-to-face interactions between mothers and infants. Their work demonstrated the potential of machine learning techniques to model and analyze complex and variable interaction sequences. These models were particularly effective in capturing the temporal dynamics of interactions.

Rehg et al. [57] advanced the field by introducing the Multimodal Dyadic Behavior (MMDB) dataset, which provided a rich resource for analyzing child-adult interactions by combining multiple data streams such as video, audio, and physiological signals. This work highlighted the value of multimodal approaches in understanding complex social behaviors.

Yu and Smith [58] utilized support vector machines (SVMs) to classify infant-parent joint attention patterns. By analyzing eye-tracking data, they elucidated how infants and parents coordinate their visual attention during object exploration, emphasizing the role of gaze and joint attention in early social development.

The advent of deep learning marked a significant shift towards more sophisticated models capable of handling large and complex datasets. Deep learning models, particularly CNNs and RNNs, allowed for more nuanced analysis of interaction dynamics.

Pusiol et al. [59] were among the first to apply 3D CNNs to detect and classify behaviors in naturalistic infant-caregiver interactions. Their approach was groundbreaking in its ability to automatically analyze complex, real-world interaction scenarios, particularly in the context of developmental disorders.

As deep learning models evolved, researchers started to explore the potential of spatio-temporal models. Cao et al. [60] introduced OpenPose. It is a real-time system for multi-person 2D pose estimation using Part Affinity Fields. Even though OpenPose is not specifically designed for mother-infant interactions, it has been widely used for analyzing body language and non-verbal cues in dyadic interactions. In recent years, there has been a growing interest in multimodal approaches that integrate different types of data to provide a deeper understanding of mother-infant interactions.

Gordon et al. [61] investigated the use of physiological and behavioral synchrony measures to predict group cohesion and performance. While their work focused on group settings, the integration of physiological and behavioral data has significant implications for studying synchrony in dyadic relationships, including mother-infant interactions.

The development of transformer-based models has further advanced the field by allowing for the integration of long-range dependencies in interaction sequences. Vaswani et al. [16] introduced the transformer architecture, which revolutionized natural language

processing and has since been adapted for various multimodal tasks, including interaction analysis.

The application of RL to mother-infant interactions is a relatively new but promising development. RL models can learn from interactions over time, making them well-suited for modeling the dynamic and reciprocal nature of mother-infant exchanges.

Self-supervised learning (SSL) has also emerged as a powerful tool in this field, particularly for extracting meaningful representations from unlabeled interaction data. Withanage Don et.al [62] presented a novel approach to automating the annotation of infant-caregiver interactions using multimodal self-supervised learning models, achieving improved accuracy in predicting engagement phases.

As the field of machine-learning continues to grow, the development of even more advanced models that leverage multimodal data, handle temporal dynamics more effectively, and provide interpretable insights into the complexities of early social interactions is possible.

## 3.4 Signal processing for mother-infant interactions

Signal processing techniques have been essential, in analyzing mother-infant interactions by helping researchers to extract meaningful information from complex multimodal data. This section provides a comprehensive overview of the historical development and current state of signal processing methods applied to this field.

The initial applications of signal processing in mother-infant interaction studies focused primarily on audio analysis. In the 1980s, researchers began using spectral analysis techniques to study infant cries and maternal vocalizations. Lester et al. [63] pioneered the use of spectrographic analysis to characterize infant cry acoustics, establishing relationships between cry features and infant health status. This work laid the foundation for further research using more advanced signal processing techniques.

In the 1990s, Papousek [64] used time-frequency analysis to study the prosodic features of infant-directed speech. These early studies relied on Short-Time Fourier Transform (STFT) to analyze the time-varying spectral content of vocalizations:

$$STFT\{x[n]\}(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{\{-j\omega n\}}$$

where: $x[n]$ is the input signal, $w[n]$is the window function, *m* is the time index, and *ω* is the angular frequency.

As technology advanced, more sophisticated audio processing techniques emerged. Fernald and Kuhl [65] utilized pitch tracking algorithms to study the exaggerated pitch contours in infant-directed speech. These methods often employed autocorrelation-based techniques for fundamental frequency estimation.

In the early 2000s, Burnham et al. [66] applied formant analysis to investigate vowel hyper articulation in infant-directed speech. Formant estimation typically involves Linear Predictive Coding (LPC) analysis.

The mid-2000s saw a shift towards multimodal analysis, integrating video and physiological data alongside audio. Cohn and Tronick [67] developed coding systems for facial expressions and body movements, which later informed automated video processing techniques.

Messinger et al. [68] introduced computer vision techniques to automatically track facial expressions in mother-infant interactions. These methods often employ facial landmark detection algorithms, such as Active Appearance Models (AAMs) or more recent deep learning-based approaches.

In the 2010s, researchers began applying more advanced time series analysis techniques to study the dynamics of mother-infant interactions. Lester et al. [69] utilized cross-recurrence quantification analysis (CRQA) to examine the temporal coordination between mother and infant behaviors.

The integration of physiological measures has provided new insights into the biological underpinnings of mother-infant interactions. Feldman et al. [70] applied heart rate variability (HRV) analysis to study physiological synchrony between mothers and infants.

More recently, Leong et al. [71] have employed EEG hyper scanning to investigate neural synchrony during mother-infant interactions. These studies often utilize coherence analysis to measure the degree of coupling between two EEG signals:

$$C_{xx}(f) = \frac{\left|G_{xy}(f)\right|^2}{G_{xx}(f) \cdot G_{yy}(f)}$$

where: $G_{xy}(f)$ is the cross-spectral density between signals *x* and *y,* $G_{xx}(f)$ and $G_{yy}(f)$are the autospectral densities of *x* and *y, and f* is the frequency of interest.

Looking ahead, researchers are exploring continuous multimodal monitoring techniques for long-term analysis of mother-infant interactions. Azhari et al. [72] reviewed recent advances in wearable technologies for studying parent-infant interactions, highlighting the potential for naturalistic, longitudinal studies.

The field of signal processing for mother-infant interactions has evolved dramatically over the past few decades, from simple audio analysis to sophisticated multimodal approaches incorporating advanced machine learning techniques. As technology continues to advance, we can expect further innovations in continuous monitoring, real-time processing, and privacy-preserving analysis strategies based on detailed interaction data.

# 4. METHODOLOGY

This chapter presents a comprehensive description of the methodological approach employed in this study to analyze mother-infant interactions and assess the predictability of maternal sensory signals. The methodology encompasses multiple interconnected components as illustrated in Figure 4.1. The chapter details the dataset preparation process, including data collection and preprocessing, followed by the novel Hybrid Deep Learning Framework that integrates PGMHA and TBSM. The implementations of VAD for auditory analysis and computer vision techniques for visual stimuli assessment are described, concluding with the approach to entropy rate calculation for quantifying maternal behavioral predictability.

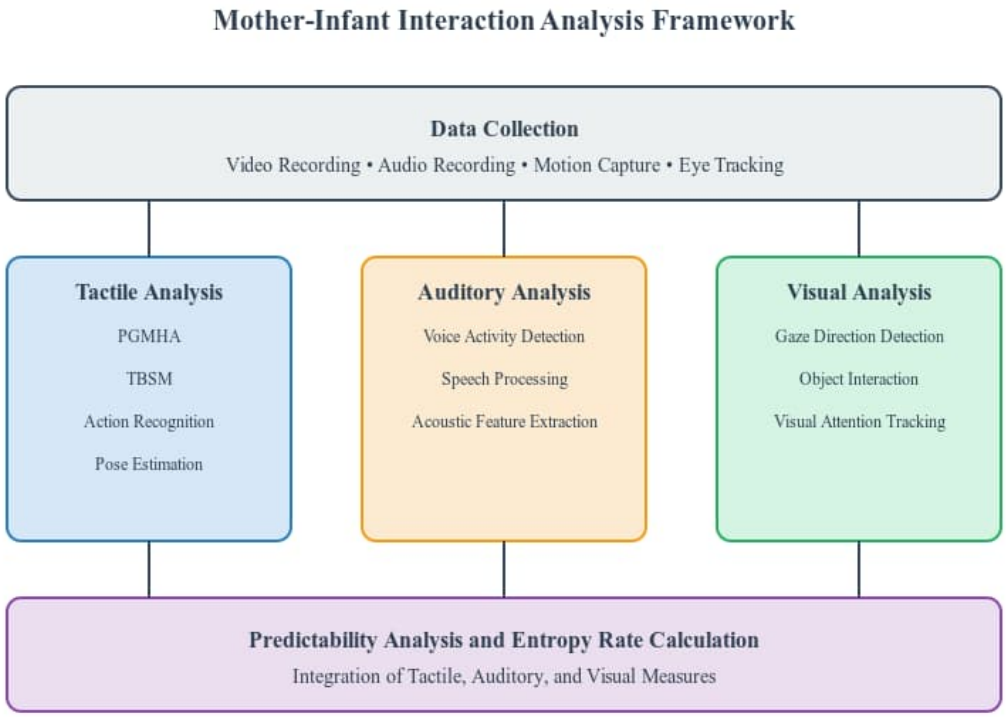


***Figure 4.1*** *Methodological framework for maternal predictability analysis.*

## 4.1 Dataset Preparation

The dataset for this study comprises observations from 63 Finnish mother-infant pairs. The mothers' ages ranged from 21 to 43 years at the time of the study, while the infants were between 5 and 8 months old. Data collection took place in a laboratory setting at Tampere University's Hervanta campus, specifically in the Centre of Immersive Visual

Technologies (CIVIT) laboratory. The Ethics Committee of the Tampere Region approved the methods and data collection in year 2021.

Each mother-infant pair participated in the following activities:

1. A 12-minute videotaped free play session.
2. The Marschak Interaction Method (MIM) for babies, consisting of 5 semi-structured interaction tasks.
3. Completion of an electronic questionnaire.

### 4.1.1 Data Types Collected

The dataset for this study consists of a comprehensive collection of multimodal data from mother-infant dyads. It includes high-fidelity audiovisual recordings, three-dimensional motion capture data, gaze information, and spatial audio recordings. This diverse dataset allows a deeper analysis of mother-infant interactions across various sensory modalities.

### 4.1.2 Data Acquisition Equipment and Collection Techniques

The study used a comprehensive multimodal data acquisition setup comprising high-fidelity video cameras, eye-tracking devices, and professional audio recording equipment. Each component was carefully selected to ensure optimal data quality while maintaining ecological validity in the mother-infant interactions.

The primary video recording setup used high-quality cameras to capture detailed footage of the mother-infant interactions. Specifically, Basler acA 1920-50gc cameras were utilized, with VS-0618H1 lenses. These cameras were strategically positioned to provide optimal coverage of the designated play area. The video was recorded at a resolution of 1920x1200 pixels, with a frame rate of 20 fps. These were initially captured in raw format to preserve maximum fidelity; the footage was later compressed to MP4 format for practical data management purposes. Figure 4.2 illustrates the experimental setup of recording environment.

The study utilized a lightweight custom 3D motion capture setup to track marker positions on both mother and infant. This system, designed to be less intrusive than traditional MOCAP suits and beanies, allowed for more natural interactions. Custom tools were developed to extract spatial positions, enabling analysis of mother-infant distance, motion over time, and touching interactions. This approach provided detailed spatiotemporal data while minimizing disruption to the participants' behavior.

Tobii Pro eye-tracking glasses were used to extract gaze information from the mothers. Due to practical constraints, gaze-tracking data was collected from 25 of the 63 mother-infant pairs. Despite this limitation, the collected data allows for quantification of the mother's gaze directed at the infant's face, providing valuable insights into visual attention patterns during interactions.

***Figure 4.2*** *Experimental Setup: Mother-Infant Interaction Recording Environment.*

The audio recording setup utilized a RØDE Wireless GO II dual-channel wireless system with RØDE Lavalier GO microphones, enabling independent vocal recording for both mother and infant. This multi-channel configuration ensured high-quality spatial audio capture and precise attribution of vocalizations during interactions, complementing the visual and motion capture data streams.

The integration of these recording modalities - video, motion capture, eye-tracking, and spatial audio - generated a rich multimodal dataset for analyzing mother-infant interactions across multiple dimensions of communication and engagement. This comprehensive data collection approach enables detailed analysis of temporal synchrony, mutual responsiveness, and interaction patterns between mother and infant while maintaining a naturalistic environment for the participants. For this study we exclusively used primary videos, Tobii pro tracking videos and audio data. Given the multifaceted nature of mother-infant interactions, video and audio data comprehensively captures the complex spatiotemporal dynamics of maternal behaviors, including tactile, visual, and auditory stimuli. Thus, enabling quantitative assessment of maternal sensory predictability.

### 4.1.3 Data Preprocessing

The collected data underwent several preprocessing steps to prepare it for analysis. From the total dataset of 63 mother-infant pair recordings, 12 videos were selected for this study - 8 for training and 4 for validation. Each 12-minute free play session was segmented into shorter clips and categorized into four predefined classes of mother-infant interactions: mother holding the baby, mother interacting with an object, mother not holding the baby, and mother not interacting with an object. The training dataset consisted of 40 video clips per class (totaling 160 clips), while the validation dataset included 15 video clips per class (totaling 60 clips).

Separate preprocessing was performed for the subset of 25 videos that included gaze-tracking data, enabling extraction of visual stimuli information. Trained psychology professionals manually coded the interactions using standardized coding systems. While accurate, this manual process was time-consuming and expensive, highlighting the need for more efficient, dynamic, and multimodal approaches to analysis. Figure 4.3 gives us an example of the free play.

***Figure 4.3*** *Mother-Infant Dyad During Free-Play Session*

### 4.1.4 Data Augmentation

To enhance the robustness of our model and mitigate potential overfitting due to the limited size of the dataset, data augmentation techniques were applied to the training dataset. These techniques included:

1. Temporal augmentation: Random temporal cropping and frame skipping

2. Spatial augmentation: Random flipping, rotation, and scaling

3. Intensity augmentation: Adjustments to brightness, contrast, and hue

This comprehensive dataset, combining various data types and augmentation techniques, forms the foundation for subsequent analysis aimed at developing cost-efficient, dynamic, and multimodal approaches to studying early dyadic parent-infant interactions.

## 4.2 Hybrid Deep Learning Framework

This section introduces a novel Hybrid Deep Learning Framework designed to enhance action recognition in mother-infant interaction videos. The framework integrates advanced machine learning and computer vision techniques to analyze the complex, multimodal nature of parent-child interactions. Our approach aims to provide a more efficient, scalable, and objective method for analyzing these critical interactions that shape early childhood development.

The Hybrid Deep Learning Framework consists of two main components:

1. Pose-Guided Motion History Analysis (PGMHA)
2. Transformer-Based Sequence Modeling (TBSM).

These components work in tandem to capture both spatial and temporal aspects of mother-infant interactions, providing a comprehensive analysis of these complex behavioral patterns.

### 4.2.1 Pose-Guided Motion History Analysis (PGMHA)

PGMHA is an innovative approach that combines pose estimation with motion history images (MHIs) to capture the nuanced movements characteristic of parent-child interactions. This component integrates skeletal representations from pose estimation with motion history, offering a detailed view of spatial-temporal dynamics.

The concept of Motion History Images was first introduced by Bobick and Davis [73]. They proposed MHIs to represent and recognize human actions in video sequences. Our approach builds upon this foundation, incorporating modern pose estimation techniques to enhance the effectiveness of MHIs in capturing subtle interactions.

The PGMHA process involves the following steps: Pose Estimation, Motion History Image Generation, Pose-Guided MHI Integration.

1. Pose Estimation: Pre-trained pose estimation models such as MoveNet [50] or MediaPipe [74] are used to extract skeletal information from each video frame. Let $P_t$ represent the set of pose keypoints at time t:

$$P_t = \{(x_i, y_i, c_i) \mid i = 1, \ldots, K\}$$

where $(x_i, y_i)$ are the coordinates of the *i*-th keypoint, $c_i$ is its confidence score, and *K* is the total number of keypoints.

2. Motion History Image (MHI) Generation: MHIs are created by assigning intensities to pixels based on the recency of motion at that location. The outline of the algorithm used to generate MHIs is as follows:

1. Start with an MHI, H(*x,y,t*), initialized to zero for all pixel positions (*x,y*) at time t. Here, t represents the current frame in the sequence.
2. For each frame at time t:
   a. Calculate the difference between the current frame and the previous frame to detect motion. This can be done using frame differencing:

      D$(x, y, t)$=|I$(x, y, t)$ - I$(x, y, t$-1)|

      where D(x,y,t) is the motion mask for the current frame, and I(x,y,t) is the intensity of the pixel at position (x,y) in the frame at time t.

   b. Apply a threshold to D(x,y,t) to identify areas with significant motion. This binary mask can be defined as:

$$M(x, y, t) = \begin{cases} 1, & \text{if } D(x, y, t) > threshold \\ 0, & \text{otherwise.} \end{cases}$$

   c. The MHI is updated based on the motion mask *M(x,y,t)*. For pixels with motion, *M(x,y,t)* = 1, set the MHI to the maximum intensity. For pixels without motion, decay the intensity of the MHI to indicate the aging of the motion information. This can be formulated as:

$$H(x, y, t) = \begin{cases} \tau, & \text{if } M(x, y, t) = 1 \\ H(\mathrm{x}, \mathrm{y}, \mathrm{t} - 1) * decay\_factor, & \text{otherwise} \end{cases}$$

   where $\tau$ is the maximum intensity value, and decay factor is a number less than 1, that represents how much of the previous intensity value at each pixel should be retained for the next time step.

3. To facilitate further processing, we normalize the MHI so that the intensities span the full range of possible values.

3. Pose-Guided MHI Integration: We modify the traditional MHI generation process to incorporate pose information. A weighting function $w(x, y, P_t)$ is introduced to prioritize regions around detected poses:

$$w(x, y, P_t) = \max_{i=1..K} \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right)$$

where $\sigma$ is a parameter controlling the spread of influence around each keypoint.

The pose-guided MHI is then defined as:

$$H_\tau^{PG}(x, y, t) = \begin{cases} \tau \cdot w(x, y, P\_t), & \text{if } D(x, y, t) = 1 \\ max(0, H_\tau^{PG}(x, y, t-1) - w(x, y, P_t)), & \text{otherwise.} \end{cases}$$

This pose-guided approach ensures that the MHI captures not just any motion in the frame, but specifically the motion related to the interacting individuals, making it more relevant for analyzing mother-infant interactions.

### 4.2.2 Transformer-Based Sequence Modeling (TBSM)

TBSM leverages the power of transformer architectures to process and understand the temporal dynamics of the interactions. This component is designed to recognize patterns corresponding to the four predefined classes of mother-infant interactions across sequences of frames.

The TBSM process includes:

1. Feature Extraction: Features are extracted from each pose-guided MHI using a CNN such as EfficientNet [75]. Let $f_t = \text{CNN}\left(H_\tau^{PG}(x, y, t)\right)$ represent the feature vector extracted from the MHI at time t.

2. Sequence Processing: The extracted features serve as input to a transformer model, which is adapted to process sequences of these features. We use a variant of the Vision Transformer (ViT) architecture [76], modified to handle temporal sequences.

Given a sequence of T feature vectors, we first add positional encodings:

$$f_t' = f_t + \text{PE}(t)$$

where PE(t) is a positional encoding vector. The transformer then processes this sequence through L layers of multi-head self-attention and feed-forward networks:

$$z_l = \text{MSA}\left(\text{LN}(z_{l-1})\right) + z_{l-1}$$

$$z_l' = \text{MLP}\left(\text{LN}(z_l)\right) + z_l$$

where $z_l$ is the layer's output after self-attention, MSA is multi-head self-attention, LN is layer normalization, and MLP is a multi-layer perceptron.

3. Classification: A dense layer is added atop the transformer model to classify the input sequence into one of the four interaction classes:

$$y = \text{softmax}(W \cdot z_L + b)$$

where *W* and *b* are learnable parameters, and *y* is the predicted probability distribution over the four classes.

By combining PGMHA and TBSM, our Hybrid Deep Learning Framework aims to capture both the spatial and temporal aspects of mother-infant interactions, providing a comprehensive analysis of these complex behavioral patterns. This approach represents a significant advancement in the automated analysis of early childhood interactions, offering the potential for more efficient and objective assessments of maternal predictability and its impact on infant development.

The integration of pose estimation with MHIs in PGMHA allows for a more focused analysis of human movements, particularly useful in the context of mother-infant interactions where subtle movements and gestures can be significant. The use of transformers in TBSM enables the model to capture long-range dependencies in the temporal sequence, which is crucial for understanding the flow and patterns of interaction over time.

This Hybrid Deep Learning Framework addresses several challenges in the analysis of mother-infant interactions. Automation of feature extraction and classification, significantly reduces the time and resources required compared to manual coding. The framework provides a consistent, data-driven approach to analyzing interactions, reducing potential biases inherent in human observations. Once trained, the model can process large volumes of video data, enabling larger-scale studies than previously feasible. The combination of spatial (through PGMHA) and temporal (through TBSM) analysis allows the framework to capture the multi-faceted nature of human interactions.

In the following sections, we will discuss how this framework is integrated with Voice Activity Detection (VAD) for auditory stimuli analysis and computer vision techniques for visual stimuli analysis, creating a comprehensive system for studying mother-infant interactions.

## 4.3. Voice Activity Detection (VAD) for Auditory Stimuli Analysis

To analyze the auditory component of mother-infant interactions, we implemented an unsupervised Voice Activity Detection (VAD) system. This approach aims to identify and

categorize vocal exchanges between the mother and infant, focusing on detecting verbal utterances including speech and other vocalizations, without the need for labeled training data.

### 4.3.1 Audio Preprocessing and Feature Extraction

The raw audio recordings underwent preprocessing to enhance signal quality. This process involved noise reduction using spectral subtraction, audio normalization to ensure consistent volume levels, and segmentation into short-time frames of 25 milliseconds with a 10-millisecond overlap.

For each frame, we extracted a set of acoustic features known for their effectiveness in characterizing speech and non-speech sounds. The primary features included Mel-frequency cepstral coefficients (MFCCs), spectral centroid, zero-crossing rate (ZCR), and spectral flux.

MFCCs were calculated using a mel-filterbank followed by a discrete cosine transform, providing a compact representation of the spectral envelope. The spectral centroid, computed as the weighted mean of the frequencies present in the signal, offered insights into the sound's brightness. The ZCR, defined as the rate of sign-changes along a signal, helped distinguish between voiced and unvoiced speech segments. Lastly, spectral flux, measured as the frame-to-frame spectral difference, captured the dynamic characteristics of the audio signal.

### 4.3.2 Unsupervised Clustering using Gaussian Mixture Models

We employed Gaussian Mixture Models (GMMs) to cluster the extracted features into homogeneous groups. The process involved determining the optimal number of mixture components using the Bayesian Information Criterion (BIC), training a GMM on the feature set using the Expectation-Maximization algorithm, and obtaining a probabilistic clustering of the audio frames.

The BIC was computed for a range of component numbers K, as follows:

$$BIC(K) = -2\ln(L) + K\ln(n)$$

where, L is the likelihood of the data given the model, K is the number of free parameters in the model, and n is the number of observations. This approach balanced model complexity with goodness of fit to prevent overfitting while capturing the underlying data structure.

### 4.3.3 Cluster Analysis and Interpretation

Following GMM clustering, we analyzed the resulting clusters by computing summary statistics for each cluster, including the mean and variance of MFCCs, mean spectral centroid, mean ZCR, and mean spectral flux. Interpretation of cluster characteristics was based on these statistics. For instance, high mean ZCR and spectral flux were typically associated with speech or vocalization, while low spectral centroid and flux often indicated background noise or silence.

### 4.3.4 Temporal Analysis of Vocal Patterns

To understand the dynamics of vocal interactions, we performed a temporal analysis examining cluster durations, frequencies, and transitions between clusters. We computed transition probabilities between different acoustic states, allowing us to quantify various aspects of the vocal interaction, such as average duration of vocalization bouts, frequency of turn-taking, and prevalence of overlapping vocalizations.

The transition probability matrix P was defined as:

$$P_{ij} = P(X_{t+1} = j | X_t = i)$$

where $X_t$ represents the cluster state at time t, and i and j are cluster indices.

### 4.3.5 Integration with Multimodal Analysis

The VAD results were temporally aligned with data from other modalities, enabling comprehensive analysis of mother-infant interactions and exploration of relationships between vocal patterns and other forms of communication.

This unsupervised approach to VAD allows for the analysis of the auditory component of mother-infant interactions without labeled training data. While it may not achieve the same specificity as supervised approaches, it provides valuable insights into the structure and dynamics of vocal exchanges. The method's flexibility allows for adaptation to diverse datasets of mother-infant interactions, accommodating different acoustic environments and interaction styles.

## 4.4 Computer Vision Techniques for Visual Stimuli Analysis

The primary objective of our visual stimuli analysis is to determine whether the infant's gaze is directed towards objects that the mother touches during their interaction. This analysis is crucial for understanding joint attention and the infant's responsiveness to maternal cues. We utilize video footage captured from Tobii eye-tracking glasses worn

by the mother, providing a first-person perspective of the infant and the interaction environment. Our analysis employs several computer vision techniques to extract relevant information from these videos.

### 4.4.1 Object Detection and Tracking

We implemented a You Only Look Once (YOLO) based object detection model to identify and localize objects in each video frame. The model is fine-tuned on a dataset of common objects present in mother-infant interactions. For each detected object, we obtain a bounding box defined by:

$$B = (x, y, w, h, c)$$

where $(x, y)$ are the coordinates of the top-left corner, $w$ and $h$ are the width and height of the box, and $c$ is the confidence score.

We then employ a Kalman filter for object tracking across frames, which helps in maintaining object identity and handling occlusions.

### 4.4.2 Maternal Touch Detection

To identify objects touched by the mother, we combine the object detection results with hand tracking. We use Openpose and Movenet based hand key point detection models to locate the mother's hands in each frame. An object is considered "touched" if the distance between any hand keypoint and the object's bounding box is below a threshold

$$\text{Touched}(o, h) = \min_{i,j} \Delta\left(O_i, H_j\right) < \text{threshold}$$

where $O_i$ is the corner points of the object's bounding box and $H_j$ is the detected hand keypoints.

### 4.4.3 Infant Gaze Estimation

We used the same Openpose and Movenet models for infant face detection and landmark localization. From these landmarks, we estimate the infant's gaze direction. The gaze direction is represented as a unit vector:

$$\mathrm{g} = \left(g_x, g_y, g_z\right)$$

We project this gaze vector onto the 2D image plane to determine the point of regard.

### 4.4.4 Joint Attention Analysis

To assess whether the infant is looking at the object the mother touched, we compute the following metrics:

1. Gaze-Object Intersection: We calculate whether the infant's projected gaze intersects with the bounding box of the touched object. This is represented as a binary variable for each frame:

$$I(t) = \begin{cases} 1, & \text{if gaze intersects object} \\ 0, & \text{otherwise.} \end{cases}$$

2. Reaction Time: We measure the time delay between the mother touching an object and the infant's gaze shifting to that object:

$$\Delta t = t_{\text{gaze}} - t_{\text{touch}}.$$

3. Attention Duration: We calculate the duration for which the infant's gaze remains on the touched object:

$$D = \sum_{t=t_{gaze}}^{t_{gaze}+T} I(t)$$

where $T$ is the maximum duration, we consider for sustained attention.

### 4.4.5 Temporal Analysis

We perform temporal analysis to understand the dynamics of joint attention:

1. Frequency of Joint Attention: We calculate the proportion of maternal object touches that result in the infant's gaze shift to the object within a specified time window.

2. Attention Pattern: We analyze the sequence of infant gaze shifts in relation to maternal touches to identify patterns of responsiveness or anticipation.

By focusing our computer vision analysis on these specific aspects, we obtain quantitative measures of how infants visually respond to maternal object manipulation. This approach provides insights into the development of joint attention and the infant's engagement with objects in their environment as mediated by maternal behavior.

In conclusion, this methodology section has presented an integrative approach to analyzing mother-infant interactions, focusing on the predictability of maternal sensory signals. Our hybrid deep learning framework, combining PGMHA and TBSM, forms the core of our video analysis. Complemented by VAD for auditory analysis and specialized computer vision techniques for visual interactions, this approach enables a comprehensive, multimodal examination of interaction dynamics. The integration of these methods, culminating in entropy rate calculations, allows for a quantitative assessment of maternal

behavior predictability. This methodology paves the way for deeper insights into the role of predictable maternal signals in early childhood development, offering advantages over traditional observational methods in terms of objectivity and scalability.

# 5. RESULTS

This chapter presents the experimental results of our framework for analyzing mother-infant interactions, with particular emphasis on quantifying maternal sensory predictability through three key measures: tactile, auditory, and visual stimuli. While our Hybrid Deep Learning Framework has demonstrated promising results in tactile stimuli analysis, the auditory and visual components remain under development.

## 5.1 Analysis of Tactile Stimuli in Mother-Infant Interactions

Our Hybrid Deep Learning Framework, integrating PGMHA and TBSM, demonstrated robust performance in recognizing and classifying tactile interactions between mothers and infants. The framework was systematically evaluated against several established baseline models using standard performance metrics. These metrics - accuracy, precision, recall, and F1 score - were chosen for their complementary nature in assessing classification performance, particularly in scenarios involving complex human interactions. The selection of these metrics was driven by the need to comprehensively evaluate the framework's performance across different aspects of interaction classification, ensuring that the model performs reliably across various interaction scenarios.

The comprehensive evaluation process revealed compelling results, as detailed in Table 5.1. The ViViT-based model demonstrated superior performance across most metrics, achieving the highest precision of 75.3%, recall of 73%, and F1 score of 67%. This consistent performance across multiple metrics is particularly noteworthy in the context of mother-infant interaction analysis. While the 3D CNN model showed a marginally higher accuracy of 75% compared to our model's 74.8%, this slight difference becomes negligible when considering the overall performance profile. Our model's stronger performance in precision and recall indicates more reliable and consistent classification capabilities, which is particularly important when analyzing subtle interaction patterns that characterize mother-infant relationships.

The performance of our model suggests its effectiveness in recognizing and classifying different types of mother-infant interactions: holding baby, not holding baby, manipulating object, and not manipulating object. The high precision and recall values indicate that the model can accurately identify instances of each interaction class while minimizing false positives and false negatives. The superior performance of our ViViT-based model compared to traditional approaches like 3D CNNs, LSTMs, and Two Stream ConvNets highlights the benefits of our hybrid approach. The integration of PGMHA for capturing

spatial information and TBSM for modeling temporal dynamics appears to provide a more comprehensive understanding of the interaction patterns.

**Table 5.1: Performance Metrics of Different Models**

| Models | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| ViViT Model | 0.748 | 0.753 | 0.73 | 0.67 |
| 3D CNN | 0.75 | 0.375 | 0.5 | 0.42 |
| LSTM | 0.68 | 0.67 | 0.66 | 0.53 |
| Two Stream ConvNets | 0.63 | 0.62 | 0.50 | 0.58 |

To illustrate the capabilities of our framework, we present results from an example video [77]. At the time of this publication, our actual dataset is not publicly accessible and is available only upon request.


***Figure 5.1*** *Interaction Classification Examples.*

The proposed framework demonstrated robust performance in detecting key interaction points within mother-infant dyads. As evidenced in Figure 5.1, the framework correctly identified the current interactions as "Not Holding Baby" while simultaneously recognizing that the mother was "Manipulating Object" and “Not Manipulating Object” (i.e., holding or moving the object). This multi-label classification demonstrates the system's ability to discern and categorize various types of mother-infant engagement scenarios.

## 5.2 Preliminary Results for VAD and Visual Stimuli Detection

While the primary focus of this research has been on tactile interaction analysis, preliminary implementations VAD and visual stimuli detection systems have yielded promising initial results.

The VAD system has been implemented with basic functionality for detecting and segmenting vocal interactions. Figure 5.2 presents the preliminary results of the VAD system, displaying both the audio waveform and corresponding classification results. The upper panel shows the raw audio waveform, while the lower panel illustrates the system's classification of audio segments into three categories: noise (gray), infant vocalizations (red), and adult speech (blue). The system demonstrates basic capabilities in distinguishing maternal vocalizations from background noise, identifying temporal boundaries of vocal segments, and detecting overlapping vocalizations between mother and infant.

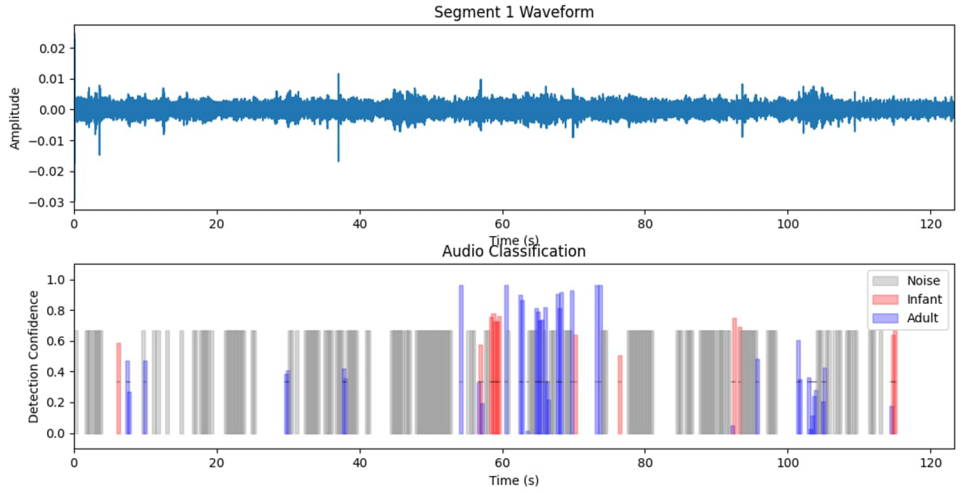


***Figure 5.2*** *Voice Activity Detection results showing audio waveform (top) and corresponding classification of audio segments (bottom)*

The visual analysis system demonstrated capability in tracking key interaction features between mother-infant dyads. Figure 5.3 illustrates the system's performance in identifying and tracking critical visual engagement markers, including infant gaze direction and maternal-infant hand positions. These interaction points, visualized through colored markers overlaid on the video frames, provide spatial reference points for analyzing engagement patterns.

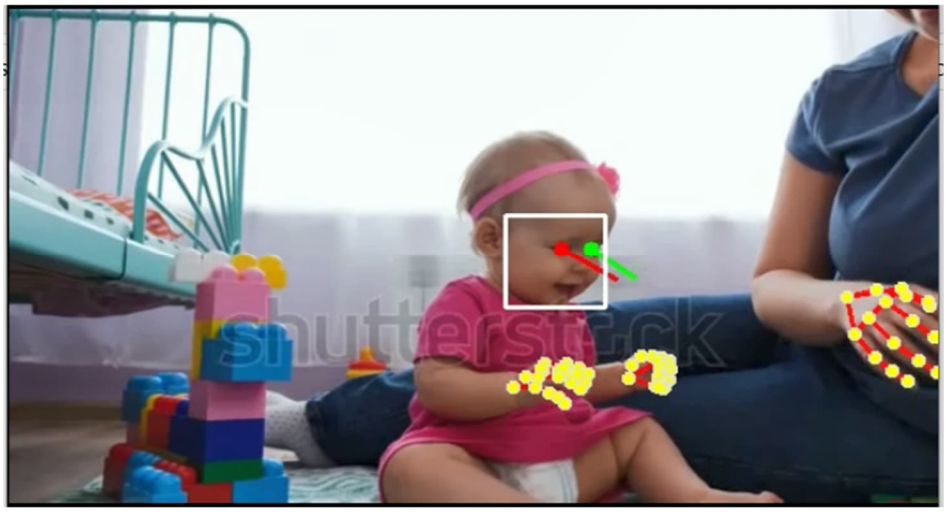

***Figure 5.3*** *Key Interaction Points Detection in Mother-Infant Interaction using Open-Pose Model.*

The implementation of the Movenet Multipose model for skeletal pose estimation proved particularly effective. As shown in Figure 5.4, the model successfully generated precise skeletal mappings for both mother and infant, enabling quantitative analysis of spatial relationships during interactions. This skeletal tracking framework provides objective measurements for analyzing interpersonal distance dynamics, postural orientations, and movement synchronization patterns between mother and infant throughout their interactions.

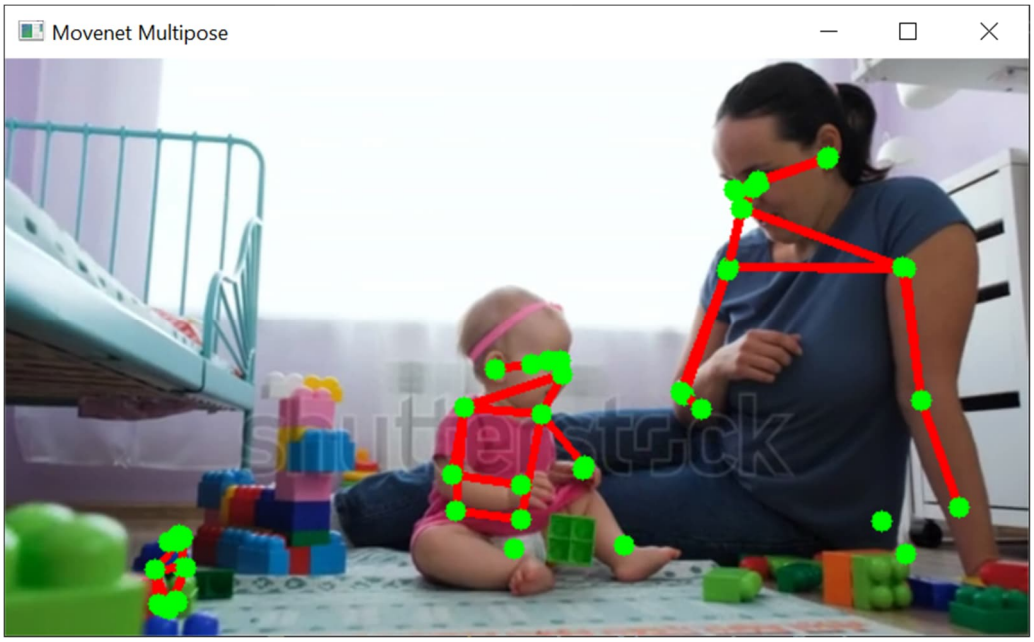


***Figure 5.4*** *Skeletal Pose Tracking using Movenet Multipose Model.*

## 5.3 Current Limitations and Future Work

While our results are promising, several aspects of the study require further development. The Voice Activity Detection (VAD) and Visual Stimuli Detection components have been implemented but not yet quantitatively assessed. Future work will focus on evaluating their performance using metrics such as accuracy, precision, recall, and F1 score for speech detection in VAD, and gaze direction accuracy, object interaction detection rate, and false positive rate for visual stimuli detection.

The entropy calculation, crucial for quantifying the predictability of maternal sensory signals, has not yet been computed due to the pending completion of voice activity detection and visual stimuli detection. This remains an important area for future work. Additionally, the current dataset, while providing valuable insights, is limited in size. Expanding the dataset will be essential for validating the model's performance across a wider range of interactions.

## 5.4 Implications and Potential Applications

The strong performance of our model in action recognition sets the stage for more advanced analyses of mother-infant interactions. Potential applications include clinical assessment for early identification of at-risk mother-infant dyads by providing objective measures of interaction quality. In the field of developmental psychology, our model could enable larger-scale longitudinal studies on the impact of early interactions on child development. Furthermore, it could be used to develop educational tools and feedback systems for parents and caregivers to enhance the quality of early childhood interactions.

## 5.5 Ethical Considerations

As we move towards more widespread application of these technologies, it is crucial to address several ethical considerations. Privacy concerns related to the collection and analysis of sensitive family data must be carefully managed. Potential biases in machine learning models and their impact on diverse populations need to be thoroughly investigated and mitigated. Additionally, the ethical implications of automated assessments of parenting behavior require careful consideration to ensure fair and beneficial use of these technologies.

# 6. CONCLUSIONS

This thesis has made significant strides towards fulfilling its primary objectives, while also laying the groundwork for future research. The key achievements and remaining challenges are summarized as follows:

1. We have developed a comprehensive framework for analyzing mother-infant interactions using advanced machine learning techniques. While we have not yet fully implemented the quantification of maternal sensory predictability, our work has established the necessary foundation for this analysis.
2. We have successfully created and evaluated a novel Hybrid Deep Learning Framework for recognizing and analyzing mother-infant interactions in video data. Our ViViT-based model, integrating PGMHA and TBSM, has demonstrated superior performance compared to existing approaches in most metrics.
3. We have explored the theoretical application of our framework for calculating the entropy rate of maternal sensory signals. However, the full implementation of this calculation, including the computation of the transition matrix, remains an area for future work.

In conclusion, this work bridges the gap between advanced machine learning techniques and early childhood development research, opening new avenues for understanding the complex dynamics of mother-infant interactions. Moving forward, the completion of VAD and visual stimuli detection components, along with the implementation of entropy rate calculations, will fully realize the potential of this research in contributing to our understanding of early childhood development. This work represents a significant step towards more comprehensive, efficient, and objective analysis of mother-infant interactions, with potential implications for both research and clinical applications in developmental psychology.

# REFERENCES

[1] N. A. C. F. Rocha, F. P. dos Santos Silva, M. M. dos Santos, and S. C. Dusing, "Impact of mother–infant interaction on development during the first year of life: A systematic review," Journal of Child Health Care, vol. 24, no. 3, pp. 365-385, 2020.

[2] T. L. Bale, T. Z. Baram, A. S. Brown, J. M. Goldstein, T. R. Insel, M. M. McCarthy, C. B. Nemeroff, T. M. Reyes, R. B. Simerly, E. S. Susser, and E. J. Nestler, "Early life programming and neurodevelopmental disorders," Biological Psychiatry, vol. 68, no. 4, pp. 314-319, 2010.

[3] M. D. S. Ainsworth, “Patterns of attachment: A psychological study of the strange situation”. Lawrence Erlbaum Associates, 1978.

[4] K. Hirsh-Pasek, L. B. Adamson, R. Bakeman, M. T. Owen, R. M. Golinkoff, A. Pace, P. K. Yust, and K. Suma, "The contribution of early communication quality to low-income children's language success," Psychological Science, vol. 26, no. 7, pp. 1071-1083, 2015.

[5] R. Feldman, "Parent–infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions," Journal of Child Psychology and Psychiatry, vol. 48, no. 3-4, pp. 329-354, 2007.

[6] L. A. Sroufe, "Attachment and development: A prospective, longitudinal study from birth to adulthood," Attachment & Human Development, vol. 7, no. 4, pp. 349-367, 2005.

[7] T.L Bale, T.Z Baram, A.S Brown, J.M Goldstein, T. R Insel, M.M McCarthy, C.B Nemeroff, T.M Reyes, R.B Simerly, E.S Susser, E. J & Nestler, "Early life programming and neurodevelopmental disorders.” in Biological Psychiatry, 2010, 68(4), 314–19.

[8] E. P. Davis, S. A. Stout, J. Molet, B. Vegetabile, L. M. Glynn, C. A. Sandman, K. Heins, H. Stern, and T. Z. Baram, "Exposure to unpredictable maternal sensory signals influences cognitive development across species," Proceedings of the National Academy of Sciences, vol. 116, no. 34, pp. 16435-16444, 2017.

[9] T. Z. Baram, E. P. Davis, A. Obenaus, C. A. Sandman, S. L. Small, A. Solodkin, and H. Stern, "Fragmentation and unpredictability of early-life experience in mental disorders," American Journal of Psychiatry, vol. 169, no. 9, pp. 907-915, 2012.

[10] B. G. Vegetabile, S. A. Stout-Oswald, E. P. Davis, T. Z. Baram, and H. S. Stern, "Estimating the entropy rate of finite Markov chains with application to behavior studies," Journal of Educational and Behavioral Statistics, vol. 44, no. 3, pp. 282-308, 2019.

[11] R. Lorenz, R. P. Monti, I. R. Violante, C. Anagnostopoulos, A. A. Faisal, G. Montana, and R. Leech, "The automatic neuroscientist: A framework for optimizing experimental design with closed-loop real-time fMRI," NeuroImage, vol. 184, pp. 186-200, 2016.

[12] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," Image and Vision Computing, vol. 60, pp. 4-21, 2017.

[13] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[14] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics, 2018, pp. 80-89.

[15] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981, pp. 674-679.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[17] G. Bertasius, H. Wang, & L. Torresani, “Is Space-Time Attention All You Need for Video Understanding?” in Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 813-824.

[18] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, J. Tighe "VidTr: Video Transformer Without Convolutions," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV, 2021, pp. 13557-13567

[19] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836-6846.

[20] J. Bowlby, Attachment and loss: Vol. 1. Attachment. New York: Basic Books, 1969.

[21] A.M Groh, R.M.P Fearon, M.H IJzendoorn, M.J Bakermans-Kranenburg, & G.I Roisman, “Attachment in the Early Life Course: Meta-Analytic Evidence for Its Role in Socioemotional Development.” in Child Development Perspectives, 2017,11(1), pp.70–76.

[22] J. Slomian, G. Honvo, P. Emonts, J-Y. Reginster, and O. Bruyère, "Consequences of maternal postpartum depression: a systematic review of maternal and infant outcomes," in Women's Health, 2019, 15.

[23] R. Feldman, "Parent-infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions," in Journal of Child Psychology and Psychiatry, 2007, 48(3-4), pp. 329-354.

[24] A.E. Bigelow, K. MacLean, and J. Proctor, "The role of joint attention in the development of infants' play with objects," Developmental Science, 2007, 7(5), pp. 518-526.

[25] E. Tronick, H. Als, L. Adamson, S. Wise, and T.B. Brazelton, "The infant's response to entrapment between contradictory messages in face-to-face interaction," in Journal of the American Academy of Child Psychiatry, 1978, 17(1), pp. 1-13.

[26] R. Abu-Zhaya, A. Seidl, and A. Cristia, "Multimodal infant-directed communication: How caregivers combine tactile and linguistic cues," Journal of Child Language, 2017, 44(5), pp. 1088-1116.

[27] R. Feldman, M. Singer, and O. Zagoory, "Touch attenuates infants' physiological reactivity to stress," Developmental Science, 2010, 13(2), pp. 271-278.

[28] C. Saint-Georges, M. Chetouani, R. Cassel, F. Apicella, A. Mahdhaoui, F. Muratori, M.C. Laznik, & D. Cohen, “Motherese in interaction: At the cross-road of emotion and cognition? (A systematic review)”. *PloS One, 8*(10), e78103, 2013.

[29] D.R. Pederson and G. Moran, "A categorical description of infant-mother relationships in the home and its relation to Q-sort measures of infant-mother interaction," Monographs of the Society for Research in Child Development, 1995, 60(2-3), pp. 111-132.

[30] P.M. Crittenden, "Abusing, neglecting, problematic, and adequate dyads: Differentiating by patterns of interaction," Merrill-Palmer Quarterly of Behavior and Development, 1981, 27(3), pp. 201-218.

[31] R. Bakeman and V. Quera, "Sequential analysis and observational methods for the behavioral sciences," Cambridge University Press, 2011.

[32] M. Mäntymaa, K. Puura, I. Luoma, R.K. Salmelin, and T. Tamminen, "Early mother-infant interaction, parental mental health and symptoms of behavioral and emotional problems in toddlers," Infant Behavior and Development, 2004, 27(2), pp. 134-149.

[33] R. Feldman, C.W. Greenbaum, and N. Yirmiya, "Mother-infant affect synchrony as an antecedent of the emergence of self-control," Developmental Psychology, 1999, 45(1), pp. 226-232.

[34] I. Bretherton and K.A. Munholland, "Internal working models in attachment relationships: Elaborating a central construct in attachment theory," in J. Cassidy and

P.R. Shaver (Eds.), Handbook of attachment: Theory, research, and clinical applications, The Guilford Press, 2008, pp. 102-127.

[35] N. Tottenham and L.J. Gabard-Durnam, "The developing amygdala: a student of the world and a teacher of the cortex," Current Opinion in Psychology, 2017, 17, pp. 55-60.

[36] N.L. Maitre, A.P. Key, O.D. Chorna, J.C. Slaughter, P.J. Matusz, M.T. Wallace, and M.M. Murray, "The dual nature of early-life experience on somatosensory processing in the human infant brain," Current Biology, 2017, 27(7), pp. 1048-1054.

[37] D.A. Abrams, T. Chen, P. Odriozola, K.M. Cheng, A.E. Baker, A. Padmanabhan, S. Ryali, J. Kochalka, C. Feinstein, and V. Menon, "Neural circuits underlying mother's voice perception predict social communication abilities in children," Proceedings of the National Academy of Sciences, 2016, 113(22), pp. 6295-6300.

[38] P. Kim, L. Strathearn, and J.E. Swain, "The maternal brain and its plasticity in humans," Hormones and Behavior, 2016, 77, pp. 113-123.

[39] I. Laptev, "On space-time interest points," International Journal of Computer Vision, vol. 64, no. 2-3, pp. 107-123, 2005.

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol. 1, pp. 886-893.

[41] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in Proceedings of the 15th ACM International Conference on Multimedia, 2007, pp. 357-360.

[42] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," International Journal of Computer Vision, vol. 103, no. 1, pp. 60-79, 2013.

[43] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732.

[44] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems, 2014, pp. 568-576.

[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489-4497.

[46] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299-6308.

[47] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in Proceedings of the European Conference on Computer Vision, 2016, pp. 20-36.

[48] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794-7803.

[49] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 203-213.

[50] C. Wei, H. Fan, S. Xie, C. Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14668-14678.

[51] G. Chen, Y.D. Zheng, J. Wang, J. Xu, Y. Huang, J. Pan, Y. Wang, Y. Wang, Y. Qia, T. Lu and L. Wang, "VideoLLM: Modeling Video Sequence with Large Language Models," arXiv preprint arXiv:2305.13292, 2023.

[52] J. Molet, K. Heins, X. Zhuo, Y. T. Mei, L. Regev, T. Z. Baram, and H. Stern, "Fragmentation and high entropy of neonatal experience predict adolescent emotional outcome," Translational Psychiatry, vol. 6, no. 1, p. e702, 2016.

[53] L. M. Glynn and T. Z. Baram, "The influence of unpredictable, fragmented parental signals on the developing brain," Frontiers in Neuroendocrinology, vol. 53, p. 100736, 2019.

[54] E. P. Davis, R. Korja, L. Karlsson, L. M. Glynn, C. A. Sandman, B. Vegetabile, E.L Kataja, S. Nolvi, E. Sinervä, J. Pelto, H. Karlsson, H.S Stern, and T. Z. Baram, "Across continents and demographics, unpredictable maternal signals are associated with children's cognitive function," EBioMedicine, vol. 46, pp. 256-263, 2019.

[55] Reindl, V., Gerloff, C., Scharke, W., & Konrad, K," Brain-to-brain synchrony in parent-child dyads and the relationship with emotion regulation revealed by fNIRS-based hyperscanning." NeuroImage, 178, 493–502, 2018.

[56] D. S. Messinger, P. Ruvolo, N. V. Ekas, and A. Fogel, "Applying machine learning to infant interaction: The development is in the details," Neural Networks, vol. 23, no. 8-9, pp. 1004-1016, 2010

[57] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, H. Rao, J. C. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye, "Decoding children's social behavior," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3414-3421.

[58] C. Yu and L. B. Smith, "Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination," PLOS ONE, vol. 8, no. 11, p. e79659, 2013.

[59] G. Pusiol, A. Esteva, S. S. Hall, M. Frank, A. Milstein and L. Fei-Fei, "Vision-based classification of developmental disorders using eye-movements," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 317-325.

[60] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 2021.

[61] I. Gordon, A. Gilboa, S. Cohen, N. Milstein, N. Haimovich, S. Pinhasi, and S. Siegman, "Physiological and behavioral synchrony predict group cohesion and performance," Scientific Reports, vol. 10, no. 1, pp. 1-12, 2020.

[62] D. S. Withanage Don, D. Schiller, T. Hallmen, S. Mertes, T. Baur, F. Lingenfelser, M. Müller, L. Kaubisch, C. Reck and E. André, “Towards automated annotation of infant-caregiver engagement phases with multimodal foundation models”. In *Proceedings of the International Conference on Multimodal Interaction (ICMI '24)*, 2024.

[63] B. M. Lester, C. F. Z. Boukydis, C. T. Garcia-Coll, W. Hole, and M. Peucker, "Infantile colic: Acoustic cry characteristics, maternal perception of cry, and temperament," Infant Behavior and Development, vol. 15, no. 1, pp. 15-26, 1992.

[64] M. Papousek, "Intuitive parenting: A hidden source of musical stimulation in infancy," in Musical beginnings: Origins and development of musical competence, I. Deliège and J. Sloboda, Eds. Oxford: Oxford University Press, 1996, pp. 88-112.

[65] A. Fernald and P. Kuhl, "Acoustic determinants of infant preference for motherese speech," Infant Behavior and Development, vol. 10, no. 3, pp. 279-293, 1987.

[66] D. Burnham, C. Kitamura, and U. Vollmer-Conna, "What's new, pussycat? On talking to babies and animals," Science, vol. 296, no. 5572, p. 1435, 2002.

[67] J. F. Cohn and E. Z. Tronick, "Mother-infant face-to-face interaction: Influence is bidirectional and unrelated to periodic cycles in either partner's behavior," Developmental Psychology, vol. 24, no. 3, pp. 386-392, 1988.

[68] D. S. Messinger, M. H. Mahoor, S. M. Chow, and J. F. Cohn, "Automated measurement of facial expression in infant-mother interaction: A pilot study," Infancy, vol. 14, no. 3, pp. 285-305, 2009.

[69] B. M. Lester, D. M. Bagner, J. Liu, L. L. LaGasse, R. Seifer, C. R. Bauer, S. Shankaran, H. Bada, R. D. Higgins and A. Das, "Infant neurobehavioral dysregulation: Behavior problems in children with prenatal substance exposure," Pediatrics, vol. 124, no. 5, pp. 1355-1362, 2009.

[70] R. Feldman, R. Magori-Cohen, G. Galili, M. Singer, and Y. Louzoun, "Mother and infant coordinate heart rhythms through episodes of interaction synchrony," Infant Behavior and Development, vol. 34, no. 4, pp. 569-577, 2011.

[71] V. Leong, E. Byrne, K. Clackson, S. Georgieva, S. Lam, and S. Wass, "Speaker gaze increases information coupling between infant and adult brains," Proceedings of the National Academy of Sciences of the United States of America, vol. 114, no. 50, pp. 13290-13295, 2017.

[72] A. Azhari, M. Lim, A. Bizzego, G. Gabrieli, M. H. Bornstein and G. Esposito, "Physical presence of spouse enhances brain-to-brain synchrony in co-parenting couples," Scientific Reports, vol. 10, no. 1, pp. 1-11, 2020.

[73] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, 2001.

[74] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. L. Chang, M. G. Yong, J. Lee, W. T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[75] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proceedings of the International Conference on Machine Learning, 2019, pp. 6105-6114.

[76] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[77] “Happy young mother and baby daughter playing together," Shutterstock, https://www.shutterstock.com/video/clip-1104907771-happy-young- mother-baby-daughter-playing-together.