Contents

List of Figures

List of Tables

Acknowledgement

At the outset, I wish to express my deep heartfelt gratitude to my dissertation guide, **Prof. Manisha Verma**, Assistant Professor, Department of Mathematics and Computing, Indian Institute of Technology (ISM) Dhanbad, for his invaluable guidance, untiring efforts, timely suggestions and constant inspiration throughout the period of our research pursuit. The knowledge and experience that I gained from him will always enlighten my life. Moreover, his open door for discussion on the problem related to my dissertation has taken my dissertation at the level of completion. Working under his supervision and guidance always remained a matter of deep satisfaction.

I am deeply appreciative of **Prof. S P Tiwari**, Head of the Department, for his unwavering commitment to providing the necessary resources and conducive environment essential for the successful execution of this project.

To my family, I owe a debt of gratitude for being my steadfast support system. Their encouragement, understanding, and unwavering belief in my abilities have been a source of strength and motivation throughout this journey. They always provided me with the moral and emotional support I needed to move forward in my life.

Finally, I express my heartfelt gratitude to God who continues to look upon me despite my many flaws. He has always guided and protected me, enabling me to overcome challenges and achieve success. I am what I am because of his unfailing love and grace toward me and I owe the greatest of gratitude to him.

Anurag (20JE0167)

Abstract

This study explores how to recognize facial expressions in children, which can be more complex than identifying emotions in adults. Children's facial expressions are often subtler and change rapidly, making it harder for existing facial expression recognition (FER) systems, mostly trained on adult data, to perform well.

To address this gap, we created a new dataset called ChildNet using web-scraped images of children's facial expressions. We also used an established dataset, TIF-DF (Tromsø Infant Faces - Developmental Faces), which contains high-quality, validated images of infant and young children's emotions. These two datasets provide a better foundation for training models specifically on children's expressions.

We studied and compared several deep learning models, including Graph Neural Networks (GNN), Graph Convolutional Networks (GCN), ResNet-50, and Siamese ResNet-50. We used the Face-Alignment library to extract facial landmarks and created various adjacency matrices, such as custom designs and Delaunay triangulation, to feed spatial information into the graph-based models.

All models were initially trained on ChildNet and TIF-DF to measure their performance on child-specific expressions. The results showed that models like ResNet50 and Siamese ResNet-50 performed better in generalization, with the Siamese ResNet-50 being particularly effective at distinguishing between similar emotions.

This research highlights the potential of GNN-based architectures and advanced convolutional models in recognizing children's emotions more accurately. With improved datasets like ChildNet and TIF-DF, and continued development of specialized models, FER systems can be better tailored for use in educational tools, therapeutic applications, and child-friendly technology.

Chapter 1

Introduction

Human Facial expressions represent one of the most fundamental and universal channels for conveying emotions, intentions, and mental states[5]. The ability to recognize and interpret facial expressions is crucial for effective social interaction, communication, and empathy development[1]. In past years, Facial Expression Recognition (FER) has been a huge topic of research in computer vision, machine learning, psychology, and human-computer interaction[10]. FER has been used in lots of industries, and it is also important for lots of applications, such as those used in the healthcare industry, education, entertainment, and security[3]. But currently existing FER systems are developed for adults, so recognizing facial expressions in children remains challenging[9], and it is also an under-explored research area.

Children's facial expressions are different from adult expressions, and they have different characteristics, such as dynamic and intense change. Some research has shown that some children's expressions are more spontaneous, dynamic, and intense, with quick changes between emotional states[15]. Moreover, children's facial morphology differs from adults because their facial features are continuously developing throughout childhood to adolescence[14]. This development trajectory of facial expression abilities in children further complicates this field. Some Studies have said, ”Children can learn new emotions from their surroundings, cultures.”

The traditional approaches to facial expression recognition have highly relied on Convolutional Neural Network (CNNs)[4], which have demonstrated impressive performance in various tasks of computer vision. These models typically process input images as grid-like structures and extract hierarchical features through convolutional operations. CNN has been effective for lots of applications in the Computer vision field, but it's failed when we use it in children’s facial expressions. It is struggling to capture the relational dynamics between facial landmarks and may require extensive training data to generalize across the variation present in children’s expressions.

Recently, in deep learning, new architectures have been introduced to solve this type of challenge. **Graph Neural Network** gives a powerful framework for modeling[11]. Those can use landmarks as nodes and spatial relationships as edges within a graph structure. This method captures the shape and structure of the face in a way that helps identify even the smallest differences in facial expressions. Similarly, Advanced CNN models like **ResNet50** perform better because they use deep residual learning, which allows them to be trained with more layers without losing accuracy or quality[6]. **Siamese networks** use a special design that compares pairs of images to learn how similar they are[8]. **Transfer Learning** is a useful technique to adapt a model, a model trained on a specific problem dataset, and fine-tune it for a similar type of problem. In our case model train on the adult facial dataset and fine-tune on the children’s dataset[16].

Children’s Facial datasets are limited compared to adult datasets because of some ethical concerns. Some datasets like CAFE[12], TIF_DF[13]. Because of a lack of data, using web scraping, I created my own dataset, ChildNet[2], which aims to solve the problem of the availability of the dataset. To solve this problem in our thesis, we use some techniques like **Transfer learning** and **Siamese networks**.

Chapter 2

Literature Review

2.1 Theoretical Framework

Facial Expression Recognition (FER) is based on psychological theories, mainly Ekman's Facial Action Coding System (FACS), which breaks expressions into muscle movements called Action Units (AUs)[5]. His universality hypothesis suggests that six basic emotions—happiness, sadness, fear, disgust, anger, and surprise—are expressed similarly across all cultures. Early facial expression recognition (FER) methods used feature engineering, where experts manually selected patterns like Gabor wavelets, LBP, and HOG to identify facial expressions based on texture and geometry.

Pathak, R. et al.(2020): Facial expressions help to understand the mood and health conditions of non-verbal groups such as babies, aiding in the early detection of issues like sleep and excessive crying disorders, which are linked to long-term developmental challenges. Traditional IoT systems for continuous monitoring are costly due to cloud infrastructure and dependency on an uninterrupted internet. Recent advancements propose low-cost IoT edge computing with multi-headed 1-dimensional CNNs, enabling local processing to classify expressions (e.g., happy, crying, sleeping) efficiently.[17]

2.2 Critical Review of Relevant Literature

2.2.1 Traditional Approaches to FER

Early FER methods used a step-by-step process: detecting faces, extracting features, and classifying emotions. Algorithms like Viola-Jones enabled automated detection, while feature extraction techniques relied on geometric and texture-based methods such as Gabor wavelets and Local Binary Patterns (LBP). Classification was performed using models like Support Vector Machines (SVMs), but these methods struggled with variations in lighting, pose, and facial differences, especially in children.[19]

2.2.2 CNN-Based Deep Learning Approaches

Deep learning transformed FER by automating feature extraction, with CNNs becoming the dominant approach. Models like VGGNet, Inception, and ResNet-50 improved recognition accuracy. However, most CNNs are trained on adult datasets, leading to biased results for children's expressions. They also rely on pixel grids, missing spatial relationships between facial landmarks.[18]

2.2.3 Graph-Based Approaches to FER

To address CNN limitations, researchers have explored graph-based models like Directed Graph Neural Networks (DGNNs)[20] and Graph Convolutional Neural Networks (GCNNs), GraphSage, and GraphAttention Networks. These models focus on facial landmarks and their spatial connections. They also improve interpretability, making them useful in education and healthcare.[7]

2.2.4 Children-Specific FER Research

Research on children's facial expressions is limited. Studies highlight age, gender, emotion type, and cultural background as factors influencing expression recognition. Children's faces change as they grow, requiring specialized models that account for these variations.

2.2.5 Transfer Learning and Domain Adaptation

Transfer learning helps improve facial expression recognition (FER) for children by fine-tuning models trained on adult data, allowing them to adapt despite facial differences. Domain adaptation techniques further refine this process by aligning feature distributions between adult and children's expressions while preserving emotion-specific details and enhancing model accuracy.

2.3 Identification of Gaps in Existing Research

1. **Limited Children-Specific Datasets** – Existing children's FER datasets are limited in scale, diversity, and coverage of emotion categories compared to adult datasets. Due to this data scarcity, to develop the children's FER-specific models, reliance on transfer learning from the adult dataset.

2. **Unexplored Graph-Based Methods** – Graph based approaches in FER are remains underexplored. The advantage of explicitly modeling the unique spatial relationship in children's expression has not been fully investigated.

3. **Limited Research on Adjacency Matrices** – For graph-based approaches, the choice of adjacency matrix representation significantly impacts performance. Yet, there are limited research papers that have discussed the construction of adjacency matrix methods for facial landmarks. Where spatial relationships differ from adults.

4. **Lack of Focus on Model Interpretability** – Most research on facial expression recognition (FER) prioritizes accuracy over understanding how models make decisions. For applications involving children, especially in education and healthcare, it's essential to know which features influence the model's choices. However, this aspect remains largely unexplored.

2.4 Addressing Research Gaps

This thesis addresses several of the identified gaps in existing research, making specific contributions to advance the field of children's facial expression recognition:

1. **Development of ChildNet Dataset**: To address the critical gap in children-specific datasets, this research introduces ChildNet, a custom dataset of children's facial expressions. While respecting ethical considerations in data collection, ChildNet provides a more comprehensive resource for training and evaluating children's FER models, complementing existing datasets like TIF_DF.

2. **Comprehensive Evaluation of Graph-Based Methods**: This research systematically explores the application of Graph Neural Networks (GNNs) and Graph Convolutional Networks (GCNs) to children's facial expressions. By explicitly modeling the spatial relationships between facial landmarks as graphs, these approaches potentially better capture the unique structural characteristics of children's expressions.

3. **Comparative Analysis of Multiple Architectures**: This study conducts a rigorous comparison of diverse architectural approaches, including GNNs, GCNs, ResNet50, MobileNetV2, and Siamese networks. This comparative analysis provides valuable insights into the relative strengths and limitations of different approaches for children's FER.

4. **Exploration of Adjacency Matrix Representations**: This study investigates various adjacency matrix construction methods for graph-based models, including manual construction, master node approaches, and Delaunay triangulation. This exploration helps identify optimal representations for capturing the spatial relationships in children's facial expressions.

5. **Analysis of Performance Across Emotion Categories**: By analyzing performance variations across different emotion categories and arousal levels, this research provides insights into the specific challenges of recognizing different emotions in children and how architectural choices affect these challenges.

2.5 Conceptual Framework Development

Based on the reviewed literature and identified gaps, this research develops a conceptual framework for children's facial expression recognition that integrates multiple perspectives and approaches. The framework is built around several key components:

1. **Developmental Sensitivity**: The framework acknowledges the developmental trajectory of facial expressions throughout childhood, recognizing that expression production and recognition evolve with age.

2. **Transfer Learning Foundation**: Given the limitations in children-specific datasets, the framework leverages transfer learning from adult datasets as a foundational strategy. Pre-trained models on large adult datasets provide a starting point that is then adapted to the unique characteristics of children's expressions.

3. **Graph Representation Flexibility**: For graph-based approaches, the framework incorporates flexible adjacency matrix representations that can be optimized for different emotion categories and age groups. This flexibility acknowledges that the most informative spatial relationships may vary across emotional expressions.

4. **Emotion-Specific Analysis**: Rather than treating all emotions uniformly, the framework adopts an emotion-specific approach that recognizes the unique challenges

associated with different emotion categories. This granular analysis informs targeted improvements for emotions that prove particularly challenging to recognize in children.

5. **Application-Oriented Evaluation**: Beyond classification accuracy, the framework incorporates evaluation criteria relevant to intended applications in educational, clinical, and human-computer interaction contexts. These criteria include interpretability, computational efficiency, and robustness to real-world conditions.

Chapter 3

Methodology

3.1 Dataset Overview

This research uses multiple datasets to develop and evaluate facial expression recognition (FER) models for children. Each dataset serves a specific purpose in the research pipeline, with some used for model training and others for validation and testing. The following sections describe the key datasets used in this study, along with their characteristics, preprocessing steps, and augmentation techniques.

3.1.1 ChildNet Dataset

The ChildNet dataset was created specifically for this research to address the significant gap in publicly available children's facial expression datasets. This dataset is collected through web scraping from the internet. This dataset contains images of children aged 1-16 years.[2]

Data Collection Process The collection process for ChildNet involved several steps:

1. **Web Scraping**: Using Python-based web scraping tools, images were collected from publicly available sources using search queries related to children's emotions and facial expressions.

2. **Age Filtering**: Only images depicting children in the 1-16 age range were retained, as determined by both automated age estimation tools and manual verification.

3. **Expression Categories**: Images were collected across seven basic emotion cate-

gories: happiness, sadness, anger, fear, disgust, surprise, and neutral.

4. **Ethical Considerations**: To ensure ethical compliance, only publicly available images were collected, with no personal identifiers retained. Additionally, the dataset was developed solely for research purposes.

Dataset Composition The final ChildNet dataset comprises:

- Approximately 442 facial images of children
- Distribution across 7 emotion categories
- Age range of 1-16 years
- Diversity in gender, ethnicity, and lighting conditions
- Images with various poses and facial orientations

Preprocessing Steps Due to the diverse sources of the collected images, extensive preprocessing was required to standardize the data for model training:

1. **Face Detection**: Each image was processed using a face detection algorithm to locate and extract the facial region. The Face-Alignment and Dlib library's face detector was employed for this purpose, providing robust detection across various poses and lighting conditions.

2. **Cropping**: Detected faces were cropped to focus only on the facial region, removing background elements that might introduce noise or bias into the model training process.

3. **Resizing**: All cropped facial images were resized to a uniform dimension of 224×224 pixels, ensuring consistency across the dataset regardless of the original image dimensions.

4. **Color Normalization**: Images were converted to RGB format with pixel values normalized to the range [0,1] to facilitate model training.

5. **Landmark Detection**: Facial landmarks were detected using Face-alignment and Dlib's 68-point facial landmark detector, providing key points for graph-based models.

3.1.2 Tromsø Infant Faces Dataset (TIF_DF)

The Tromsø Infant Faces Dataset (TIF_DF) was utilized for experiments, providing valuable data on infant facial expressions. This dataset complements ChildNet by focusing on a younger age group (0-12 months), allowing the investigation of emotion recognition across the developmental spectrum.

Dataset Characteristics TIF_DF offers several important features:

- 119 high-quality images of 19 infants (0-12 months)
- Six facial expressions: happiness, sadness, anger, fear, disgust, and surprise, neutral
- Controlled imaging conditions with standardized lighting
- Ethnically homogeneous sample (Norwegian infants)
- Validated by expert raters for expression clarity and intensity

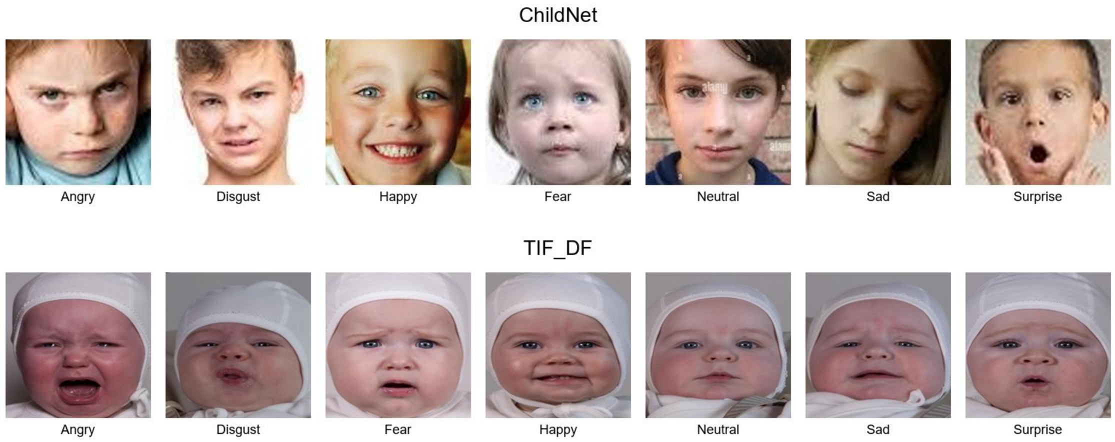

Figure 3.1: ChildNet and TIF_DF Samples

3.2 Dataset Comparative Analysis

To understand the unique characteristics of children's facial expressions compared to adults, we conducted a comparative analysis across the datasets used in this study. This analysis revealed several important differences:

- **Facial Morphology**: Children's faces in ChildNet and TIF_DF exhibited different proportions compared to adult faces in FER2013 and AffectNet, with larger eyes and foreheads relative to lower facial features.

- **Expression Intensity**: Children's expressions, particularly in ChildNet, were generally more intense and uninhibited compared to adult expressions in FER2013 and AffectNet.
- **Expression Variability**: Within the same emotion category, children showed greater variability in expression manifestation, potentially due to less developed social display rules.
- **Landmark Distribution**: The spatial distribution of facial landmarks differed between children and adults, with relatively larger distances between eye and mouth landmarks in children.

These observed differences underscore the importance of developing specialized models for children's facial expression recognition rather than simply applying adult-trained models. The unique characteristics of children's expressions necessitate both specialized datasets like ChildNet and TIF_DF and appropriate modeling approaches that can capture these distinctive patterns.

3.3 Model Architectures

This research employs multiple deep learning architectures to comprehensively evaluate different approaches to children's facial expression recognition. Each architecture offers distinct advantages and represents different paradigms in deep learning for computer vision. The following sections detail the models implemented in this study.

3.3.1 Graph-based Method: Graph Convolutional Network (GCN)

Graph Convolutional Networks (GCNs) are leveraged in this study to model facial landmarks as graph-structured data, effectively capturing both local and global dependencies through message passing. GCNs are particularly suited for structured representations like facial keypoints, where topological relationships between features are more informative than raw pixel values.

Graph Construction Facial landmark graphs are constructed by:

- Detecting 68 facial landmarks using Face-Alignment and Dlib's facial landmark detector.

- Representing each landmark as a node, initialized with spatial coordinates and localized features.
- Defining edges based on anatomical connectivity (e.g., jawline, eyebrows, eyes, nose, and mouth), forming a sparse adjacency matrix shared across samples.

Architecture Overview The GCN model used in this study consists of:

- **Input Layer**: Applies a graph convolution to the node feature input.
- **Message-Passing Layers**: Six intermediate GCN layers with batch normalization, ReLU activations, and dropout, each connected via residual connections to stabilize learning.
- **Final GCN Layer**: Processes the accumulated features from previous layers.
- **Global Pooling**: Combines max and mean pooling across nodes to form a holistic graph-level representation.
- **Fully Connected Layers**: Two dense layers convert the pooled features into the final emotion class prediction.

Mathematical Formulation The GCN layers follow the formulation:

$$H^{(l+1)} = \text{ReLU}\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3.1}$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-connections, $\tilde{D}$ is the degree matrix, $H^{(l)}$ is the activation at layer l, and $W^{(l)}$ is the layer-specific trainable weight matrix. After the final GCN layer, global mean and max pooling are applied:

$$H_{\text{global}} = \text{concat}\left(\text{meanpool}(H), \text{maxpool}(H)\right) \tag{3.2}$$

The output is then passed through two linear layers with ReLU and dropout before classification.

Adaptation for FER For children's facial expression recognition:

- Landmark graphs are consistent across samples, enabling efficient batch processing.
- Node features incorporate both geometry and localized texture cues.

- The output layer size is adapted to match the number of emotion classes (7).

3.3.2 CNN-based Method: ResNet50

Model Overview ResNet50, a 50-layer deep residual network, was adopted in this study as a strong convolutional baseline for children's facial expression recognition (FER). Its deep architecture with skip connections allows efficient training of very deep networks by addressing the vanishing gradient problem.

Architecture Overview The architecture is composed of:

- **Initial Layers**: A 7×7 convolutional layer with stride 2, followed by batch normalization, ReLU activation, and a 3×3 max pooling layer with stride 2.
- **Residual Blocks**: Four stages of residual blocks containing convolutional and identity mappings, with channel dimensions of 64, 128, 256, and 512, respectively. Each residual block includes bottleneck units of the form $1 \times 1 \rightarrow 3 \times 3 \rightarrow 1 \times 1$.
- **Global Average Pooling**: The spatial feature maps are reduced to a single vector using global average pooling.
- **Fully Connected Head**:
  - A fully connected layer reducing the feature dimension to 512
  - ReLU activation
  - Dropout layer with rate 0.5
  - Final linear layer mapping to the number of emotion classes

Adaptation for FER To adapt the standard ResNet50 for facial expression recognition:

- The input layer was retained to accept RGB images of size $224 \times 224 \times 3$.
- The final fully connected layer was replaced to output probabilities over 7 emotion categories.
- Transfer learning was applied by initializing with ImageNet-pretrained weights. Early layers were frozen except for the last 30 parameter groups to allow fine-tuning of higher-level features.

- Dropout was included before the final layer to mitigate overfitting.

3.3.3 Siamese Network Architecture: Siamese ResNet50

Overview The Siamese network consists of two identical branches based on ResNet-50, designed to compute the similarity between two input images. Each branch extracts high-level features, and a final fully connected head computes a similarity score from the absolute difference of these features. The architecture builds on a reusable base class for training and evaluation logic.

SiameseBase Class

The `SiameseBase` class defines the essential methods for training and evaluating Siamese models:

- **Training Step:** For a batch containing image pairs (x_1, x_2) and labels y, the model outputs a similarity score $\hat{y} = f(x_1, x_2)$. Binary cross-entropy loss is computed as:

$$\mathcal{L}_{\text{bce}} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

- **Validation Step:** Computes loss and classification accuracy by thresholding the predicted score at 0.5:

$$\hat{y}_{\text{bin}} = \begin{cases} 1 & \text{if } \hat{y} > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- **Epoch-End Methods:** Aggregate loss and accuracy over all batches to compute validation statistics and log training progress.

SiameseResNet50 Architecture

The `SiameseResNet50` network uses a pretrained ResNet-50 model as its backbone:

- **Feature Extractor:** ResNet-50 is truncated before the final fully connected layer. All layers are initially frozen, and only the final residual blocks (`layer3` and `layer4`) are unfrozen for fine-tuning.

- **Forward Pass:**

  - Each image is passed through the same feature extractor to obtain embeddings $f(x_1)$ and $f(x_2)$.

- The absolute difference $\Delta = |f(x_1) - f(x_2)|$ is computed.
- This difference is passed through a fully connected network:

  $$\text{FC Layers: } \Delta \to \text{ReLU} \to \text{Dropout} \to \text{Sigmoid}$$

  to obtain a similarity score in $[0, 1]$.
- **Similarity Head:** The final MLP architecture used for pairwise similarity prediction is as follows:

  Input: Concatenated feature vector of size 2048
  $\to$ Linear(2048, 512)
  $\to$ ReLU
  $\to$ Dropout(0.5)
  $\to$ Linear(512, 256)
  $\to$ ReLU
  $\to$ Dropout(0.3)
  $\to$ Linear(256, 1)
  $\to$ Sigmoid (outputs similarity score in [0,1])

This design allows the Siamese network to learn fine-grained differences between feature vectors, making it suitable for pairwise comparison tasks such as face or expression similarity classification.

3.4 Adjacency Matrix Construction Methods

For graph-based models (GNN and GCN), the choice of adjacency matrix is crucial as it defines the connectivity pattern between facial landmarks. This research explored multiple adjacency matrix construction methods to identify optimal representations for children's facial expressions.

3.4.1 Manual Adjacency Matrix Construction

This approach used domain knowledge about facial anatomy to define connections between landmarks:

- **Facial Components**: Landmarks were grouped by facial components (jawline, eyebrows, eyes, nose, mouth)
- **Within-Component Connections**: Landmarks within the same facial component were connected sequentially
- **Between-Component Connections**: Strategic connections between components were established based on their anatomical relationships
- **Mathematical Representation**: The adjacency matrix A was defined as:

$$A_{ij} = \begin{cases} 1, & \text{if landmarks } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

3.4.2 Master Node Adjacency Matrix

This method designates a central landmark (typically the nose tip, landmark 31) as a ”master node” connected to all other landmarks:

- **Master Node**: The nose tip landmark was chosen as it remains relatively stable across expressions
- **Star Topology**: The master node connects to all other nodes, creating a star-shaped graph
- **Mathematical Representation**:

$$A_{ij} = \begin{cases} 1, & \text{if } i = 31 \text{ or } j = 31 \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

3.4.3 Delaunay Triangulation-Based Adjacency Matrix

Delaunay triangulation was used to construct edges based on geometric principles automatically:

- **Triangulation Algorithm**: For the set of landmark points, a triangulation was computed such that no point is inside the circumcircle of any triangle

- **Edge Extraction**: Edges were derived from the resulting triangulation, connecting landmarks that form the sides of triangles

- **Adaptation to Expression Changes**: This method automatically adjusts the graph structure based on the geometric arrangement of landmarks, potentially adapting to expression-specific deformations

- **Mathematical Representation**:

$$A_{ij} = \begin{cases} 1, & \text{if landmarks } i \text{ and } j \text{ are connected by Delaunay triangulation} \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

3.4.4 Distance-Based Thresholding

This method created edges based on Euclidean distances between landmarks:

- **Distance Computation**: The Euclidean distance between each pair of landmarks was calculated

- **Thresholding**: Connections were established between landmarks whose distance was below a threshold τ

- **Mathematical Representation**:

$$A_{ij} = \begin{cases} 1, & \text{if } \|p_i - p_j\|_2 < \tau \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

  where p_i and p_j are the 2D coordinates of landmarks i and j, and τ is a threshold parameter determined empirically.

Each adjacency matrix construction method was systematically evaluated for its effectiveness in capturing the structural information relevant to children's facial expressions. The comparative analysis of these methods provided insights into the optimal graph representation for different emotion categories and age groups.

Chapter 4

Experiments and Results

4.1 Experimental Setup

The performance of the proposed models was evaluated using two child-centric facial expression datasets: *TIF-DF* and *ChildNet*. To assess classification quality, we employed standard evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrices. These metrics provided a detailed view of model behavior across various emotion classes. The experiments compared the performance of graph-based neural networks (GNNs) and convolutional neural network (CNN) architectures, including a Siamese CNN variant designed for pairwise similarity learning.

4.1.1 Training Strategy Overview

To overcome the inherent challenges of facial expression recognition in children, a multi-stage training strategy was adopted. Initially, models were trained on adult facial expression datasets to learn general facial representations. These pretrained models were then fine-tuned on child-specific datasets to adapt to the distinct morphological and expressive characteristics of children. This transfer learning approach helped mitigate the limitations caused by the scarcity of large-scale pediatric emotion datasets and facilitated better feature generalization and specialization.

4.1.2 Model-Specific Training Parameters

Graph Convolutional Network (GCN)

- Optimizer: Adam with a learning rate of 0.001

- Batch Size: 32; Training Epochs: 100 with early stopping
- Weight Decay: 5×10^{-4}
- Regularization Techniques: Batch Normalization and Dropout (0.2 in intermediate layers, 0.3 in the output layer)

ResNet50

- Architecture: Pre-trained ResNet50 with a total of 24,560,711 parameters; 15,492,615 parameters were trainable
- Optimizer: Adam, learning rate = 0.001
- Batch Size: 32; Epochs: 50 with early stopping
- Weight Decay: 0.01; Gradient Clipping: 0.1

Siamese ResNet50 Network

- Base Architecture: ResNet50 with ImageNet pre-trained weights
- Total Parameters: 59,324,481; Trainable Parameters: 16,145,409
- Training Data: Balanced pairs of similar (positive) and dissimilar (negative) samples
- Optimizer: Adam, learning rate = 0.0001
- Batch Size: 32; Training Epochs: 40
- Loss Function: Binary Cross-Entropy
- Distance Function: Absolute difference
- Hard Negative Mining: Performed every 5 epochs to enhance pairwise discrimination

4.1.3 Performance Monitoring

Model selection during training was driven by validation performance. The following criteria were used:

- Overall validation accuracy
- Per-class and macro-averaged F1 scores
- Confusion matrices to visualize class-specific errors
- Multi-class ROC-AUC for assessing probabilistic prediction quality

4.2 Results

4.2.1 TIF_DF Dataset Results

Performance of Graph-Based Neural Networks

1. GCN with Delaunay Triangulation: The GCN using Delaunay triangulation achieved an accuracy of **38%**. It performed best on the *Happiness* category, achieving a high recall of 0.86 and an F1-score of 0.57. Some performance was also observed in the *Sad* class (F1-score: 0.67). However, other categories such as *Angry*, *Disgust*, and *Fear* were not correctly classified, indicating poor generalization across underrepresented emotions.

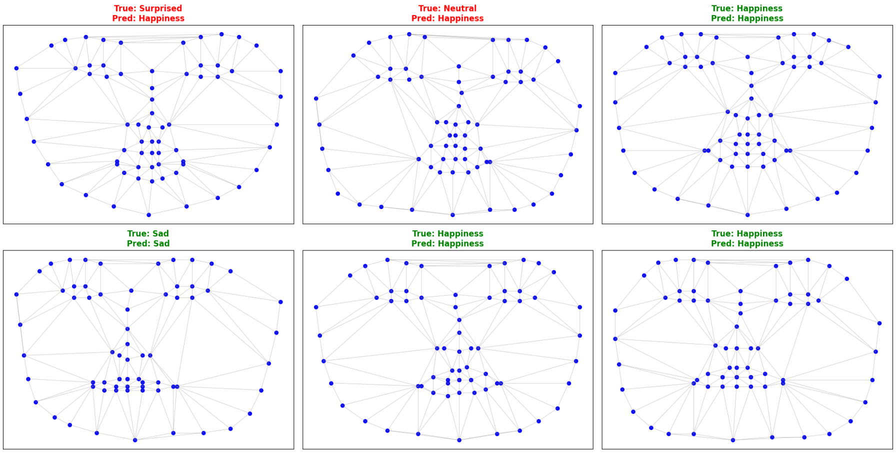

Figure 4.1: Result of GCN with Delaunay Triangulation Method (TIF_DF)

2. GCN with Delaunay + Master Node: Introducing a master node to the Delaunay graph slightly decreased the performance, yielding an accuracy of **33%**. The model correctly identified only the *Happiness* class with perfect recall but failed across all other classes. This suggests that the global node did not improve representation in low-sample conditions.

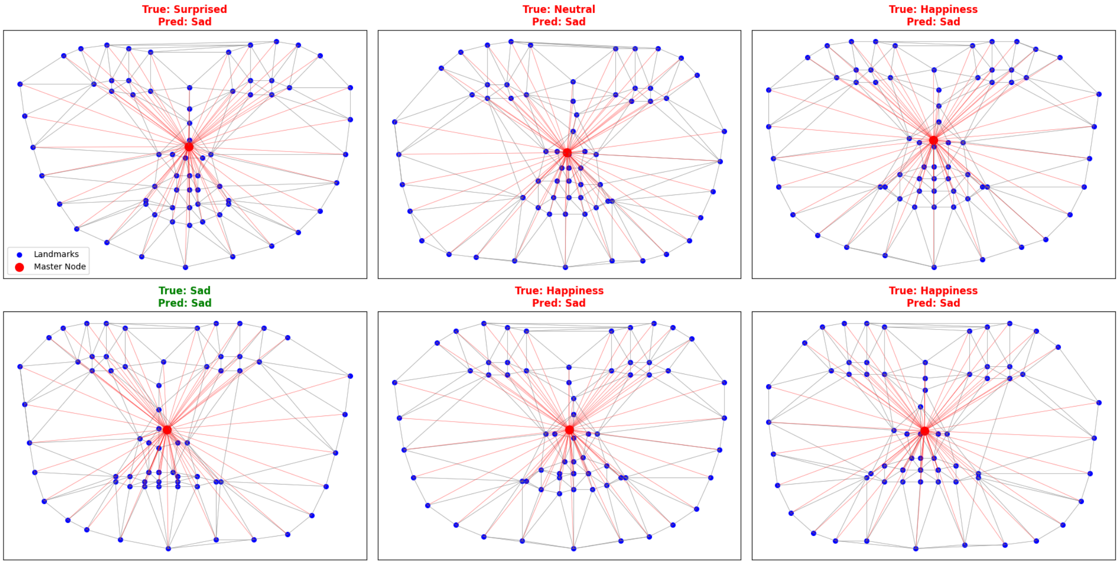

Figure 4.2: Result of GCN with Delaunay Triangulation and Master Node Method (TIF_DF)

3. GCN with Custom Adjacency Matrix: With a manually constructed adjacency matrix, the model achieved an accuracy of **38%**. It showed strong results on *Happiness* (F1-score: 0.56) and some recognition in the *Sad* category (F1-score: 0.33). However, other classes were completely misclassified, likely due to limited training examples or insufficient structural variation.

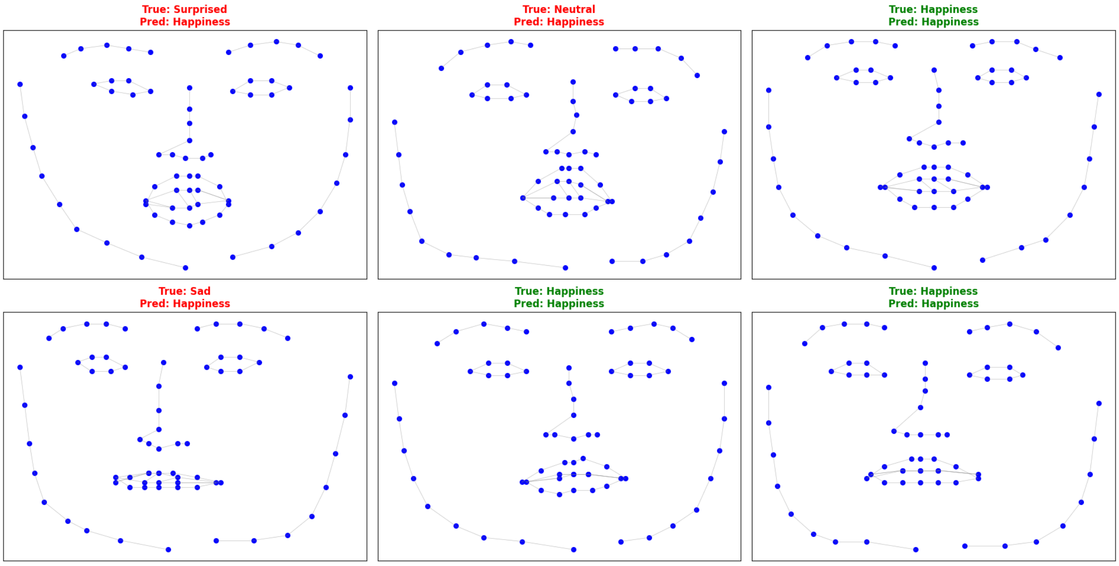

Figure 4.3: Result of GCN with Custom Adjacency Matrix Method (TIF_DF)

4. GCN with Distance-Based Thresholding: Using Euclidean distance to define graph connections led to an accuracy of **33%**. The *Happiness* class again stood out (F1-score: 0.50), but similar to other configurations, the model failed to generalize to most other emotions. This reinforces the sensitivity of GCNs to the structure and quality of the input graph, especially in small-sample scenarios.

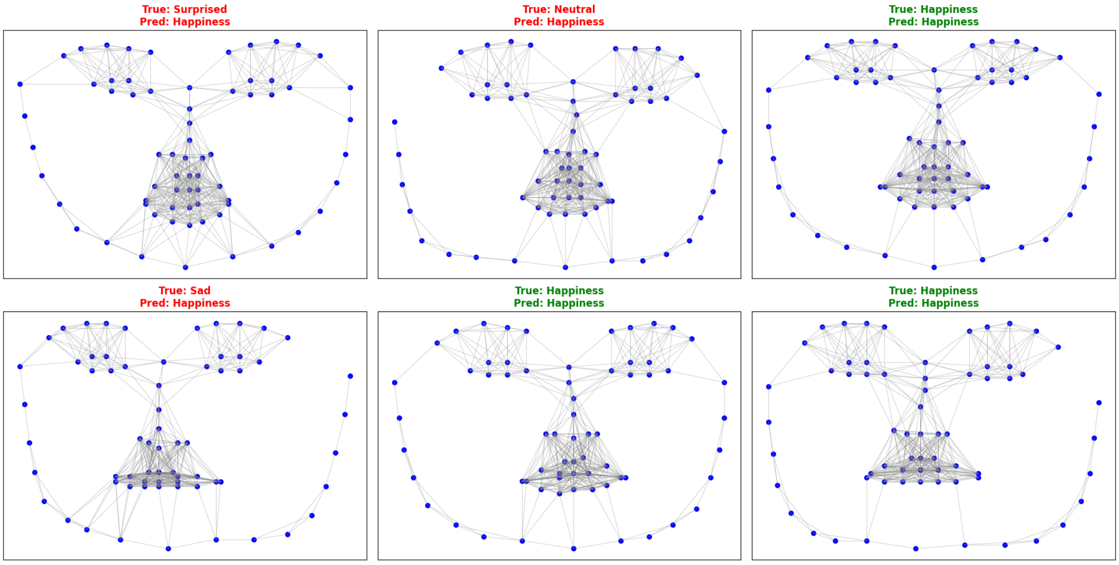

Figure 4.4: Result of GCN with Distance-Based Thresholding Method (TIF_DF)

Performance of CNN-Based Models

1. ResNet50: The fine-tuned ResNet50 model achieved the best performance among all tested methods, with a test accuracy of **55.17%**. It performed well on multiple classes, especially *Happiness* (F1-score: 0.77), *Neutral* (F1-score: 0.75), and *Sad* (F1-score: 0.55). This reflects the advantage of convolutional architectures in extracting pixel-level spatial features for emotion recognition tasks.

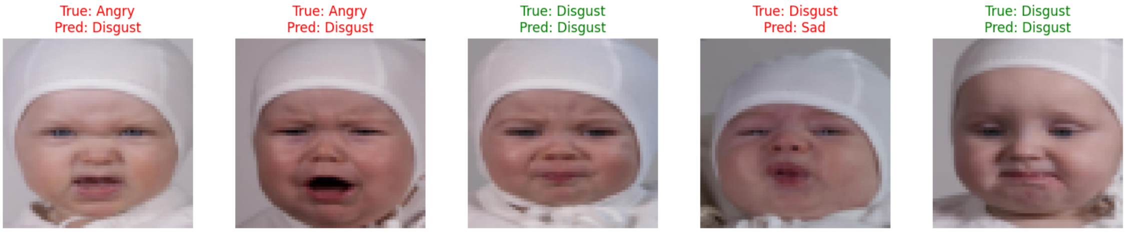

Figure 4.5: Result of ResNet50 (TIF_DF)

Performance of Siamese Network

1. Siamese ResNet50: A pair-wise similarity analysis was carried out to examine the model's ability to distinguish between facial expressions depicting the same or different emotions. The model exhibited high accuracy on most cross-emotion pairs, such as *Angry-Happiness*, *Disgust-Angry*, and *Neutral-Happiness*, each correctly identified with a success rate of 100%. This suggests the model is capable of recognizing visual differences between distinct emotional categories.

However, the model consistently failed to detect similarity within the same emotion class. Examples include *Disgust-Disgust*, *Fear-Fear*, *Happiness-Happiness*, and others, all resulting in 0% accuracy. This highlights a limitation in the model's ability to generalize intra-class emotional features.

Out of a total of **31** evaluated pairs, the model correctly predicted **19**, yielding an overall accuracy of **61.29%**. These results emphasize the model's strength in distinguishing different emotions but also underline its difficulty in recognizing consistent emotional traits across visually varied instances of the same emotion.

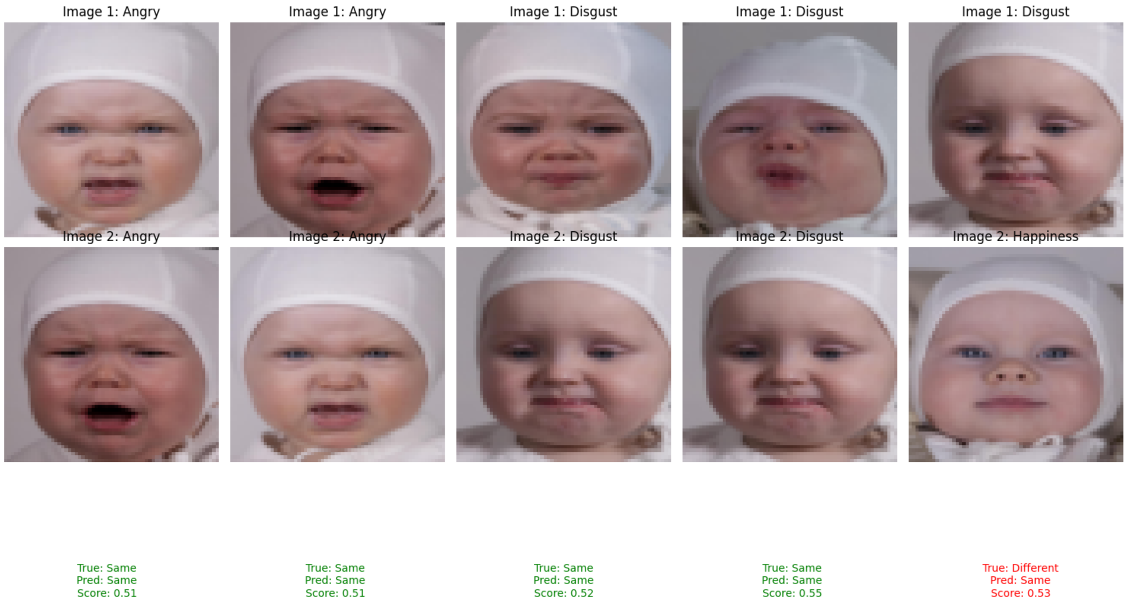

Figure 4.6: Result of Siamese ResNet50 (TIF_DF)

Table 4.1: Summary of Model Performance on TIF_DF Dataset

Model / Method	**Acc.**	**Top F1 Class(es)**	**Key Observations**
GCN (Delaunay Triangulation)	38%	Sad(0.67), Happiness(0.57)	Strong recall for Happiness; failed on Angry, Disgust, Fear
GCN (Delaunay + Master Node)	33%	Happiness(1.00)	Only Happiness correctly predicted; all other classes misclassified
GCN (Custom Adjacency Matrix)	38%	Happiness(0.56), Sad(0.33)	Slight generalization improvement; poor across underrepresented emotions
GCN (Distance-Based)	33%	Happiness(0.50)	Consistently recognizes Happiness; fails to generalize other emotions
ResNet50 (CNN)	55.17%	Happiness(0.77), Neutral(0.75), Sad(0.55)	Best performing model overall; captures rich spatial features
Siamese ResNet50	61.29%	Cross-emotion pairs(100%)	Excels at distinguishing emotion pairs; 0% for same-emotion recognition

4.2.2 ChildNet Dataset Results

Performance of Graph-Based Neural Networks

1. GCN with Delaunay Triangulation: When the adjacency matrix was constructed using Delaunay triangulation, the GCN maintained the same accuracy of **32.48%**. The *Happiness* class again showed the highest recall (0.78) and a relatively strong F1-score of 0.46. Although slight improvements were observed in precision for some classes (e.g., *Disgust*: 0.40), recognition performance for *Fear* remained at 0.00.

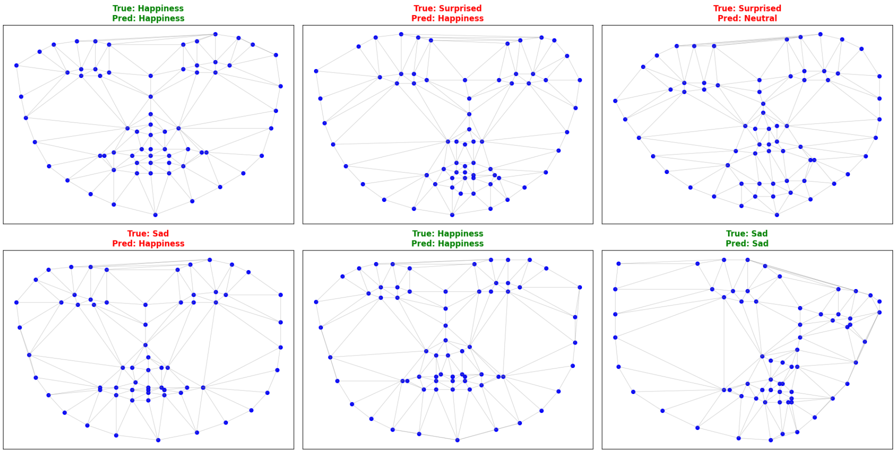

Figure 4.7: Result of GCN with Delaunay Triangulation Method (ChildNet)

2. GCN with Delaunay + Master Node: Introducing a global master node into the Delaunay graph structure improved the model's performance, achieving a test accuracy of **39.60%**. This configuration enhanced the classification of *Angry* (F1-score: 0.47), *Happiness* (0.50), and *Surprised* (0.42). While recall scores improved for multiple classes, *Fear* and *Disgust* continued to present challenges, with F1-scores of 0.00, suggesting further work is needed to improve representation of less distinct emotions.

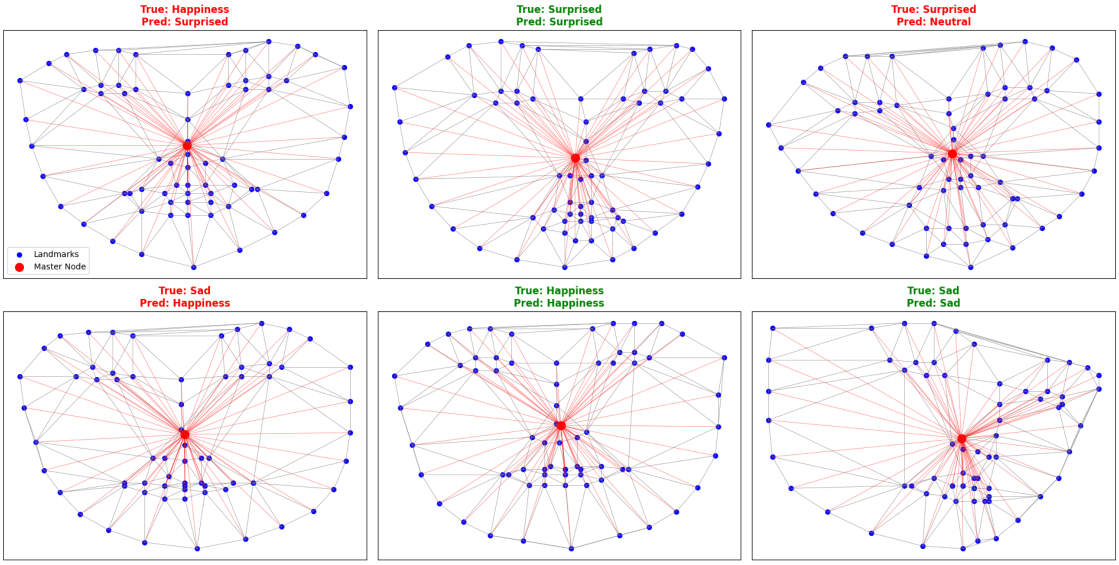

Figure 4.8: Result of GCN with Delaunay Triangulation and Master Node Method (ChildNet)

3. GCN with Custom Adjacency Matrix: Using a manually defined adjacency structure, the GCN achieved a test accuracy of **32.48%**. It showed the best recognition performance for the *Happiness* class (F1-score: 0.43) and moderate results for *Surprised* (F1-score: 0.33) and *Neutral* (F1-score: 0.30). However, the model failed to classify the *Fear* category correctly, recording an F1-score of 0.00. This outcome indicates limitations in generalizing across underrepresented or visually subtle emotions.

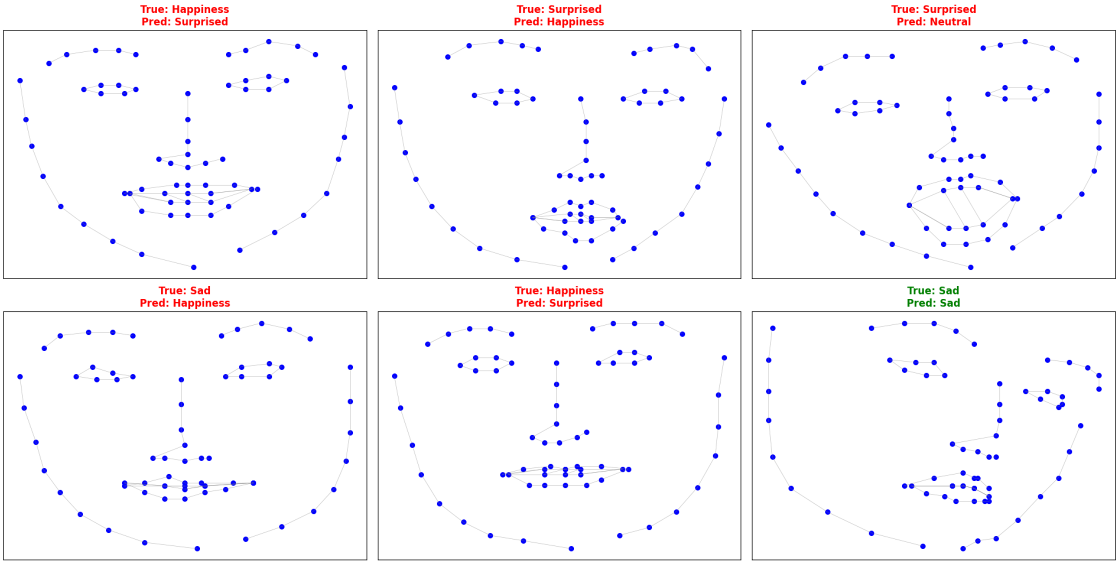

Figure 4.9: Result of GCN with Custom Adjacency Matrix Method (ChildNet)

4. GCN with Distance-Based Thresholding: Using Euclidean distances between key points to construct the adjacency matrix led to a test accuracy of **38.46%**. The model performed best on the *Happiness* (F1-score: 0.48) and *Surprised* (F1-score: 0.46) categories. Although this method increased classification balance across emotions compared to others, *Disgust* and *Fear* remained poorly predicted, both yielding F1-scores of 0.00.

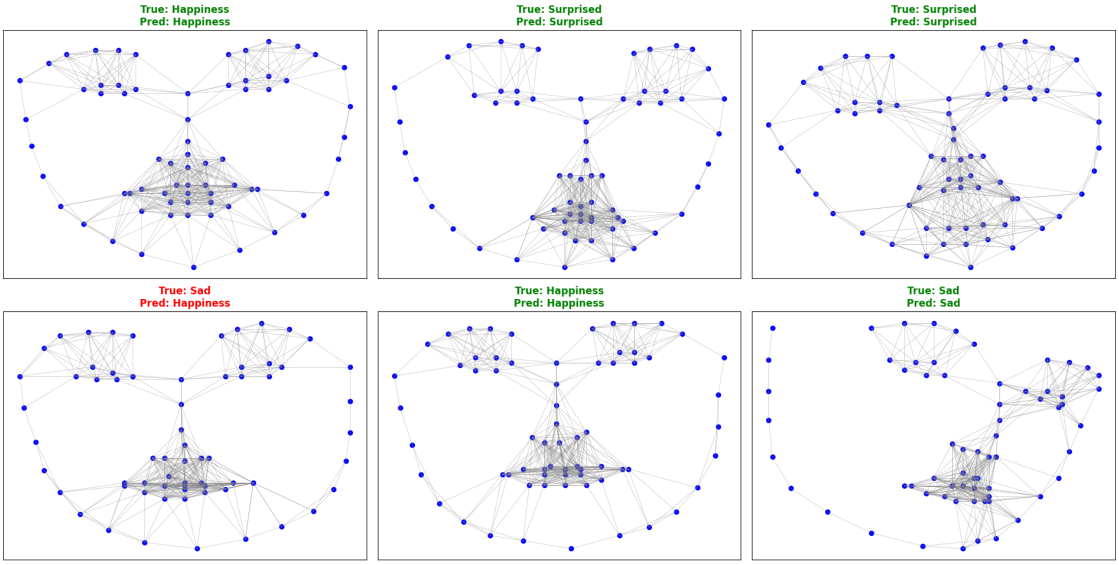

Figure 4.10: Result of GCN with Distance-Based Thresholding Method (ChildNet)

Performance of CNN-Based Models

1. ResNet50: The model achieved a test accuracy of **46.15%** on the emotion classification task. It performed best in recognizing the **Surprised** category, achieving an F1-score of **0.60**, followed by **Happiness** (F1-score: **0.49**) and **Neutral** (F1-score: **0.46**). While moderate performance was observed for some emotions, the model struggled significantly with the **Fear** category, resulting in an F1-score of **0.00**, indicating that none of the fear instances were correctly classified.

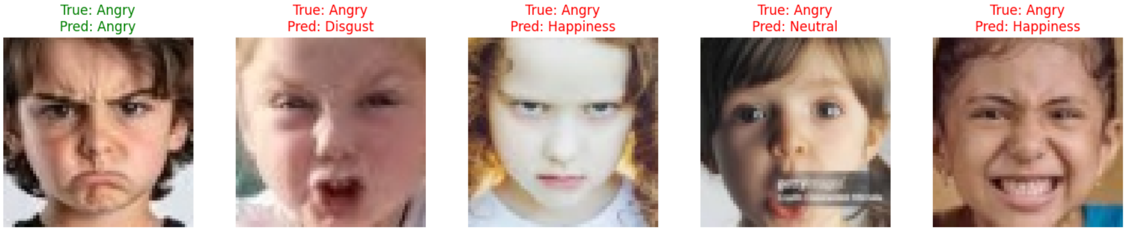

Figure 4.11: Result of ResNet50 (ChildNet)

Performance of Siamese Network

1. Siamese ResNet50: A pair-wise emotion similarity evaluation was conducted to assess the model's ability to determine whether two facial expressions represent the same emotion. The model performed well on different-emotion pairs, achieving **100% accuracy** in most cases. However, it consistently failed to recognize same-emotion pairs, with **0% accuracy** in categories such as Angry-Angry, Happy-Happy, and Sad-Sad.

Out of a total of **91** comparisons, **44** were correctly classified, resulting in an overall accuracy of **48.35%**. These results indicate that the model is more effective at identifying emotional differences than similarities within the same category.

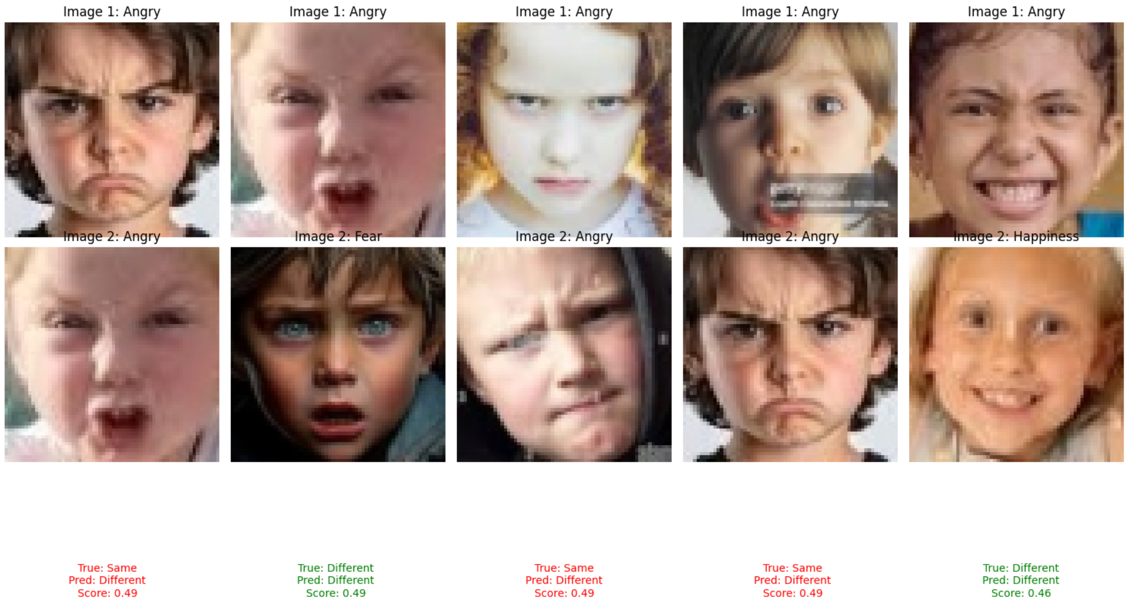

Figure 4.12: Result of Siamese ResNet50 (ChildNet)

Table 4.2: Summary of Model Performance on ChildNet Dataset

Model / Method	**Acc.**	**Top F1 Class(es)**	**Key Observations**
GCN (Delaunay)	32.48%	Happiness(0.46), Disgust(0.40)	Strong on Happiness; Fear misclassified (F1: 0.00)
GCN (Delaunay + Master)	39.60%	Happiness(0.50), Angry(0.47), Surprised(0.42)	Best GCN result; Fear and Disgust remain unrecognized
GCN (Custom Adj. Matrix)	32.48%	Happiness(0.43), Surprised(0.33), Neutral(0.30)	Fear not detected; slight gains for Neutral/Surprised
GCN (Distance-Based)	38.46%	Happiness(0.48), Surprised(0.46)	Balanced results; continued poor performance on Disgust and Fear
ResNet50 (CNN)	46.15%	Surprised(0.60), Happiness(0.49), Neutral(0.46)	Best overall accuracy; Fear still 0.00 F1-score
Siamese ResNet50	48.35%	Cross-emotion pairs (100%)	Strong at cross-emotion distinction; fails on same-emotion detection

Chapter 5

Conclusion and Summary

5.1 Conclusions

This study conducted a comparative evaluation of various facial expression recognition models using two child-focused datasets, *TIF-DF* and *ChildNet*. The analysis spanned across three main model families: Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), and Siamese architectures, each offering distinct strengths and limitations in handling emotion classification tasks.

Graph-based models, while offering interpretable spatial relationships through graph structures, exhibited limited performance—particularly with subtle or infrequently represented emotions. GNNs were generally effective in detecting highly expressive emotions such as *Happiness*, but consistently underperformed on less pronounced emotions like *Fear* and *Disgust*. Attempts to enhance performance using techniques such as master nodes and custom adjacency matrices yielded inconsistent results. These outcomes highlight the strong dependency of GNNs on graph topology and balanced data distribution.

Conversely, the ResNet50-based CNN model demonstrated the most reliable and generalizable performance across both datasets, achieving top classification accuracies of 55.17% on *TIF-DF* and 46.15% on *ChildNet*. The strength of CNNs lies in their ability to capture detailed spatial features and texture patterns, which are critical in modeling expressive facial variations. Nonetheless, CNNs also showed reduced accuracy on underrepresented emotion classes, indicating the potential benefit of techniques like class balancing and advanced feature augmentation.

Siamese networks provided a different approach by learning emotional similarity through image pair comparisons. These models performed well in distinguishing between emotion pairs but struggled with recognizing pairs belonging to the same emotional class. This limitation, stemming from an inability to model intra-class variability effectively, resulted in moderate overall performance. While valuable in verification scenarios, Siamese networks were less suited for multi-class classification in this context.

In conclusion, while GNNs and Siamese models offer niche advantages for structured data and similarity-based tasks, respectively, CNN-based architectures currently provide the most effective and robust solution for direct classification of facial expressions in children. Future work may explore hybrid models or incorporate attention mechanisms and synthetic data generation to further improve recognition accuracy, particularly for low-frequency emotional categories.

5.2 Future Directions

In the coming years, research on facial expression recognition (FER) in children can be significantly enhanced by exploring hybrid deep learning architectures. A promising approach involves combining the capabilities of Convolutional Neural Networks (CNNs) with Graph Neural Networks (GNNs), enabling models to extract both visual texture and geometric relationships from facial landmarks. Incorporating transformer-based attention modules can further refine these models by allowing them to selectively focus on critical facial regions that may hold subtle yet important emotional cues—something particularly relevant when working with children, whose facial expressions tend to be more animated and rapidly shifting. Additionally, integrating temporal models such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks can help in capturing the evolution of emotions across video frames, which is essential for understanding the natural, dynamic nature of children's expressions.

Another important direction involves multimodal emotion recognition, where facial data is complemented by other behavioral or physiological indicators such as vocal expressions, body language, and bio signals (e.g., heart rate or eye movement). Since children often express emotions through multiple channels, combining these inputs can create a more holistic and accurate system. Enhancing the quality of node features in graph-based models—by embedding information like facial region salience or temporal shifts between landmarks—can also lead to better recognition accuracy. Moreover, applying contrastive learning techniques, which help models learn discriminative features by comparing similar and dissimilar pairs, can address challenges associated with limited annotated data. This is especially useful in child-centric FER, where dataset availability is often constrained.

Collectively, these future enhancements could lead to more flexible, generalizable, and interpretable FER systems tailored specifically to children's emotional behavior, facilitating their use in practical domains such as education, healthcare, and interactive technologies.

Bibliography

[1] Ralph Adolphs. Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, 1(1):21–62, 2002.

[2] Anurag. Childnet: A facial expression dataset for children, 2025. Dataset of children's facial expressions containing seven distinct emotions across different age ranges.

[3] Catalina Alina Corneanu, Moisés Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on facial expression recognition in the wild: Databases, methods, and challenges. *Computer Vision and Image Understanding*, 159:1–20, 2017.

[4] Alexandre Dapogny, Kévin Bailly, and Séverine Dubuisson. Benchmarking deep facial expression recognition in children. *IEEE Transactions on Affective Computing*, 2020.

[5] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Hozaifa Kassab, Mohamed Bahaa, and Ali Hamdi. Gcf: Graph convolutional networks for facial expression recognition. *arXiv preprint arXiv:2407.02361*, 2024.

[8] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015.

[9] Abhishek Kumar and S K Singh. Facial expression recognition for children: A review. *Artificial Intelligence Review*, 54(2):1439–1477, 2021.

[10] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[11] Zhi Li, Weihong Deng, Junping Du, et al. Graph-based facial expression recognition with semantic relations. *Pattern Recognition*, 117:107994, 2021.

[12] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability. *Behavior Research Methods*, 47(4):1191–1202, 2015.

[13] Johanna K. Maack, Astrid Bohne, Dag Nordahl, Astrid K. Lindahl, Unni K. Haukvik, Karin V. Kvernenes, Åsa Hammar, Ragnhild S. Høifødt, Lars Wichstrøm, and Hans M. Nordahl. The tromsø infant faces (tif) database: Development, validation, and application to emotion perception. *Frontiers in Psychology*, 8:409, 2017.

[14] Carol Z Malatesta and Jeannette M Haviland. Learning display rules: The socialization of emotion expression in infancy. *Child Development*, pages 991–1003, 1982.

[15] Daniel S Messinger et al. *The development of facial expressions in infancy*, pages 123–147. Springer, 2012.

[16] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[17] R. Pathak and Y. Singh. Real time baby facial expression recognition using deep learning and iot edge computing. In *2020 International Conference on Computing, Communication and Security (ICCCS)*, pages 1–6. IEEE, 2020.

[18] Fadhil Yusuf Rahadika, Novanto Yudistira, and Yuita Arum Sari. Facial expression recognition using residual convnet with image augmentations. *Jurnal Ilmu Komputer dan Informasi*, 14(2):127–135, 2021.

[19] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition using lbp and svm. *Pattern Recognition Letters*, 29(6):931–938, 2009.

[20] Xu Xu, Zhou Ruan, and Lei Yang. Facial expression recognition based on graph neural network. In *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, pages 95–99. IEEE, 2020.