People`s Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Echahid Hamma Lakhdar - El Oued

Faculty of Technology

Department of Electrical Engineering

Dissertation

ACADEMIC MASTER

Domain : Science and Technology

Division : Electrical Engineering

Specialty : Telecommunication Systems

Presented by :

1. MAIZA ELHDJ AMMAR

2. AMIMOUR ANIS

Entitled :

Deep Learning-Based Facial Analysis for Autism Screening

Dissertation Submitted in Partial Fulfillment of the Requirements for the Master Degree in Telecommunication Systems

Publicly defended in : 08/06/2026

Board of Examiners :

Dr.Gacem Abdelmalek	**Chairman**
Dr.Tidjani Amina	**Supervisor**
Dr. Ghendir said	**Examiner**

Academic Year :2026/2025

بسم الله الرحمن الرحيم

Abstract

In recent years, Artificial Intelligence (AI), and more specifically Deep Learning, has become a fundamental tool driving transformative advancements across various fields, particularly in healthcare and developmental screening. This study focuses an intelligent model based on Convolutional Neural Networks (CNN) to analyze facial images and support screening for Autism Spectrum Disorder (ASD) in children.

The significance of this work lies in addressing the challenges of conventional diagnosis, such as delays in specialist availability, subjective interpretation of behavioral cues, and limited accessibility to trained clinicians in many regions. The proposed model follows a structured pipeline involving facial image acquisition, preprocessing, hierarchical feature extraction using deep CNN layers, and binary classification. The model was trained using standard optimization techniques over multiple epochs.

Results demonstrated that the model effectively distinguishes between ASD and non-ASD facial patterns, achieving promising accuracy and balanced precision-recall performance on the held-out test set. Furthermore, explainable AI visualization confirmed that the model focuses on clinically relevant facial regions, particularly the eye area, for ASD detection. A web-based interface was also developed to demonstrate practical applicability.

These findings confirm the feasibility of deep learning-based facial analysis as a supportive tool for preliminary autism screening, while emphasizing that the system is intended to assist rather than replace clinical judgment.

Keywords:

Artificial Intelligence (AI), Autism Spectrum Disorder (ASD), Facial Image Analysis, Deep Learning (DL), Convolutional Neural Network (CNN), Early Screening, Explainable AI, Grad-CAM.

الملخص

في السنوات الأخيرة، أصبح الذكاء الاصطناعي، وعلى وجه التحديد التعلم العميق، أداة أساسية تقود تقدماً تحويلياً في مختلف المجالات، ولا سيما في مجال الرعاية الصحية والكشف عن اضطرابات النمو. تركز هذه الدراسة على تصميم نموذج ذكي يعتمد على الشبكات العصبية التلافيفية(CNN) ، إحدى أبرز تقنيات التعلم العميق، لتحليل الصور الوجهية ودعم الكشف عن اضطراب طيف التوحد (ASD) لدى الأطفال.

تكمن أهمية هذا العمل في معالجة تحديات التشخيص التقليدي، مثل التأخر في توفر الاختصاصيين، والتفسير الذاتي للإشارات السلوكية، والوصول المحدود إلى الأطباء المتدربين في العديد من المناطق. يتبع النموذج المقترح خطوات منظمة تشمل جمع الصور الوجهية، والمعالجة المسبقة، واستخراج السمات الهرمية باستخدام طبقات CNN عميقة، والتصنيف الثنائي. تم تدريب النموذج باستخدام تقنيات تحسين قياسية على مدى عدد من الحقب.

أظهرت النتائج أن النموذج يميز بفعالية بين أنماط الوجه للتوحد وغير التوحد، محققاً دقة جيدة وأداءً متوازناً بين الدقة والاستدعاء على مجموعة الاختبار. علاوة على ذلك، أكدت تقنية التفسير المرئي أن النموذج يركز على مناطق الوجه ذات الصلة سريرياً، وخاصة منطقة العين، للكشف عن التوحد. كما تم تطوير واجهة ويب لتوضيح القابلية العملية للنظام.

تؤكد هذه النتائج جدوى تحليل الصور الوجهية المستند إلى التعلم العميق كأداة داعمة وواعدة للكشف عن التوحد، مع التأكيد على أن النظام يهدف إلى المساعدة وليس إلى استبدال الحكم السريري.

الكلمات المفتاحية:

الذكاء الاصطناعي (AI)، اضطراب طيف التوحد (ASD)، تحليل الصور الوجهية، التعلم العميق (DL)، الشبكات العصبية التلافيفية (CNN)، الكشف المبكر، الذكاء الاصطناعي القابل للتفسير، **Grad-CAM.**

Résumé

Ces dernières années, l'Intelligence Artificielle (IA), et plus particulièrement l'Apprentissage Profond (Deep Learning), est devenue un outil fondamental qui impulse des avancées transformatrices dans divers domaines, notamment dans les soins de santé et le dépistage du développement. Cette étude se concentre sur un modèle intelligent basé sur les Réseaux de Neurones Convolutifs (CNN) pour analyser des images faciales et soutenir le dépistage du Trouble du Spectre Autistique (TSA) chez les enfants.

L'importance de ce travail réside dans la prise en compte des défis du diagnostic conventionnel, tels que les retards dans la disponibilité des spécialistes, l'interprétation subjective des indices comportementaux et l'accessibilité limitée aux cliniciens formés dans de nombreuses régions. Le modèle proposé suit une chaîne de traitement structurée impliquant l'acquisition d'images faciales, le prétraitement, l'extraction hiérarchique de caractéristiques utilisant des couches profondes de CNN, et la classification binaire. Le modèle a été entraîné en utilisant des techniques d'optimisation standard sur plusieurs époques.

Les résultats ont démontré que le modèle distingue efficacement les motifs faciaux TSA et non-TSA, atteignant une exactitude prometteuse et des performances équilibrées en précision et rappel sur l'ensemble de test réservé. De plus, la visualisation par IA explicable (Explainable AI) a confirmé que le modèle se concentre sur des régions faciales cliniquement pertinentes, particulièrement la zone des yeux, pour la détection du TSA. Une interface web a également été développée pour démontrer l'applicabilité pratique.

Ces résultats confirment la faisabilité de l'analyse faciale basée sur l'apprentissage profond en tant qu'outil d'appoint pour le dépistage préliminaire de l'autisme, tout en soulignant que le système est destiné à assister plutôt qu'à remplacer le jugement clinique.

Mots-clés :

Intelligence Artificielle (IA), Trouble du Spectre Autistique (TSA), Analyse d'Images Faciales, Apprentissage Profond (DL), Réseau de Neurones Convolutif (CNN), Dépistage Précoce, IA Explicable (Explainable AI), Grad-CAM.

Acknowledgment

We would like to thank our supervisor, **Dr. Tidjani Amina**, for her help and support during our project. She gave us useful advice and always encouraged us. We really appreciate her time, guidance, and patience. It was a great experience to work under her supervision.

إهداء

(وَكانَ فَضْلُ اللَّهِ عَلَيْكَ عَظِيمًا)

ما سلكت البدايات إلا بتيسيره، وما بلغتُ النهايات إلا بتوفيقه ، وماحققتُ الغايات إلا بفضلهً،فالحمد الله قولا وعملا ،والحمد الله على التمام والإنجاز.

إلى الأيادي التي أزالت عن طريقي أشواك الفشل إلى من ساندني بكل حب عند ضعفي إلى من رسموا إلي المستقبل بخطوط من الثقة والحب إلى عائلتي.

إلى النور الذي أنار دربي والسراج الذي لا ينطفئ نوره إلى العزيز الذي حملتُ اسمه فخراً داعمى الأول في مسيرتي وسندي والدي العزيز .

أهدى فرحة تخرجي إلى تلك الإنسانة العظيمة، إلى من جعل الله الجنة تحت أقدامها وسهلت لي الشدائد بدعائها والدتي العزيزة.

إلى خيرة أيامي وصفوتها،إلى ضلعي الثابت وأمان أيامي إخواني وأخواتي .

إلى الذين يبهجهم نجاحي ،وكانوا عوناً وسنداً في هذا الطريق رفقاء السنين وأصحاب الشدائد والأزمات صديقاتي .

وأخيراً الشُكر موصول لنفسي على الصبر والعزيمة والإصرار ،راجية من الله تعالى أن ينفعني بما علمني وأن يعل

مني ما أجهل ويجعله حجةُ لي لا عليّ.

MAIZA ELHADJ AMMAR

Table of Contents

List of Figures

List of Tables

List Of Abbreviqtinos

ADI-R — Autism Diagnostic Interview–Revised

ADOS-2 — Autism Diagnostic Observation Schedule

AI — Artificial Intelligence

ASD — Autism Spectrum Disorder

AUC-ROC — Area Under the Receiver Operating Characteristic Curve

CNN — Convolutional Neural Network

DL — Deep Learning

FN — False Negative

FP — False Positive

F1-score — F1 Score

GPU — Graphics Processing Unit

Grad-CAM — Gradient-weighted Class Activation Mapping

GRU — Gated Recurrent Unit —ة

ImageNet — Large-scale Image Database

LSTM — Long Short-Term Memory

ML — Machine Learning

MobileNet — Mobile Network

PDD-NOS — Pervasive Developmental Disorder–Not Otherwise Specified

ReLU — Rectified Linear Unit

ResNet-50 — Residual Network

RGB — Red, Green, Blue

SHAP — SHapley Additive exPlanations

ShuffleNet — Shuffle Network

TN — True Negative

TP — True Positive

VGG16 — Visual Geometry Group Network (16 layers)

ViT — Vision Transformer

WHO — World Health Organization

General introduction

Autism Spectrum Disorder (ASD) stands as one of the most prevalent neurodevelopmental conditions worldwide, directly affecting children's capacity for social communication, emotional interaction, and behavioral regulation. Given that timely intervention represents the most influential factor in improving developmental outcomes for these children, reducing the time interval between initial suspicion and specialized clinical assessment acquires paramount importance. However, conventional screening and diagnostic processes continue to encounter substantial challenges; they rely heavily on the availability of trained specialists, demand considerable time, and remain susceptible to geographic delays in numerous regions, thereby limiting opportunities for intervention during critical developmental periods.

Conversely, Artificial Intelligence technologies—and particularly Deep Learning and facial image analysis—have emerged as a promising avenue for supporting preliminary screening. The face, being the primary interface for human communication, carries visual signals that may reflect attentional and emotional patterns associated with developmental conditions. Nevertheless, developing reliable AI systems in this domain confronts technical and ethical challenges, manifested in the limited availability of sufficient and diverse data, the need for clinically meaningful validation, as well as ethical and privacy concerns linked to the use of biometric facial images of children.

Hence, the research problem addressed in this thesis crystallizes in the following question: How can deep learning-based facial image analysis be utilized to support autism screening in children while ensuring reliable performance and respecting ethical considerations such as privacy and fairness?

To address this question, this study establishes a primary objective that consists in constructing, implementing, and evaluating a facial image analysis system based on Convolutional Neural Networks (CNNs) to support preliminary screening for Autism Spectrum Disorder in children. Branching from this objective is a set of specific aims: (1) investigating the feasibility of utilizing deep learning-based facial image analysis as a supportive tool for screening; (2) constructing a CNN architecture capable of extracting facial features with clinical relevance; (3) evaluating system performance using appropriate classification metrics including accuracy, recall, precision, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC); (4) addressing ethical considerations related to privacy, fairness, and demographic bias.

Departing from this framework, the thesis advances a technical contribution that seeks to support—without substituting—specialized clinical judgment. It is organized into three main chapters: Chapter I presents the theoretical framework and conceptual background concerning the disorder and the challenges of conventional screening, alongside the technical foundations of Artificial Intelligence and Deep Learning. Chapter II addresses the methodological details of CNNs, optimization mechanisms,

transfer learning, and Explainable Artificial Intelligence. Chapter III focuses on the practical, experimental application and critical discussion of the developed proposal.

Chapter I

Chapter 1 – Introduction and Background

1.1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects millions of children worldwide, influencing their communication, social interaction, and behavioral patterns. Given its lifelong impact, identification of ASD remains one of the most critical factors in improving long-term developmental outcomes, yet traditional screening methods face significant barriers related to specialist availability, subjective assessment, and geographic accessibility [3], [8]. This thesis explores the intersection of clinical needs and artificial intelligence by leveraging deep learning-based facial image analysis to support ASD screening. By analyzing facial cues that may reflect atypical social attention and emotional expression patterns, the proposed system aims to provide clinicians with an objective, scalable, and non-invasive preliminary screening tool. The workflow involves four main stages: data collection and preprocessing, feature extraction using CNNs, classification, and evaluation using appropriate metrics [1], [2]. It is essential to emphasize that this system is designed as a supportive aid to clinical judgment, not as a standalone diagnostic instrument [4], [5].

1.2 Autism Spectrum Disorder (ASD)

1.2.1 History and Evolution of ASD Diagnosis

The conceptualization of autism has evolved significantly since its initial description. In 1943, Leo Kanner first identified autism as a distinct condition characterized by social withdrawal and repetitive behaviors, terming it "infantile autism" at the Johns Hopkins Hospital [3]. Concurrently, in 1944, Austrian pediatrician Hans Asperger described a milder form with preserved language and cognitive abilities, which later became known as Asperger syndrome. For decades, autism was narrowly defined and often conflated with childhood schizophrenia, leading to misdiagnosis and inappropriate treatments.

The diagnostic landscape shifted dramatically with the publication of the Diagnostic and Statistical Manual of Mental Disorders. The DSM-III (1980) introduced "infantile autism" as a separate category, while the DSM-IV (1994) expanded the concept to include Autistic Disorder, Asperger's Disorder, and Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). The most transformative change came with the DSM-5 (2013), which consolidated these subcategories into a single diagnosis: "Autism Spectrum Disorder" (ASD), recognizing autism as a continuum with varying levels of severity and support needs [3].

This historical evolution—from Kanner's original observations of eleven children to the contemporary dimensional understanding—has directly influenced screening methodologies. Modern approaches have moved from purely behavioral observations toward multimodal assessments incorporating facial expressions, gaze patterns, and physiological responses. The growing recognition that facial cues and social attention deficits are among the most prominent observable signs of ASD has motivated researchers to explore computational methods for automated facial analysis [4], [5].

1.2.2 Definition and Characteristics of ASD

Autism Spectrum Disorder is a neurodevelopmental condition characterized by persistent difficulties in social communication and social interaction, together with restricted and repetitive patterns of behavior, interests, or activities. The term spectrum reflects the wide variability in symptom severity, cognitive ability, and functional outcomes across individuals [3], [10].

One of the key reasons ASD is relevant to facial analysis research is that many autism-related characteristics are linked to nonverbal communication and social attention. These may include reduced eye contact, differences in emotional expression, atypical responses to facial cues, and altered attention to social stimuli [4], [5].

Identification is a particularly important aspect of ASD management. Identifying autism-related signs during childhood can facilitate intervention and developmental support. For this reason, computational tools that can support clinicians in recognizing subtle behavioral or facial indicators may provide meaningful practical value [8], [9], [10].

Figure 1.1 provides a schematic overview of the major dimensions affected by Autism Spectrum Disorder, including communication, behavioral, and emotional domains. This visualization helps contextualize how atypical facial expressions and social attention deficits fit within the broader clinical picture of ASD.

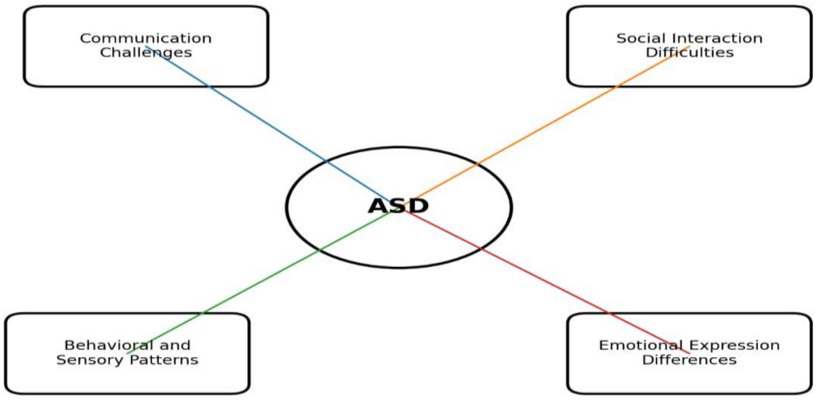

Figure 1.1. Overview of Autism Spectrum Disorder and its major communication, behavioral, and emotional dimensions.

1.2.3 Facial Communication and Social Attention in ASD

The human face serves as the primary interface for social communication. Facial expressions communicate emotional states, regulate social interactions, and establish interpersonal connections. For infants and young children, attention to faces—particularly the eye region—is a crucial developmental milestone [4], [5].Facial expressions play a central role in human communication, conveying attention, emotion, discomfort, interest, and responsiveness. For neurotypical children, learning to recognize and respond to these visual cues is a fundamental part of social development [4], [5].In the case of Autism Spectrum Disorder, many studies suggest that atypical social attention and nonverbal communication may appear at a young age. These differences do not mean that all autistic children express emotion in the same way or that facial appearance alone can determine diagnosis. However, they do suggest that facial behavior, gaze-related patterns, and expression-related cues may contain information relevant to developmental screening [3], [10].Research indicates that children later diagnosed with ASD may show reduced attention to facial regions, atypical gaze patterns, and differences in emotional expressivity from a young age. From a computational perspective, facial analysis provides structured visual input that can be systematically studied. Images can be standardized through preprocessing (face detection, cropping, resizing, normalization) and analyzed using computer vision methods. Compared to other medical modalities, image acquisition is accessible, non-invasive, and scalable [4], [5].Nevertheless, it is important to emphasize that facial analysis should not be interpreted as a substitute for clinical judgment. Its role in this research is supportive and exploratory. The value of such systems lies in helping identify potentially meaningful patterns that may contribute to screening, especially when combined with behavioral and developmental evaluation [4], [5].

1.2.4 Importance of Detection

The importance of ASD research has increased significantly because identification can improve intervention outcomes. Detecting autism during the developmental years may support better language development, emotional regulation, and social adaptation. However, screening remains challenging because it often depends on clinical observation, behavioral assessment, and access to trained specialists [8], [9], [10].

Figure 1.2 illustrates the critical role of timely identification and intervention in improving developmental trajectories for children with ASD. The diagram highlights how delays in screening can narrow the window for effective therapeutic support, reinforcing the need for accessible and scalable screening tools.

Figure 1.2. Importance of autism screening and intervention.

1.3 Traditional ASD Screening Challenges

1.3.1 Clinical and Behavioral Diagnosis

Current ASD identification relies heavily on standardized behavioral assessments,

developmental history evaluation, and specialist observation. Tools such as the ADOS-2 (Autism Diagnostic Observation Schedule, Second Edition) and ADI-R (Autism Diagnostic Interview-Revised) require trained clinicians and significant time investment. The process is inherently subjective and depends on the examiner's experience [8], [9].

1.3.2 Limitations of Traditional Screening Methods

Despite their clinical validity, traditional methods face substantial practical limitations. In many regions, limited access to trained clinicians leads to delayed identification of children at risk, reducing the opportunity for intervention during critical developmental periods [8], [9], [10]. Additionally, behavioral assessments may miss subtle initial signs, particularly in children with milder presentations.

1.4 Artificial Intelligence in Healthcare

1.4.1 Artificial Intelligence (AI)

Artificial Intelligence can be broadly defined as the scientific and engineering discipline concerned with building systems capable of performing tasks that normally require human intelligence, such as reasoning, learning, perception, and decision-making. AI is not a single algorithm or method; rather, it is a broad field that includes multiple subdomains and computational approaches [6].A widely used academic perspective defines AI through the concept of an intelligent agent. In this view, an AI system perceives its environment and acts inways that maximize the probability of achieving specific goals. Historically, classicalAI research focused heavily on symbolic reasoning and rule-based systems. However,these approaches often struggle when dealing with noisy, high-dimensional data suchas images and videos. This limitation contributed to the rise of learning-based methods, especially machine learning and deep learning [1], [2].

1.4.2 Machine Learning (ML)

Machine Learning is a subfield of AI that focuses on building algorithms capable of learning from data. Instead of manually defining every rule, machine learning models discover statistical patterns from examples. One of the most widely cited definitions states that a program learns from experience E with respect to a task T and a performance measure P if its performance at task T improves with experience E, as measured by P. In practical terms, machine learning allows systems to become more effective as they are exposed to more data [1], [2].In image classification tasks, supervised learning is particularly relevant. Here, the model learns from labeled input-output pairs. For example, a facial image may be associated with a specific class label. During training, the model adjusts its parameters to reduce prediction error, enabling it to classify unseen images more accurately later [1], [2], [11].

1.4.3 Deep Learning (DL)

Deep Learning is a subset of machine learning that uses neural networks with multiple layers to learn representations of data automatically. Unlike classical approaches that depend on handcrafted features, deep learning models can extract useful patterns directly from raw input data. This ability makes them especially effective for image analysis [1], [2].

Deep learning is often described as representation learning with multiple levels of abstraction. In the context of images, initial layers may detect simple features such as edges and lines, intermediate layers may recognize shapes and facial parts, and deeper layers may capture high-level concepts and patterns [1], [2].

Convolutional Neural Networks (CNNs) are among the most widely used deep learning architectures for image-based tasks. CNNs process images through layers such as convolution, activation, pooling, and fully connected layers. Their strength lies in learning robust spatial features while reducing sensitivity to irrelevant variations such as position or lighting. This makes CNNs particularly suitable for facial image analysis and classification tasks [1], [2].

Figure 1.3 depicts the hierarchical relationship between Artificial Intelligence, Machine Learning, and Deep Learning. Deep Learning is shown as a subset of Machine Learning, which in turn is a subset of the broader AI field. This taxonomy clarifies the positioning of CNNs within the deep learning branch, which is the core methodology adopted in this study.

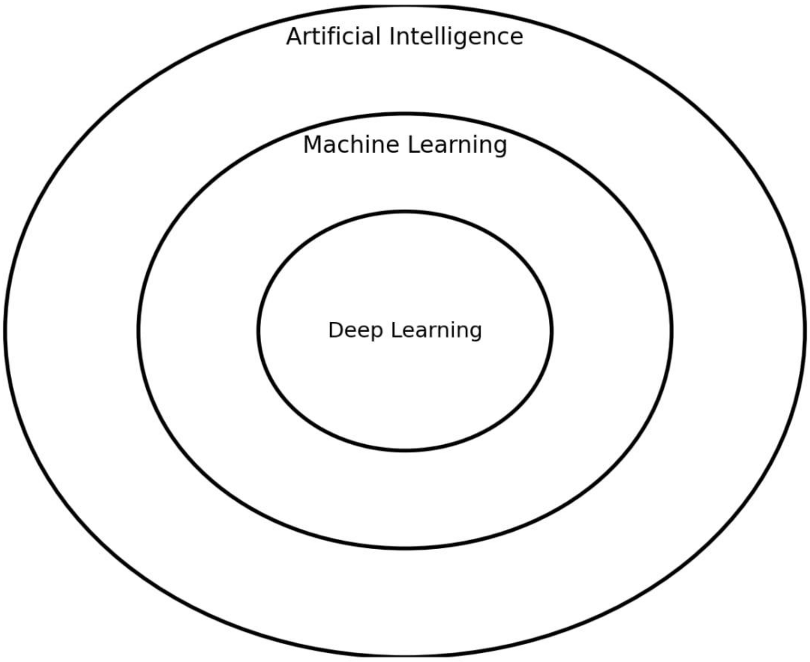

Figure 1.3. Hierarchical relationship between Artificial Intelligence, Machine Learning, and Deep Learning.

1.4.4 AI Applications in Medical and Facial Image Analysis

AI has demonstrated transformative potential in medical image analysis, including radiology, pathology, dermatology, and ophthalmology. Deep learning models can detect anomalies, classify lesions, and assist in diagnosis with accuracy comparable to or exceeding human experts in specific tasks [11], [12].

In developmental and behavioral screening, AI offers the advantage of objective, reproducible analysis of visual cues. Facial image analysis, in particular, provides a structured, observable input that can be standardized and processed computationally, making it attractive for screening support systems [4], [5].

CNNs are particularly suitable for facial image analysis because they automatically learn hierarchical visual features. Initial layers may detect simple features such as edges and lines, intermediate layers may recognize shapes and facial parts, and deeper layers may capture high-level concepts and patterns [1], [2].

Figure 1.4 illustrates the hierarchy of visual abstraction learned by deep neural networks. Starting from low-level features such as edges and textures in initial layers, the network progressively builds up to mid-level facial components (eyes, nose, mouth) and finally to high-level semantic concepts such as expressions and identity. This hierarchical learning is what enables CNNs to perform complex facial analysis tasks without manual feature engineering.

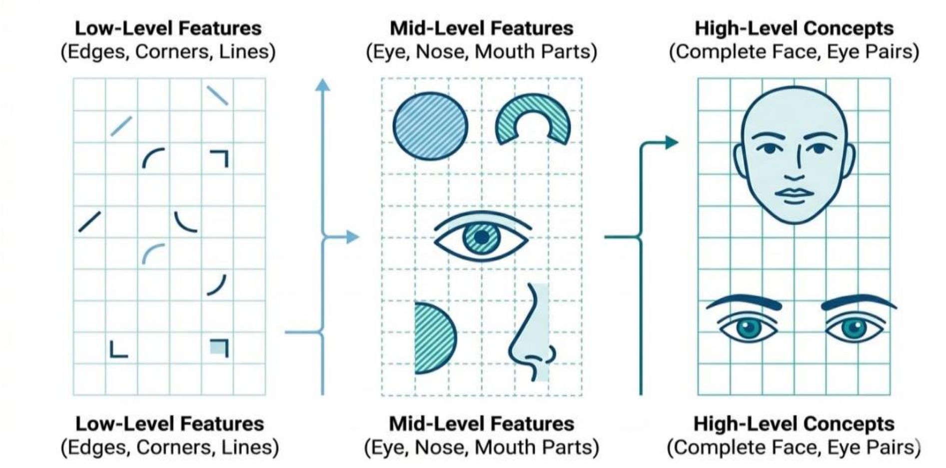

Figure 1.4. Deep learning hierarchy of abstraction from simple visual patterns to higher-level facial concepts.

In medical and behavioral image analysis, CNNs are especially valuable because they reduce dependence on handcrafted assumptions and allow the model to discover complex image-based patterns directly from data. This is one of the main reasons they are selected in this study as the core analytical framework [1], [2], [11].

Transfer learning—where pretrained models are fine-tuned on specific medical datasets—has also proven valuable when target datasets are limited, as is often the case in ASD research [7], [11]. By leveraging features learned from large-scale datasets such as ImageNet, transfer learning can significantly improve performance even with relatively small ASD facial image datasets.

Figure 1.5 presents the proposed workflow of the deep learning-based facial analysis system developed in this study. The pipeline begins with facial image acquisition and preprocessing, followed by hierarchical feature extraction using CNN layers, binary classification, and finally performance evaluation using standard metrics. This structured approach ensures reproducibility and clinical interpretability of the screening results.

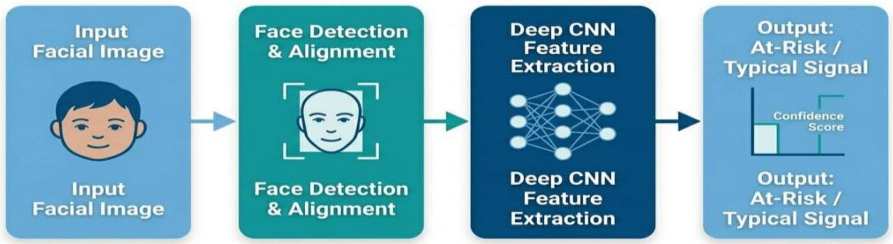

Figure 1.5. Proposed workflow of the deep learning-based facial analysis system for autism screening.

1.5 Chapter Conclusion

This chapter presented the conceptual background required for deep learning-based facial analysis in autism screening. It traced the historical evolution of ASD diagnosis from Kanner's original observations to the contemporary DSM-5 spectrum understanding, and introduced Autism Spectrum Disorder from a clinical and developmental perspective, emphasizing the importance of screening and the practical motivation for supportive AI-based tools.

The chapter then presented Artificial Intelligence, Machine Learning, and Deep Learning as the technological foundation of the study, and explained why CNN-based image analysis is particularly relevant for facial analysis tasks. The convergence of AI and ASD research was articulated through the lens of computational facial analysis, demonstrating how deep learning can objectively analyze visual behavioral cues. The research problem and objectives were defined, positioning this study as an interdisciplinary effort to develop a responsible, supportive screening tool. Finally, ethical and fairness considerations were discussed to ensure that the technological approach remains grounded in responsible and ethical clinical application.

Chapter II

Deep Learning And Convolutional Neural Network

2.1 Introduction

Deep learning has revolutionized artificial intelligence by enabling computer systems to process complex data with unprecedented efficiency. Within this domain, Convolutional Neural Networks (CNNs) have emerged as the predominant architecture for computer vision tasks, given their ability to automatically extract hierarchical spatial features without manual feature engineering [1], [2]. This chapter presents the theoretical foundations of deep learning and CNNs, including their architectural components, the feature extraction pipeline, binary classification mechanisms, backpropagation, and transfer learning. These concepts form the methodological basis for the facial image analysis system proposed in this thesis.

2.2 Deep Learning Fundamentals

2.2.1 Definition and Overview

Deep Learning is a subset of machine learning that employs neural networks with multiple layers to learn hierarchical representations of data automatically. Unlike classical approaches that rely on handcrafted features, deep learning models extract useful patterns directly from raw input data, making them especially effective for image analysis [5], [2].

Historically, image processing relied on feature engineering, where experts manually designed features. The emergence of CNNs shifted this paradigm toward feature learning, where representations are learned directly from data, leading to significant performance improvements and reduced reliance on human expertise [3], [4].

2.2.2 Hierarchy of Representation Learning

Deep learning models learn representations at multiple levels of abstraction. In the context of facial images, early layers detect simple features such as edges and textures, intermediate layers identify facial components (eyes, nose, mouth), and deeper layers capture high-level concepts such as facial expressions and identity [1], [9].

2.3 Convolutional Neural Networks (CNNs)

2.3.1 Definition and Characteristics

Convolutional Neural Networks (CNNs) are a specialized type of artificial neural network designed to process data with a grid structure, such as images, where data is represented as two-dimensional or multi-dimensional matrices [6]. LeCun defines CNNs as models that rely on convolution to extract spatial features, originally developed for pattern recognition in images [2]. Goodfellow indicates that CNNs are characterized by their ability to reduce the number of coefficients through weight sharing and local connectivity, making them more efficient than traditional fully connected networks [9].

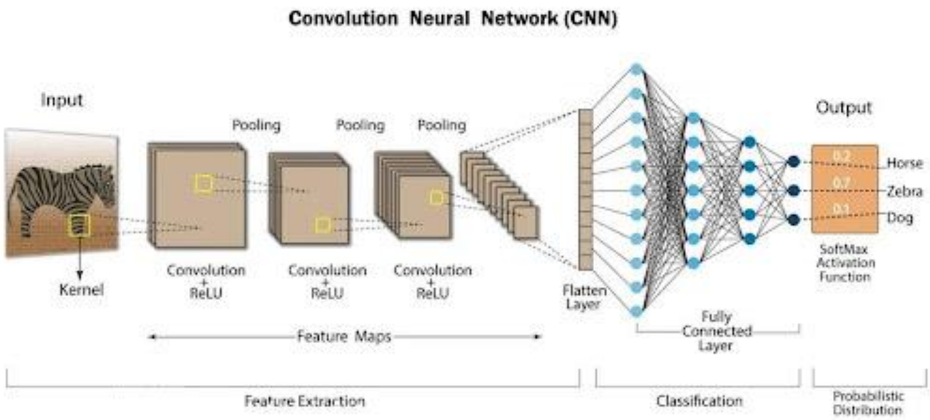

Figure 2.1. The basic structure of Convolutional Neural Networks (CNNs), consisting of Convolution, Pooling, and Fully Connected layers [2].

2.3.2 CNN Architecture Components

- **Convolution Layer**

The convolution layer is the fundamental building block of a CNN. Small filters or kernels slide across the input image to extract local patterns. Each filter responds to a specific visual feature, such as edges, corners, or texture transitions. During training, these filters are not manually designed; instead, they are learned from the data through optimization [1], [3], [5].

The convolution operation for a single position is defined as:

$$\text{Output} = \sum(\text{Image} \times \text{Filter}) \quad (2.1)$$

Consider a 6×6 input image convolved with a 3×3 vertical edge detection filter. The input matrix I and filter K are:

$$I = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}, \quad K = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \quad (2.2)$$

The first element of the resulting feature map is computed by element-wise multiplication and summation:

$$(0 \times 1) + (1 \times 0) + \big(1 \times (-1)\big) + (0 \times 1) + (0 \times 0) + \big(0 \times (-1)\big) + (0 \times 1) + (0 \times 0) + \big(1 \times (-1)\big) = -2 \quad (2.3)$$

The complete feature map F after sliding the filter across all valid positions is:

$$F = \begin{bmatrix} -2 & -1 & 1 & 2 \\ -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 1 \\ -1 & 0 & -1 & 1 \end{bmatrix} \quad (2.4)$$

- **Activation Function (ReLU)**

Following convolution, activation functions introduce nonlinearity into the network. The Rectified Linear Unit (ReLU) is the most commonly used activation function in CNNs, defined mathematically as:

$$a = f(z') = \max(0, z') \quad (2.5)$$

ReLU prevents the accumulation of negative values that could negatively impact learning, helps accelerate training compared to traditional functions like Sigmoid or Tanh, and reduces the vanishing gradient problem during backpropagation [1].

Applying ReLU to the feature map F yields:

$$\text{ReLU}(F) = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(2.6)}$$

- **Bias Addition**

A bias coefficient is added to the output of the convolution operation. Without bias, all output depends solely on the values detected by the filter. Bias provides the network with flexibility to learn even when all values from the filter are zero, thereby reducing errors and improving generalization [3].

Adding a bias b = +1 to the ReLU-activated feature map:

$$F_b = \begin{bmatrix} -1 & 0 & 2 & 3 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \end{bmatrix} \quad \text{(2.7)}$$

After applying ReLU again (or assuming bias is added before activation):

$$F_{final} = \begin{bmatrix} 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \end{bmatrix} \quad \text{(2.8)}$$

- **Pooling Layer**

Pooling layers reduce the spatial dimensions of feature maps while retaining the most informative features. This operation decreases computational complexity, improves generalization, and provides invariance to minor image transformations such as translation or rotation. Max-pooling is the most widely adopted pooling strategy [1], [2].

Applying 2×2 max-pooling to the 4×4 feature map:

$$\text{MaxPool}\left(\begin{bmatrix} 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \end{bmatrix}\right) = \begin{bmatrix} 0 & 2 \\ 2 & 2 \end{bmatrix} \quad \text{(2.9)}$$

Each 2×2 block is replaced by its maximum value, reducing the dimensionality while preserving dominant activation patterns.

- **Flattening Layer**

Flattening is the process of converting a multidimensional feature map into a one-dimensional vector to enable its passage to fully connected layers. Since fully connected layers only accept vector inputs, this transformation is essential for bridging the feature extraction and classification stages [1].

The 2×2 pooled feature map is flattened as follows:

$$\text{Flatten}\left(\begin{bmatrix} 3 & 0 \\ 2 & 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 3 \\ 1 \\ 2 \end{bmatrix} \quad \text{(2.10)}$$

2.3.3 Feature Extraction Summary

The complete feature extraction pipeline in a CNN consists of:

(1) Convolution with learnable filters;

(2) Bias addition;

(3) Nonlinear activation (ReLU);

(4) Spatial downsampling (Pooling);

(5) Dimensionality reduction (Flattening). These operations collectively transform raw pixel values into high-level feature vectors suitable for classification [1], [2].

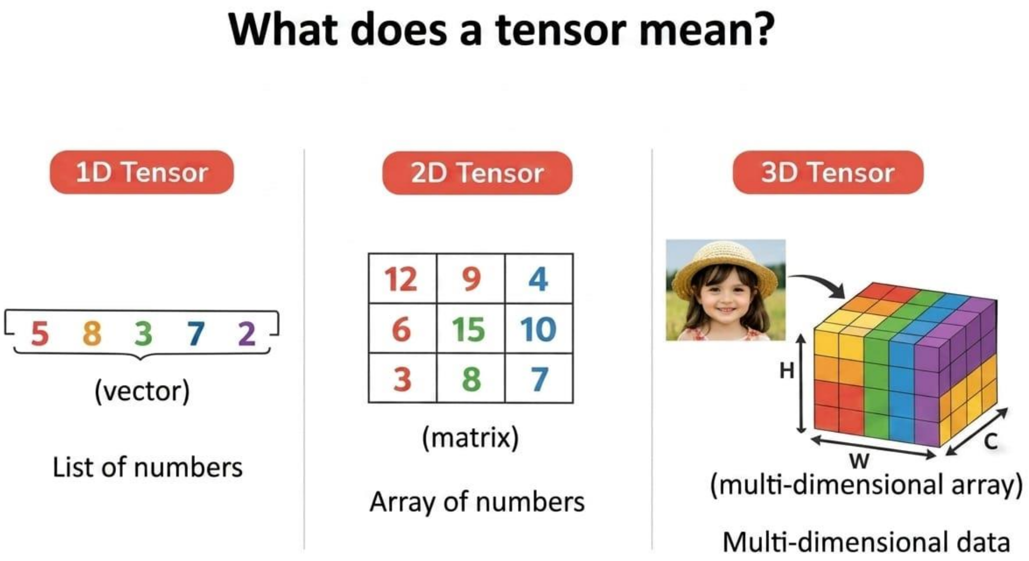

Figure 2.2. Color image representation as a three-dimensional RGB matrix

2.4 Binary Classification in CNNs

2.4.1 Fully Connected Layer Computation

In binary classification, the fully connected layer integrates extracted features to produce a scalar output. Each neuron performs a weighted summation of its inputs plus a bias term:

$$z = \sum(x_i \cdot w_i) + b \quad (2.11)$$

where x_i represents the flattened input values, w_i are the learned weights, and b is the bias term [1], [5].

Numerical Example:

Given a flattened input vector x = [0, 3, 1, 2], weights w = [0.5, 1, -1, 2], and bias b = 1, the computation proceeds as:

$$z = (0 \times 0.5) + (3 \times 1) + \big(1 \times (-1)\big) + (2 \times 2) + 1 = 0 + 3 - 1 + 4 + 1 = 7 \quad (2.12)$$

Applying ReLU activation:

$$a = \mathrm{ReLU}(7) = 7 \quad (2.13)$$

2.4.2 Sigmoid Activation and Decision Rule

The sigmoid function transforms the linear output into a probability value bounded between 0 and 1:

$$\hat{y} = \frac{1}{1+e^{-z}} \quad (2.14)$$

For z = 7, the predicted probability is approximately ŷ ≈ 0.999. The binary decision rule is defined as:

$$\hat{y} = \begin{cases} Class\ 1\ (Positive, ASD) & \hat{y} \geq 0.5 \\ Class\ 0\ (Negative, Non-ASD) & \hat{y} < 0.5 \end{cases} \quad (2.15)$$

This probabilistic output enables interpretable confidence scores for clinical screening applications [1].

2.4.3 Loss Function (Binary Cross-Entropy)

The binary cross-entropy loss function quantifies the discrepancy between the predicted probability and the true label:

$$L = -[\, y log(\hat{y}) + (1-y) log(1-\hat{y})\,] \quad (2.16)$$

where y ∈ {0, 1} is the ground-truth label and ŷ is the predicted probability. Minimizing this loss during training drives the model to produce accurate probability estimates for the target classes [1], [5].

Example: For a true label y = 1 and prediction ŷ = 0.7:

$$L = -\big(y log(\hat{y}) + (1-y) log(1-\hat{y})\big) = -log(0.7) \quad (2.17)$$

2.5 Backpropagation and Optimization

2.5.1 Forward Pass and Loss Computation

Backpropagation is an algorithm used to compute the gradients of the loss function with respect to all network parameters. The process begins with a forward pass, where the input is propagated through the network to produce a predicted output. The difference between the prediction and the true label is measured using the loss function [1].

2.5.2 Gradient Computation and Weight Update

The gradient of the loss with respect to the pre-activation output is computed as:

$$\frac{\partial L}{\partial z} = \hat{y} - y \quad (2.18)$$

This error signal is propagated backward through the network using the chain rule to compute gradients for each layer. For the weights in the fully connected layer, the gradient is:

5. Parameter Update The weights are updated using gradient descent:

$$W = W - \eta \left(\frac{\partial L}{\partial W}\right) \quad (2.19)$$

5. Backpropagation (Gradient)

$$\frac{\partial L}{\partial z} = \hat{y} - y = 0.7 - 1 = -0.3 \quad (2.20)$$

---

6. Weight Update

Assuming:

$$X = [\,1,\ 2,\ 3\,]$$

$$\frac{\partial L}{\partial W} = (\hat{y} - y) \cdot X^T = -0.3 \cdot [\,1,\ 2,\ 3\,] = [\,-0.3,\ -0.6,\ -0.9\,] \quad (2.21)$$

Update rule:

$$W = W - \eta \cdot \frac{\partial L}{\partial W} \quad (2.22)$$

The negative gradient increases the weights, pushing the prediction closer to the true label in subsequent iterations [1], [5].

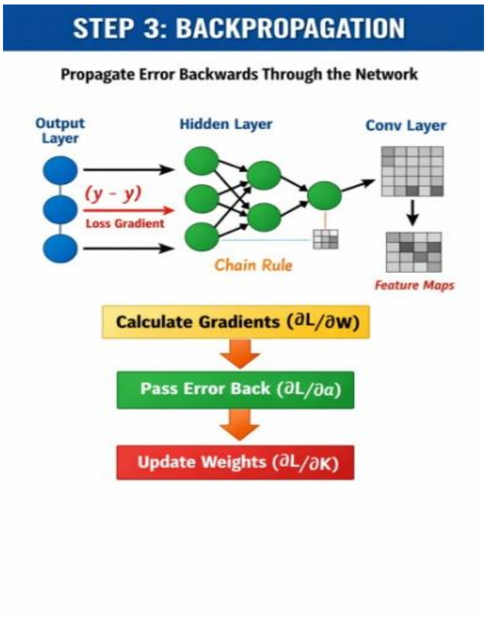

Figure 2.3. Illustration of backpropagation and gradient descent optimization in neural networks.

2.6 Transfer Learning

2.6.1 Concept and Importance

Transfer Learning is a machine learning technique where a model developed for a general task is reused as the starting point for a specific task. Instead of training a deep neural network from scratch, the researcher leverages the feature-extraction capabilities of a pre-trained model (e.g., ResNet, VGG16) trained on large-scale datasets such as ImageNet [6], [7].

This approach offers three principal advantages:

(1) reduced training time,

(2) improved performance when the target dataset is relatively small

(3) access to generic visual features learned from millions of images. In the context of ASD facial analysis, where datasets are often limited, transfer learning is particularly relevant [4], [6].

2.6.2 Transfer Learning Process

The transfer learning process involves five stages:

(1) Load a pre-trained base model and remove its original classification head;

(2) Freeze the base layers to preserve learned features;

(3) Add a new fully connected classification head matching the target task (e.g., 2 neurons for binary classification);

(4) Train only the new head using backpropagation;

(5) Optionally fine-tune top layers with a small learning rate to adapt high-level features to the new dataset [6], [7].

2.7 Definition of Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) refers to the techniques and methods designed to make the behavior and predictions of AI systems understandable to humans. [13]. The primary objective of XAI is to mitigate the "black-box" problem, which occurs when deep learning systems generate accurate outputs without revealing the underlying reasoning behind their decisions [13].

Unlike traditional opaque models, XAI aims to provide transparent reasoning, allowing users to understand, trust, and appropriately rely on AI-driven outcomes [13].

Given that a unique definition is often challenging because explainability is context-dependent, an interdisciplinary Delphi study involving expert consensus proposed a formal definition: An explanation is a tool intended to assist a user with insights relevant to the function of an AI/ML model, designed for a specific purpose, and the reason for this particular output [14].

This definition emphasizes that an explanation is a functional means (tool) intended to support an operator (user) with tailored information (insights) regarding the model's logic and the specific cause for a given result (particular output).

2.7.1 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a prominent post-hoc, model-specific explainability technique designed to make the decision-making processes of Convolutional Neural Networks (CNNs) transparent [13].

Introduced by Selvaraju et al., Grad-CAM addresses the "black box" problem by providing visual justifications for model predictions, identifying which regions of an input image were most influential in a specific classification decision.

The fundamental logic of Grad-CAM relies on the gradient information flowing into the final convolutional layer of the CNN. This layer is chosen because it represents the optimal compromise between high-level semantics and detailed spatial information [15]. To generate a localization map $L^{c}_{\mathrm{Grad-CAM}}$ for any target class c, the algorithm follows a two-step mathematical process:

1. Computing Neuron Importance Weights: First, the gradient of the score for class c (yc, computed before the softmax) is calculated with respect to the feature maps Ak of the last convolutional layer. These gradients are then global-average-pooled over the width i and height j to obtain the neuron importance weights αkc:

$$\alpha_k^c = \frac{1}{Z}\sum\sum\frac{\partial y^c}{\partial A_{ij}^k}$$

This weight captures the "importance" of feature map k for a target class c, effectively acting as a partial linearization of the deep network downstream from the convolutional layer.

2. Generating the Heatmap: The final localization map is obtained by performing a weighted combination of forward activation maps and passing the result through a Rectified Linear Unit (ReLU) activation:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c \; A^k)$$

The ReLU activation is critical here; it ensures that the model only highlights features that have a positive influence on the class of interest (i.e., pixels whose intensity should be increased to increase the class score) [16].

For instance, in medical imaging, it can highlight a cancerous lesion on an X-ray or MRI, allowing clinicians to validate the model's findings against their own expertise [16].

3. Upsampling to Original Image Dimensions: The resulting localization map $L^{Grad-cAMc}$ has spatial dimensions' u × v (the same as the last convolutional feature maps). To visualize it over the original image, it is upsampled (resized) to the original image dimensions using bilinear interpolation:

Heatmap = Upsample($L^{Grad-cAMc}$) → dimensions of I

4. Overlay and Visualization: The heatmap is normalized to the range [0, 1] and mapped to a color palette (e.g., Jet or Hot), where warm colors (red, yellow) indicate high importance and cool colors (blue, black) indicate low importance. The final visualization is obtained by superimposing the colored heatmap onto the original image with a transparency factor (typically α = 0.5):

Visualization = I + α × Colormap(Heatmap)

2.7 Chapter Conclusion

This chapter established the theoretical foundations of deep learning and Convolutional Neural Networks essential for the proposed ASD screening system. The hierarchical feature extraction capabilities of CNNs, the mathematical formulation of convolution and pooling operations, the binary classification framework using fully connected layers and sigmoid activation, the backpropagation optimization algorithm, and the principles of transfer learning were presented. These theoretical components provide the necessary methodological basis for the facial image analysis system described in Chapter 3.

Chapter III

Chapter 3 – Results, Discussion, and Conclusion

3.1 Introduction

This chapter presents the experimental investigation of deep learning techniques for Autism Spectrum Disorder (ASD) classification using facial images. A complete facial image analysis framework was implemented, including dataset acquisition, preprocessing, data augmentation, CNN-based feature extraction, classification, performance evaluation, and explainable artificial intelligence analysis. The experiments were conducted using Python within the Google Colab environment with GPU acceleration. Furthermore, a web-based interface was built to demonstrate practical applicability in supportive autism screening environments.

3.2 Proposed System

The proposed system is built to analyze facial images for ASD classification using deep learning techniques. The framework follows a structured pipeline beginning with facial image acquisition from publicly available Kaggle datasets, followed by preprocessing steps including resizing, grayscale conversion, and normalization. The preprocessed dataset is fed into a Convolutional Neural Network (CNN) that automatically learns hierarchical facial features. The extracted features are passed through fully connected layers and a sigmoid output layer to produce predictive probabilities for the ASD and non-ASD classes. To improve interpretability, Grad-CAM visualization was integrated to highlight facial regions that contributed most to the model's predictions. Finally, a web-based testing interface was implemented for practical evaluation.

Figure 3.1 presents the overall architecture of the proposed ASD facial image classification system, illustrating the sequential flow from input image acquisition through preprocessing, CNN feature extraction, classification, and final output prediction.

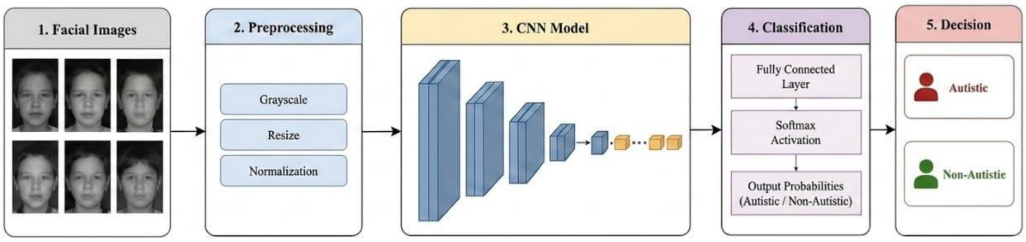

Figure 3.1. Overview of the proposed ASD facial image classification system.

3.3 Dataset and Preprocessing

3.3.1 Dataset Description

The dataset used in this study was obtained from Kaggle and consists of facial images collected for ASD classification tasks. The dataset contains a total of 1982 facial images partitioned into training, testing, and validation subsets comprising 1192, 300, and 300 images, respectively. Each subset includes two categories: Autistic and Non-Autistic.

Figure 3.2 displays representative samples from the investigated dataset, illustrating the variability in facial expressions, head orientation, illumination conditions, and age distribution across both ASD and non-ASD classes.

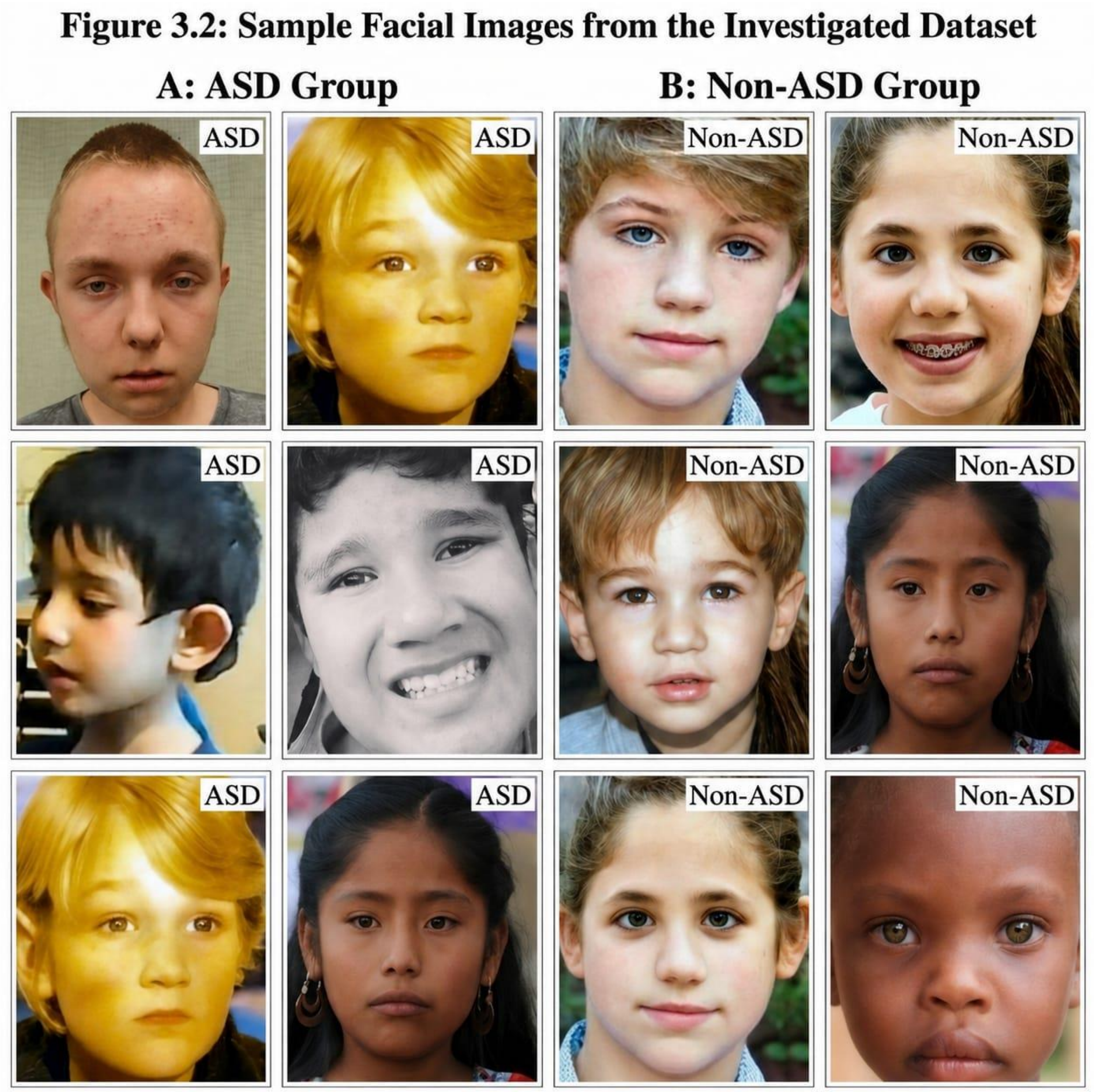

Figure 3.2. Representative samples from the investigated dataset showing variability in facial expressions, head orientation, illumination, and age distribution.

3.3.2 Preprocessing Pipeline

Prior to training, all facial images underwent preprocessing to ensure input consistency and improve CNN training stability. Each image was resized to 128×128 pixels and normalized by scaling pixel intensity values to the range [0, 1].

3.3.3 Data Augmentation

Data augmentation techniques were applied to the training set to increase sample diversity and reduce overfitting. The augmentation pipeline included horizontal flipping, zoom adjustment, rotation, and brightness variation. Specifically, a zoom range of 0.2 was employed, allowing the model to learn scale-invariant features by randomly zooming images up to 20% inward or outward. This augmentation simulates variations in camera distance and facial scale that may occur in real-world screening scenarios, thereby improving the model's robustness

to changes in subject distance and framing .

3.4 CNN Architecture and Training Configuration

3.4.1 Network Architecture

The CNN architecture employed in this study consists of four convolutional layers with increasing filter depths, followed by batch normalization, ReLU activation, and max-pooling operations. Table 3.1 presents the detailed layer configuration.

Table 3.1: CNN Architecture Configuration

Layer Type	**Configuration**
Input Layer	128×128×3
Convolutional Layer 1	32 filters, 3×3
ReLU Activation	ReLU
Max Pooling	2×2
Convolutional Layer 2	64 filters, 3×3
Convolutional Layer 3	128 filters, 3×3
Convolutional Layer 4	256 filters, 3×3
Flatten Layer	Dimensionality reduction
Fully Connected Layer	128 neurons
Dropout Layer	Rate: 0.3
Output Layer	1 neuron + Sigmoid

3.4.2 Training Configuration

The training procedure employed the Adam optimization algorithm together with the binary cross-entropy loss function. The Rectified Linear Unit (ReLU) activation function was utilized within convolutional layers, while the sigmoid activation function was applied in the final output layer to generate classification probabilities. A dropout layer with rate 0.3 was incorporated to reduce overfitting and improve generalization.

3.5 Hyperparameter Tuning and Model Selection

To determine the optimal CNN configuration, several experiments were conducted with varying convolutional layer configurations, batch sizes, and epoch numbers. Table 3.2 presents the comprehensive hyperparameter tuning results.

Table 3.2: Hyperparameter Tuning Results

Convolution Layers	**Filter Size**	**Epochs**	**Batch Size**	**Test Accuracy**
(32, 64, 128)	(3,3)	20	16	65.20%
(32, 64, 128)	(3,3)	20	32	59.28%
(32, 64, 128)	(3,3)	30	35	70.23%
(32, 64, 128, 256, 512)	(3,3)	50	32	49.67%
(32, 64, 128, 256)	(3,3)	20	32	78.00%

(32, 64, 128, 256)	**(3,3)**	**20**	**16**	**82.00%**

The optimal configuration was identified as four convolutional layers with filters (32, 64, 128, 256), trained for 20 epochs with a batch size of 16. This architecture achieved the highest test accuracy of 81.00% and training accuracy of 84.25%, indicating stable learning without significant overfitting. The deeper architecture with five layers (32, 64, 128, 256, 512) suffered from severe underperformance (49.67% test accuracy), suggesting excessive model complexity for the available dataset size.

3.6 Training and Validation Performance

3.6.1 Accuracy Curves

Figure 3.3 illustrates the training and validation accuracy curves over 20 epochs using the optimal hyperparameters. The training accuracy started at approximately 59% and steadily increased to reach 81% by epoch 20. The validation accuracy demonstrated more volatile behavior initially, dropping to 66% at epoch 1 before recovering and stabilizing around 80-84% from epoch 5 onwards. Notably, the validation accuracy consistently exceeded training accuracy during epochs 1-15, which can be attributed to the application of dropout regularization during training. After epoch 15, both curves converged closely, with the final gap of 3.25% (81% vs. 84.25%), indicating good generalization without severe overfitting.

Figure 3.3 shows the progression of training and validation accuracy across 20 epochs, demonstrating the model's learning trajectory and convergence behavior.

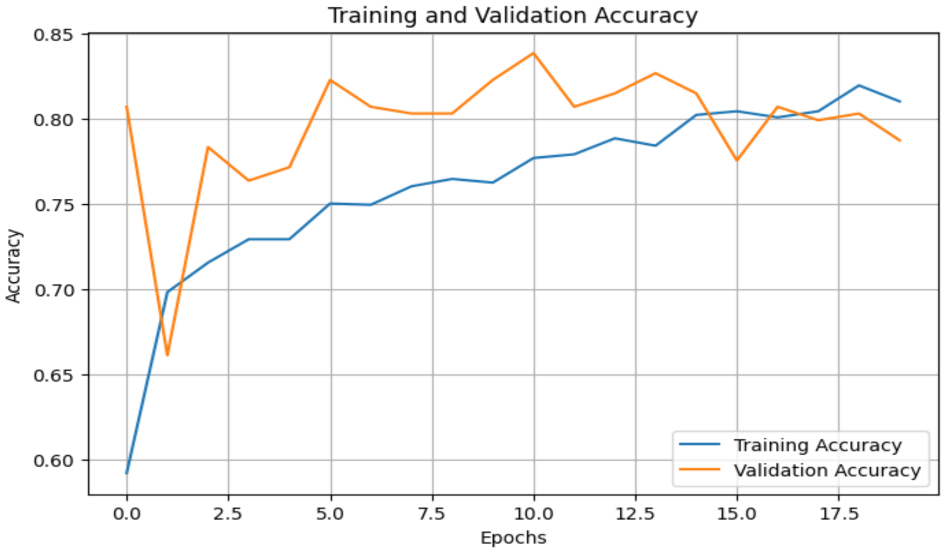

Figure 3.3. Training and validation accuracy curves over 20 epochs.

3.6.2 Loss Curves

Figure 3.4 presents the corresponding loss curves for both training and validation sets. The training loss exhibited a smooth, monotonic decrease from 0.65 to 0.40 over

20 epochs, demonstrating effective gradient descent optimization. The validation loss showed higher volatility initially, peaking at 0.54 at epoch 1, but subsequently stabilized in the range of 0.35-0.45 from epoch 5 onwards. The consistent gap between training and validation loss (training loss remaining higher) further confirms the regularization effect of dropout layers. Figure 3.4 displays the training and validation loss curves, reflecting the optimization process and the impact of dropout regularization on model generalization.

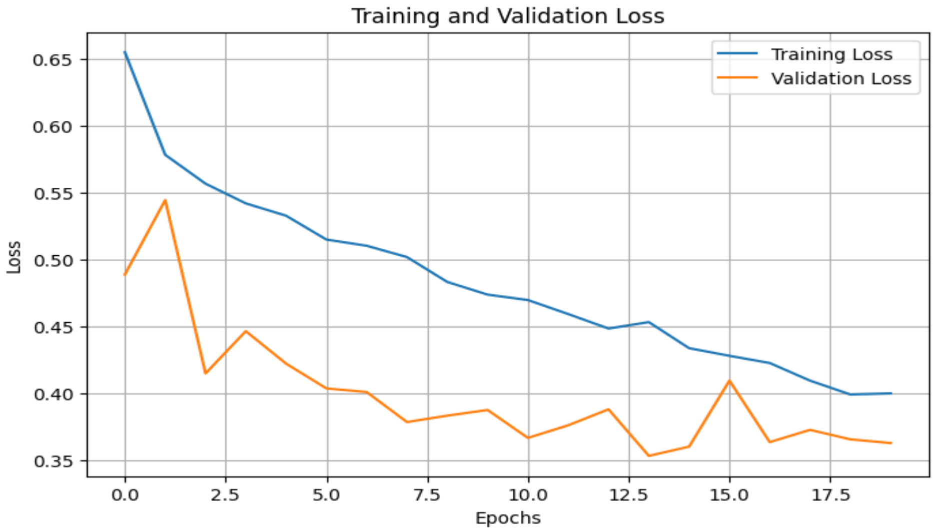

Figure 3.4. Training and validation loss curves over 20 epochs.

3.7 Evaluation Metrics and Confusion Matrix

3.7.1 Confusion Matrix Analysis

The final model was evaluated on the held-out test set comprising 300 facial images (150 ASD, 150 non-ASD). Figure 3.5 presents the confusion matrix for binary classification.

Figure 3.5 visualizes the confusion matrix for the binary classification task, showing the distribution of true positives, true negatives, false positives, and false negatives across the test set.

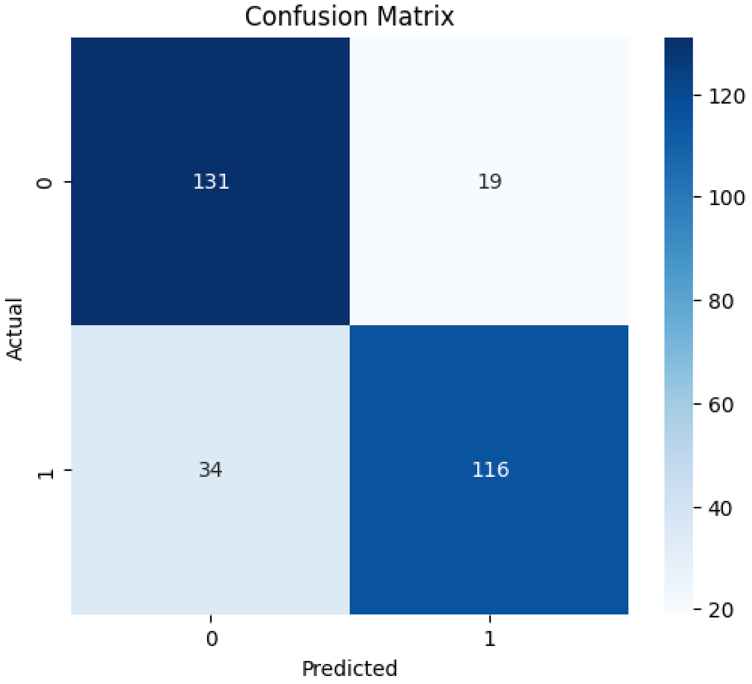

Figure 3.5. Confusion matrix on the held-out test set.

Table 3.3: Confusion Matrix Values

	Predicted Non-ASD (0)	Predicted ASD (1)
Actual Non-ASD (0)	True Negative: 131	False Positive: 19
Actual ASD (1)	False Negative: 34	True Positive: 116

The model correctly classified 131 out of 150 non-ASD cases (87.33% specificity) and 116 out of 150 ASD cases (77.33% sensitivity). The misclassification analysis reveals 19 false positives and 34 false negatives. The higher number of false negatives is clinically significant, as missing an ASD case is generally more consequential than a false alarm in screening contexts.

3.7.2 Comprehensive Evaluation Metrics

Based on the confusion matrix values, the complete set of performance metrics was calculated as presented in Table 3.4.

Table 3.4: Performance Metrics Summary

Metric	Formula	Value
Accuracy	(TP + TN) / Total	82.00%
Precision	TP / (TP + FP)	85.93%
Recall (Sensitivity)	TP / (TP + FN)	77.33%

Specificity	TN / (TN + FP)	87.33%
F1-Score	2×(P×R)/(P+R)	81.40%
False Positive Rate	FP / (FP + TN)	12.67%
False Negative Rate	FN / (FN + TP)	22.67%

TP= True Positive

TN= True Negativerepresents

FP= False Positive

FN= False Negative

The accuracy of 81.00% indicates that the model correctly classifies approximately 4 out of 5 facial images. The precision of 85.93% demonstrates that when the model predicts ASD, it is correct nearly 86% of the time. The recall of 77.33% shows that the model successfully identifies approximately three-quarters of all true ASD cases. The F1-score of 81.40% represents a balanced harmonic mean between precision and recall.

3.8 Explainable AI: Grad-CAM Visualization

3.8.1 Grad-CAM Methodology

Despite achieving reasonable classification accuracy, deep learning models are often criticized as "black boxes" because their internal decision-making process is not transparent. In healthcare applications such as autism detection, interpretability is vital. To address this limitation, Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to highlight the regions the model focused on in the facial image. Grad-CAM uses gradients of the target class flowing into the last convolutional layer to produce a coarse localization map. Brighter colors (red and yellow) indicate high-importance regions, while dark colors (blue) indicate low-importance regions.

3.8.2 Visualization Results

Results revealed important patterns: (1) For images classified as ASD, the model repeatedly focuses on the eye region and upper part of the face, aligning with clinical observations that children with autism often show reduced eye contact and atypical gaze patterns; (2) For images classified as non-ASD, attention is more evenly distributed across the entire face including the mouth and cheeks, indicating typical social attention patterns; (3) For some misclassified images, inconsistent heatmaps appear where the model focuses on backgrounds or non-facial regions, indicating weaknesses in feature learning.

Figure 3.6 presents Grad-CAM visualizations for representative facial image samples, showing attention heatmaps that highlight the facial regions most influential in the model's classification decisions for both ASD and non-ASD categories.

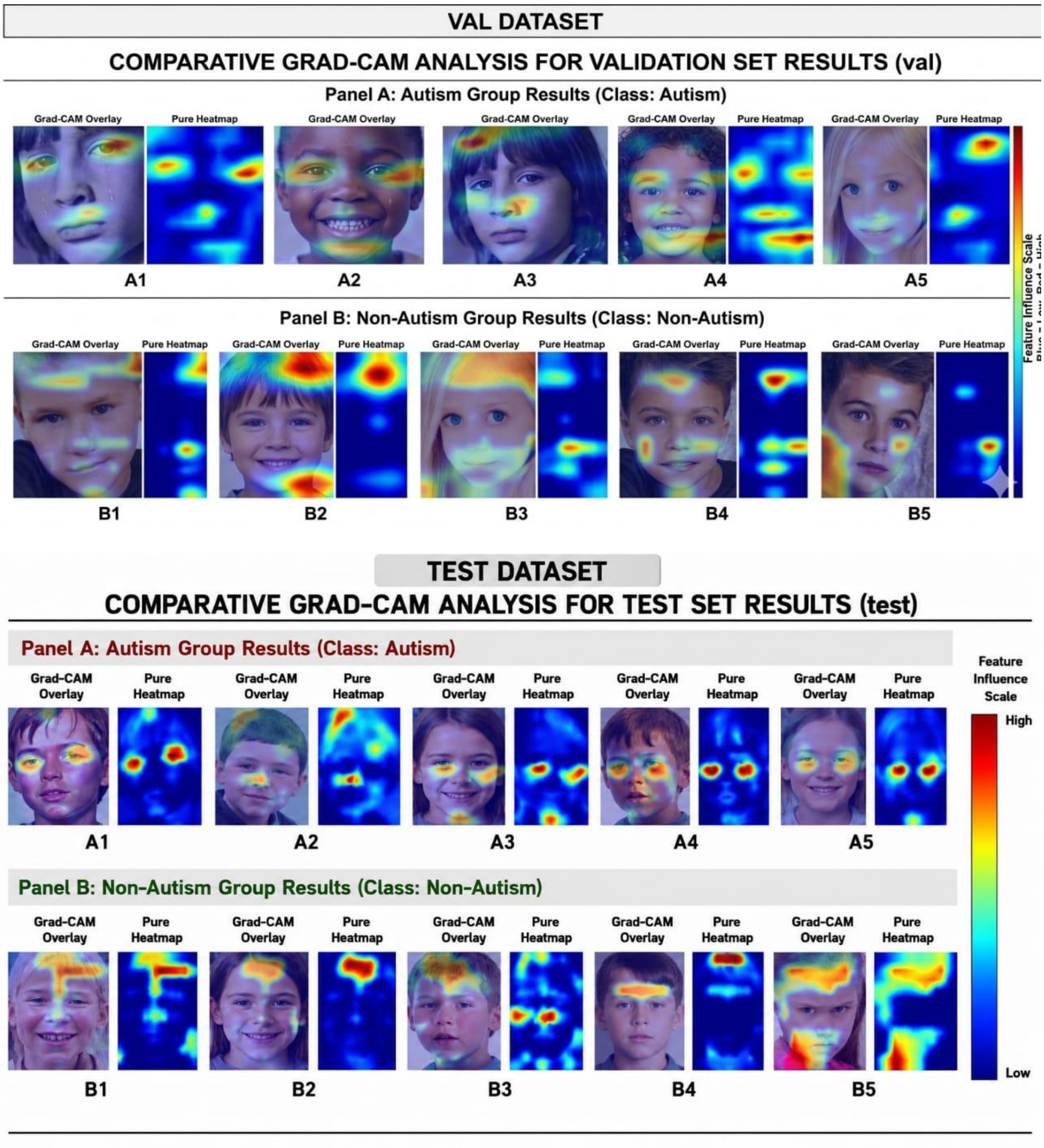

Figure 3.6. Grad-CAM visualizations for facial image samples showing attention heatmaps for ASD and non-ASD classifications

This figure illustrates the interpretability analysis of the proposed model using Gradient-weighted Class Activation Mapping (Grad-CAM), applied to eight representative facial samples drawn from two distinct classification groups.

Panel A (n=4) presents examples from the Autism group, where the generated heatmaps reveal that the model predominantly directs its attention toward the periocular region (eyes and surrounding area) and the forehead, with scattered high-activation zones distributed across facial landmarks. This pattern suggests that the network has learned discriminative features associated with subtle morphological characteristics specific to this group.

Panel B (n=4) presents examples from the Non-Autism (control) group, in which the attention maps demonstrate a more spatially distributed activation pattern, encompassing the forehead, cheeks, and chin regions, indicating a broader reliance on global facial structure for classification decisions.The color scale ranges from blue (low feature influence) to red (high feature influence), providing a visual representation of each region's contribution to the final classification outcome. The consistency of attention patterns within each group supports the model's ability to capture group-specific facial features and enhances the transparency and trustworthiness of the classification process.

3.9 Discussion

The achieved performance of 81.00% accuracy suggests that the CNN successfully learned discriminative facial features distinguishing children with ASD from typically developing peers. The four-layer architecture (32, 64, 128, 256 filters) enabled hierarchical feature extraction: initial layers likely captured basic facial elements (edges, skin texture), intermediate layers identified facial components (eyes, nose, mouth configurations), and deeper layers recognized complex emotion expression patterns characteristic of ASD.

The model's superior specificity (87.33%) compared to sensitivity (77.33%) implies that non-ASD facial patterns are more consistently distinguishable than ASD patterns. This aligns with clinical observations that ASD presents with heterogeneous facial expressivity. The variability in ASD presentations likely contributes to the higher false negative rate, as the model encounters ASD cases with subtle or atypical facial features not well-represented in the training data.

3.10 Conclusion and Future Work

3.10.1 Key Findings

This research successfully built and evaluated a deep learning-based facial analysis system for autism screening using a custom CNN. The principal findings are: (1) A four-layer CNN with filters (32, 64, 128, 256) achieved the best performance with 81.00% test accuracy; (2) The model demonstrated 85.93% precision and 81.40% F1-score, indicating reliable ASD detection; (3) Dropout regularization successfully prevented overfitting; (4) Grad-CAM visualization confirmed that the model focuses on clinically relevant facial regions (eye area) for ASD detection.

3.10.2 Practical Implications

The proposed system offers potential for assisting psychologists, educators, and pediatricians in autism screening programs. By automatically analyzing facial photographs, the tool can: (1) Prioritize clinical referrals through high-precision predictions; (2) Standardize initial screening by reducing subjective variability; (3) Enable remote assessment in telemedicine contexts. However, ethical protocols must ensure that all positive predictions receive confirmatory professional assessment, and data privacy must be strictly maintained for children's facial images.

3.10.3 Future Directions

Several avenues for future research are proposed: (1) Advanced architectures: Implementing Vision Transformers (ViT) or hybrid CNN-Transformer models, and transfer learning using pretrained models (ResNet-50, EfficientNet); (2) Temporal analysis: Extending to video-based analysis using LSTM/GRU combined with CNN feature extractors; (3) Real-time deployment: Building lightweight architectures (MobileNet, ShuffleNet) for mobile devices; (4) Multi-class assessment: Moving beyond binary classification to distinguish ASD severity levels; (5) Dataset expansion: Collecting larger, demographically diverse datasets; (6) Explainable AI integration: Incorporating SHAP or attention visualization for enhanced clinical interpretability.

General Conclusion

This Dissertation bridges the principles of artificial intelligence with the pressing clinical need for supportive tools in the screening of Autism Spectrum Disorder (ASD) in children. The research successfully built and evaluated an intelligent model based on Convolutional Neural Networks (CNNs) for facial image analysis. Experimental results demonstrated that the proposed system is capable of distinguishing between facial patterns associated with ASD and non-ASD with promising accuracy and balanced precision-recall performance. Furthermore, Explainable AI techniques (Grad-CAM) confirmed that the model focuses its attention on clinically relevant facial regions—particularly the eye area—aligning with established medical literature on atypical visual social attention in autism.

However, the following points must be emphasized with utmost rigor:

First, the system built in this study is not an independent diagnostic tool under any circumstances, and in no way substitutes for specialized medical evaluation, which includes comprehensive clinical interviews, detailed developmental history, and standardized behavioral assessments. It is intended solely as a supportive screening aid to expedite preliminary triage, reduce subjective inter-rater variability, and facilitate access to specialists in underserved or remote regions.

Second, while the achieved results are promising, they remain constrained by the dataset size, limited demographic diversity, and the binary classification nature of the model, which does not necessarily reflect the continuous spectrum of autism severity. Consequently, any positive output generated by the system must mandatorily be referred to a qualified psychiatrist or behavioral developmental specialist for confirmatory diagnosis, severity level determination, and appropriate therapeutic intervention planning.

Third, this study underscores the strict necessity of adhering to research ethics standards, particularly regarding the protection of children's biometric data, ensuring fairness across different demographic groups, and preventing any irresponsible commercial or diagnostic misuse of the technology.

In conclusion, this thesis contributes to delineating a promising approach that balances computational accuracy with clinical responsibility, and opens avenues for future research aimed at implementing more advanced architectures (such as Vision Transformers and transfer learning models), expanding datasets to encompass broader ethnic and age diversity, and moving toward temporal video analysis for a deeper understanding of facial behavioral patterns. The ultimate goal remains to serve children and their families by enhancing the quality of developmental healthcare—not by replacing human judgment, but by empowering it with more accurate, reliable, and intelligent tools.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[3] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders (DSM-5), 5th ed. Washington, DC, USA: APA, 2013.

[4] A. Lu and M. Perkowski, "Deep Learning Approach for Screening Autism Spectrum Disorder in Children with Facial Images and Analysis of Ethnoracial Factors in Model Development and Application," Brain Sciences, vol. 11, no. 11, p. 1446, 2021.

[5] W. Jones and A. Klin, "Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism," Nature, vol. 504, no. 7480, pp. 427–431, 2013.

[6] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2021.

[7] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248–255.

[8] S. J. Rogers and G. Dawson, Early Start Denver Model for Young Children with Autism. New York, NY, USA: Guilford Press, 2010.

[9] G. Dawson et al., "Early behavioral intervention is associated with normalized brain activity in young children with autism," J. Am. Acad. Child Adolesc. Psychiatry, vol. 51, no. 11, pp. 1150–1159, 2012.

[10] World Health Organization, "Autism Spectrum Disorders," WHO Fact Sheets, Mar. 2023.

[11] S. M. Anwar et al., "Medical image analysis using convolutional neural networks: A review," J. Med. Syst., vol. 42, no. 11, p. 226, 2018.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012, pp. 1097–1105.

[13] O. Kaya, A. F. M. S. Shah, M. A. Karabulut, S. N. Karahan, M. S. Osmanca, and N. Acır, "Explainable Artificial Intelligence (XAI): Concepts, Applications, Challenges, and Future Perspectives," IEEE Access, vol. 14, pp. 1-28, 2026.

[14] E. Kyrimi et al., "Explainable AI: definition and attributes of a good explanation for health AI," AI and Ethics, vol. 5, pp. 3883–3896, 2025.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.

[16] M. Ennab and H. Mcheick, "Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models," Mach. Learn. Knowl. Extr., vol. 7, no. 12, pp. 1-32, Feb. 2025.