VirtuAlly: Early Detection of Virtual Autism in Toddlers

Anushka Korlapati
anushka22085@iiitd.ac.in
IIIT Delhi
Delhi, India

Parth Garg
parth22351@iiitd.ac.in
IIIT Delhi
Delhi, India

Shagun Yadav
shagun22466@iiitd.ac.in
IIIT Delhi
Delhi, India

Abstract

With the growing prevalence of screen exposure among young children, concerns have emerged about its impact on early development. Notably, excessive screen time has been increasingly linked to Virtual Autism—a condition where toddlers begin to display Autism-like symptoms due to limited real-world social interaction. Virtu-Ally is a one of it's kind AI-powered mobile application designed to detect early signs of Virtual Autism through multimodal analysis of a child's behavior during screen interactions. The system leverages a DenseNet121-based CNN for facial expression classification, Whisper for accurate speech transcription, and Gemini 2.0 for clinically-oriented behavioral evaluation. Built with a user-centric approach, the application ensures transparency and child-friendliness, offering explainable insights through Grad-CAM and timestamp-based reasoning for speech. Evaluated on a dataset of labeled images and child speech samples, Virtu-Ally is supported by strong test accuracy, F1, ROC AUC scores. This shows its strong real-world potential as an early intervention tool, enabling parents to make timely, informed decisions about their child's development.

ACM Reference Format:
Anushka Korlapati, Parth Garg, and Shagun Yadav. 2025. VirtuAlly: Early Detection of Virtual Autism in Toddlers. In *Proceedings of Human-Centered AI (HCAI'25).* ACM, New York, NY, USA, 11 pages. https://doi.org/virtu-ally

1 Problem Statement

In recent years, excessive screen time in children aged 3 to 5 has been linked to Virtual Autism—ASD-like symptoms caused by reduced real-world interaction. Virtu-Ally is an AI-powered application that analyzes facial expressions and speech to detect early signs of this condition during screen use.

2 The Problem

Screens have become a big part of children's lives, especially after COVID-19 when more kids spent time on devices for entertainment and learning. Unfortunately, this excessive screen time has led to a condition called Virtual Autism where children show signs similar to autism like difficulty with communication and social interaction. Parents often don't notice these signs early and finding help can be slow and costly.

HCAI'25, Okhla, DEL

ACM ISBN 085-351-466-02/2025
https://doi.org/virtu-ally

3 The Solution: Virtu-Ally | AI-Powered Detection of Virtual Autism

Instead of fighting against screens, this project aims to use them to detect early signs of Virtual Autism. Virtu-Ally leverages AI to analyze a child's on-screen behavior and provide parents with early warnings about potential developmental concerns. The key features of Virtu-Ally include:

- **Facial Expression Recognition:** We periodically capture images of the child and use a DenseNet121-based model to classify them as Viturally Autistic or not.
- **Speech Based Evaluation:** As the child reads out phrases shown on our application, we use whisper to transcribe the audio file and Gemini 2.2 Flash model to analyse and classify the audio as Viturally Autistic or not.
- **Parental Insights:** Parents receive personalized feedback and recommendations to support their child, such as adjusting screen time or increasing real-world interactions.

Virtu-Ally will integrate these to develop a smart application that helps parents identify early indicators of Virtual Autism. By combining technology with early intervention strategies, our goal is to empower caregivers with actionable insights, enabling them to make informed decisions for their child's well-being. Instead of seeing screens as the problem, Virtu-Ally turns them into a tool for early detection, helping children get back on track faster.

3.1 Novelty of the Solution

A diagnostic tool disguised under a gamified interface which records data non-intrusively. Lack of literature in the domain of virtual autism, a formal tool for diagnosis is even harder to find: our research gap. Usage of multimodality which increases the reliability and trust in the evaluations. Architecture uses multimodal approach with focus on explainability. An early intervention which is most importantly built with the user-first strategy; usability is the first goal. Diagnosis tools may exist but parents are reluctant to use those due to the stigma around the mental health; our approach steers away from a stress-inducing medical environment to a child-friendly gamified interface from the comfort of their home.

4 Literature Review

4.1 Impact of Excessive Screen Exposure on Communication Skills

The rise of digital media consumption has led researchers to investigate its effects on early childhood development, particularly in relation to communication and social interaction. "Virtual Autism" is a term used to describe autism-like symptoms in children exposed to excessive screen time at a young age. A study by Yadav et al. (2024) examined a cohort of children aged 1.5 to 6 years who

were exposed to more than four hours of screen time per day. The findings indicated significant delays in language acquisition, reduced eye contact, and impaired social engagement, which closely resemble symptoms associated with Autism Spectrum Disorder (ASD).

However, the study also explored the reversibility of these effects. After a structured intervention involving a drastic reduction in screen exposure, increased parent-child interaction, and speech-language therapy, many children exhibited noticeable improvements in their communication and socialization skills. This suggests that excessive screen time during early developmental years can contribute to ASD-like behaviors, but timely intervention can potentially mitigate its impact. These findings emphasize the importance of regulating screen time for children, particularly during critical periods of cognitive and linguistic development.[1]

4.2 Distinguishing Virtual Autism from ASD

In the thesis titled "*Virtual Autism or Autism? How can we prevent misdiagnosis?*" by Walaa Almusawi (2024), the author investigates the potential for misdiagnosing autism spectrum disorder (ASD) in children who exhibit autism-like symptoms due to excessive early exposure to digital media. The study reviews existing literature linking prolonged screen time to behavioral, developmental, and cognitive issues in children, suggesting that such exposure can impair cognitive and language development, leading to symptoms that mimic ASD. Almusawi discusses interventions like "electronic fasting" or digital detoxification, which involve removing screens and digital content from children's environments. These interventions have been shown to mitigate autism-like symptoms and improve overall mental well-being, including addressing sleep and nutritional issues. To prevent misdiagnosis, the thesis suggests integrating behavioral history assessments into ASD diagnostic procedures. Evaluating a child's exposure to digital media and their response to screen-time reduction could help differentiate Virtual Autism from ASD, ensuring that children receive appropriate interventions rather than unnecessary ASD-specific treatments. [2]

4.3 Virtual Autism Evaluation

Given the challenges in diagnosing Virtual Autism, researchers have explored alternative assessment methods that leverage digital tools. One of the most promising approaches is the use of telehealth-based assessments, such as the TELE-ASD-PEDS (TAP) model. This method allows clinicians to evaluate children remotely by having parents record structured video interactions of their child performing specific tasks or engaging in social behaviors.

The TAP framework is particularly beneficial in regions where access to specialized autism diagnostic services is limited. By enabling experts to analyze a child's behaviors remotely, the system improves early detection rates and ensures timely intervention. Studies indicate that remote evaluations using TAP have comparable accuracy to in-person clinical assessments, making them a viable alternative in resource-constrained settings.

Moreover, integrating AI-driven behavioral analysis tools into telehealth assessments could further enhance diagnostic precision. Machine learning models can analyze video recordings to detect subtle social and communicative impairments that may not be immediately apparent to human evaluators, thus improving the reliability of Virtual Autism screening.[3]

4.4 Autism Screening Using AI

The paper "Autism AI: A New Autism Screening System Based on Artificial Intelligence" by Shahamiri & Thabtah (2020) presents an AI-powered autism screening system that replaces traditional scoring-based methods with a deep learning model. The system utilizes a convolutional neural network (CNN) trained on a large dataset of ASD cases and controls, achieving higher accuracy, sensitivity, and specificity compared to conventional screening tools. Unlike traditional diagnostic methods, which rely on lengthy questionnaires and expert evaluations, this AI-driven approach offers a faster, more accessible, and automated pre-diagnosis system via a mobile application. The study also compares existing autism screening apps, highlighting the superiority of AI-based detection over static rule-based scoring. The results suggest that AI can significantly improve early autism detection, making screening more efficient and widely available. [4]

4.5 Autism Screening Framework

This paper presents a cloud-based automated system for autism screening and confirmation, targeting children aged 0-17 years. The framework is divided into three stages:

(1) **Screening Phase:** Parents use a mobile application to complete a pictorial questionnaire about their child's behaviors. AI algorithms analyze responses to determine whether further evaluation is needed.
(2) **Virtual Assessment:** If screening results indicate possible ASD traits, parents record videos of their child interacting in different scenarios. These videos are analyzed by experts using machine learning models that assess nonverbal cues, eye contact, and facial expressions.
(3) **Final Confirmation:** If autism is suspected, the child is referred to an ARC for an in-person assessment by trained clinicians.

This approach aims to streamline autism detection, reduce dependence on expert availability, and increase accessibility, particularly in developing countries with limited resources. The system integrates AI-driven decision-making and cloud storage to enhance efficiency and early intervention possibilities. [5]

4.6 Application of Machine Learning in Autism Detection

The study by Liao et al. (2022) titled *"Application of Machine Learning Techniques to Detect Children with Autism Spectrum Disorder"* explores the use of artificial intelligence (AI) to enhance autism diagnosis. The research focuses on leveraging machine learning (ML) techniques to improve the accuracy and efficiency of autism screening. The paper introduces a model that processes behavioral and physiological data from children to predict ASD presence. The methodology includes:

- **Feature Extraction**: The system collects behavioral traits such as social responsiveness, eye gaze patterns, and repetitive behaviors.

- **Machine Learning Classifiers**: The study compares multiple ML models, including Support Vector Machines (SVM), Decision Trees, and Neural Networks, to determine which approach provides the highest diagnostic accuracy.
- **Performance Evaluation**: Experimental results demonstrate that ML-based models outperform traditional questionnaire-based ASD assessments in accuracy and speed.

This research highlights the potential of AI-driven diagnostic tools in early autism detection, advocating for a shift from conventional diagnostic approaches to more data-driven, automated methodologies. The findings support the integration of ML in clinical practice to provide faster and more precise autism screening. [6]

4.7 Facial Analysis for Autism Detection

The study "VITASD: Robust Vision Transformer Baselines for Autism Spectrum Disorder Facial Diagnosis" by Cao et al. (2023) presents a Vision Transformer (ViT)-based approach for detecting Autism Spectrum Disorder (ASD) using facial images. The proposed ViTASD model extracts features from pediatric patients' facial expressions and employs a Gaussian Process layer to enhance robustness in ASD analysis. Compared to traditional CNN-based models, ViTASD offers superior accuracy and interpretability, achieving state-of-the-art performance on ASD facial analysis benchmarks.

One of the key advantages of ViTASD is its robustness against variations in lighting, pose, and facial expressions, making it more reliable for real-world applications. The study also incorporates a Gaussian Process layer to quantify uncertainty in model predictions, ensuring that cases with low confidence scores can be flagged for further clinical evaluation.

Experimental results demonstrate that ViTASD outperforms previous deep learning models in ASD facial recognition tasks, achieving state-of-the-art performance on benchmark datasets. This highlights the potential of non-invasive, image-based ASD diagnostics, which could be integrated into telehealth platforms for rapid screening. [7]

5 Methodology

This section includes all of our user research and design techniques used in the development of Virtu-Ally.

5.1 User Research

To understand the problem space and the impact of excessive screen time on young children, we conducted secondary research on Virtual Autism and its symptoms in children aged 3–5. Our key stakeholders include Toddlers, Parents, Early Educators and Health Care Professionals.

From our research, we developed the following POVs:

5.2 Point of View Analysis

Parents of **toddlers** with excessive screen exposure **need a way to** become aware of early signs of Virtual Autism in a reliable and timely manner **because** traditional screening methods can be subjective and delayed.

- **We met parents** concerned about their child's social and communication development but unsure whether excessive screen time is the cause.
- **We were surprised** to see how difficult it is to timely diagnose Virtual Autism without expert intervention, leading to misdiagnoses and anxiety.
- **We wonder if** an AI-driven solution could provide objective behavioural analysis to help parents make informed decisions about screen time and developmental milestones.
- **It would be game-changing** to develop an AI-powered tool that detects early signs of Virtual Autism using facial expression recognition parents, actionable insights, and early intervention strategies.

This led us to our guiding **Empathy Maps** and **How Might We (HMW)** question:

Toddlers: The primary users of the application whose interactions and gaze patterns provide critical data for predicting virtual autism and improving early intervention methods.

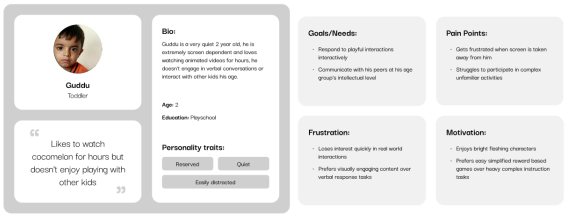

Figure 5.2.1: User Persona - Toddler

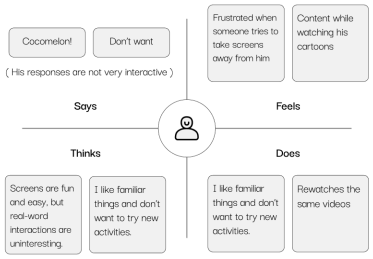

Figure 5.2.2: Empathy Map - Toddler

HMW for Toddlers:

(1) **HMW make AI-based screening for children feel like a fun and engaging activity rather than a test while still extracting meaningful data for detecting early signs of autism?**
We can make the screening process enjoyable and stress-free for children by using interactive games and animations. In the background, we collect visual data of children's facial expressions for diagnosis.

(2) **HMW create a non-intrusive, child-friendly, safe, accessible and easy-to-use tool for detecting signs of virtual autism?**
Children are familiar and comfortable with using screens - which are ubiquitous in every household. Hence we might use screens as a mode of interaction, which will passively monitor their facial expressions. This should provide a seamless and non-intrusive experience and child-friendly method for the detection of Virtual Autism.

Parents of Toddlers: Parents play the most important role in their child's early development and are often the first to notice behavioral changes that could indicate virtual autism.

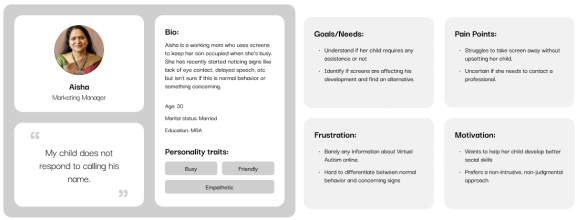

Figure 5.2.3: User Persona - Parent

Figure 5.2.4: Empathy Map - Parent

Early Educators: Teachers and caregivers in preschools and daycare centers interact with toddlers daily making them well-positioned to observe signs of attention issues.

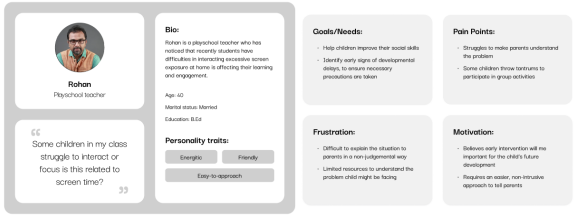

Figure 5.2.5: User Persona - Early Educator

HMW for Parents and Educators:

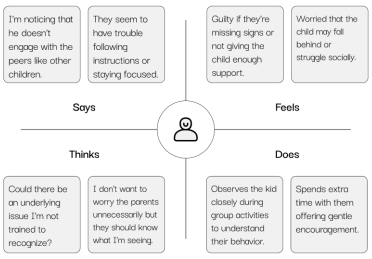

Figure 5.2.6: Empathy Map - Early Educator

(1) **How might we gain parent's trust in such a tool considering it involves collecting data involving their children?**
We plan on obtaining parent consent before collecting data and give them full access and transparency on how their children's data will be processed and to verify AI's evaluation with formal evaluation by a medical professional.

(2) **How might we ensure that our AI tool for parents and educators offers guidance without causing unnecessary fear or guilt?**
By providing clear insights and explaining the reasoning behind them, we can offer parents and educators proactive steps to adjust screen habits, ensuring their child's health without unnecessary fear or guilt. We need to offer reassurance and not alarm.

(3) **How might we explain technical insights to parents and educators that help them understand the child's development?**
Through simple and easy-to-read reports that summarize our key findings in simple language, we let the parents and educators know of our preliminary findings if they should consult a paediatrician or not and give them small actionable insights to benefit the child.

Health Care Professionals: Pediatricians, child psychologists, and therapists provide expert diagnosis and intervention strategies, helping validate our application's effectiveness in early detection.

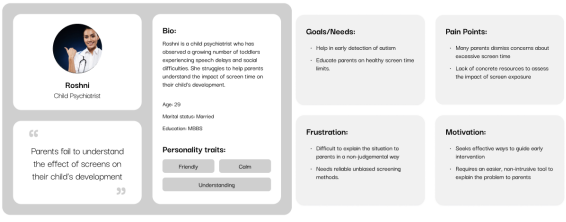

Figure 5.2.7: User Persona - Health Care Professional

HMW for Health Care Professionals:

(1) **How might we distinguish between diagnosis of autism and virtual autism?**

Virtual Autism is a reversible condition, whereas Autism isn't. Since symptoms of Autism and Virtual Autism are very similar we can provide a preliminary evaluation subject to formal evaluation for autism by medical healthcare professionals. If autism is not present through medical diagnosis, we can conclude it is a case of virtual autism.

(2) **How might we create AI-generated reports for doctors that are concise, reliable, and easily verifiable through their expertise?**
The reports we generate for healthcare professionals, unlike the ones we generated for the parents, should be detailed and technical and contain our collected data along with identified signs of Virtual Autism and their reasoning/metric, allowing doctors to verify results easily while ensuring transparency.

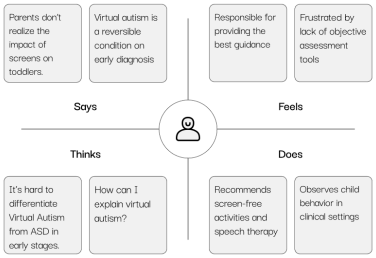

Figure 5.2.8: Empathy Map - Health Care Professional

5.3 User Flows

Based on the user needs and stakeholder goals, we mapped a simple user flow:

The task flow in **(Fig. 5.3.1)** represents the toddler's interaction with the application. It outlines how a child engages with interactive games/videos to assess behavior and responsiveness.

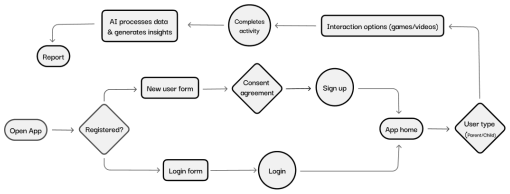

Figure 5.3.1: Task Flow - Toddler

The toddler's task flow **(Fig 5.3.2)** focuses on interactive play, where the child engages in games that assess eye contact, attention, and responsiveness. The AI silently tracks behaviour, processes insights, and sends stores a detailed report that can be accessed by the parent

The parent task flow **(Fig 5.3.3)** is centered on monitoring and analysis. Parents log in, provide consent, and input child details. After the child completes activities, the AI generates a comprehensive report with behavioural insights and recommendations, accessible on the parent's dashboard. Based on these insights, parents can choose to consult a healthcare professional for further evaluation and guidance.

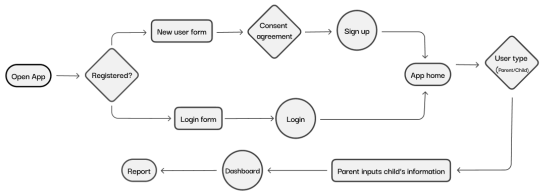

Figure 5.3.2: Task Flow - Parent

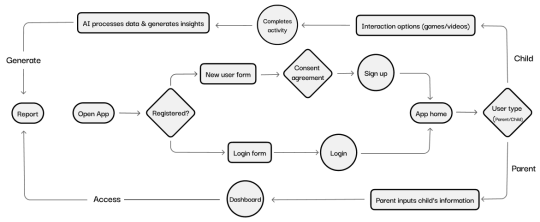

Figure 5.3.3: Combined Task Flow

The combined user task flow **(Fig 5.3.3)** visualizes a smooth process where parents manage setup and reports while children engage in gamified assessments. AI tracks behaviour, generating insights for parents to support early detection and informed decisions on Virtual Autism.

5.4 Design Process

We followed an iterative design process starting with low-fidelity sketches to outline core features and screen structure. These were refined into mid-fidelity wireframes focusing on usability and layout. Finally, we developed high-fidelity mockups showcasing the final UI, emphasizing child-friendly visuals and a clean, informative parent dashboard.

5.4.1 ***Low-Fidelity Prototype.*** The prototype (Fig. 5.4.1) illustrates the core screens and user journey, ensuring an intuitive and user-friendly experience for both parents and children. The design focuses on usability, accessibil- ity, and AI-driven behavioral assessment while maintaining a child-friendly interface for the children. For the parents, focus is on making the data insights simple to comprehend and easy to understand. Focus has also been placed on ensuring informed consent from parents to collect and process their children's data. Transparency on how data will be processed is also an important feature in line with the People and AI guidelines.

5.4.2 ***Mid-Fidelity Prototype.*** The mid-fidelity prototype (Fig. 5.4.2) & (Fig 5.4.3) refine the layouts established in the low-fidelity version, providing a more accurate visual structure. It introduces interactivity and incorporates basic UI elements, offering a more tangible and functional representation of the final product to facilitate

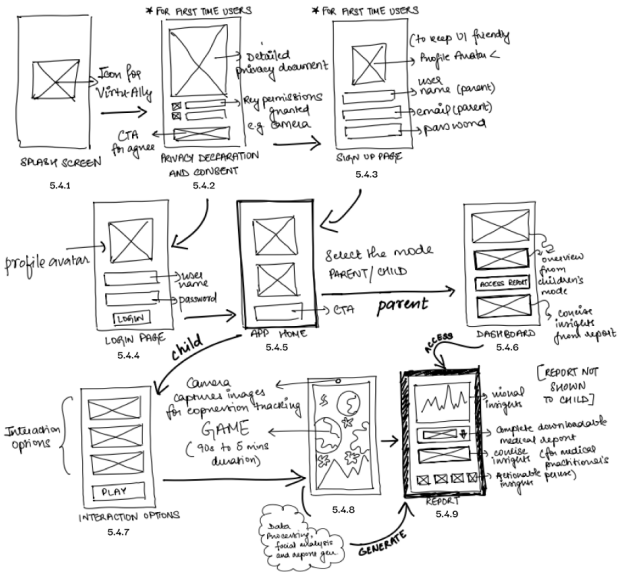

Figure 5.4.1: Lofi Prototype

user testing of VirtuAlly's core functionalities. This implementation serves as a critical tool for evaluating our design in real-world contexts with actual users, prioritizing usability principles particularly relevant for the parents of toddlers and toddlers themselves. The focus is on enhancing the user experience of the interaction while keeping the cognitive load to a minimum.

Key Screens & Flow:

- **Splash Screen:** Displays the app logo and a welcoming loading screen designed to mimic the interface of a game, making children comfortable using the application as they would with any other mobile game.
- **Sign-Up Page:** Parents enter basic details such as name, email, and password. The interface features a user-friendly profile avatar to enhance personalization and ease of onboarding.
- **Login Page:** Returning users can enter their credentials to access the app. Authentication is conducted via email, and the process is kept short and intuitive to ensure a seamless user experience.
- **Privacy & Consent:** First-time users are shown a privacy declaration outlining permissions required for AI-driven analysis, such as camera access. Special attention is given to **informed consent**, building trust through complete transparency about data processing, storage, and usage. Parents can opt out at any point.
- **App Home:** Users select their mode: Parent (dashboard access) or Child (interactive activities). Access to the Parent mode is password-protected to ensure children do not view sensitive information without supervision or consent.
- **Dashboard [Parent Mode]:** Offers a concise overview of the child's behavioral assessment, providing access to detailed reports. The dashboard prioritizes clarity and uses simple language to communicate findings. Emphasis is placed on explaining observed behaviors in a non-alarming, empathetic manner to foster a supportive environment.
- **Interaction Options [Child Mode]:** Presents gamified tasks designed to engage the child in behavioral assessments. The experience is akin to playing a regular mobile game, with passive data collection via the device camera ensuring minimal distraction or discomfort.
- **Game Session:** During these sessions, the application uses the camera to capture facial expressions and verbal responses. AI models analyze factors such as engagement, attention span, and responsiveness. Due to mobile device limitations, data processing is offloaded to the cloud using vision transformer-based models.
- **Report Generation [Parent Mode]:** The AI compiles all observations into a comprehensive behavioral report. Reports include visual insights, risk analysis, and tailored recommendations. They are designed for parental and professional use only and are not shown to the child, preserving a positive and pressure-free environment.

Minor usability adjustments were made based on user feedback and incorporated into the high-fidelity.

Figure 5.4.2: Mid-fi Prototype

Figure 5.4.3: Mid-fi Prototype

5.4.3 ***High-Fidelity Prototype.*** Final screens include branding, icons, and color. The design prioritizes clarity for parents and engagement for children. The report screen features risk-level, recommended course of action and facial expression and speech modality

flow results We developed a functional **Android** application. It is built using **Jetpack Compose** in **Android Studio** to facilitate user testing of VirtuAlly's core functionalities. Android application is designed following a **user-centered design approach**.

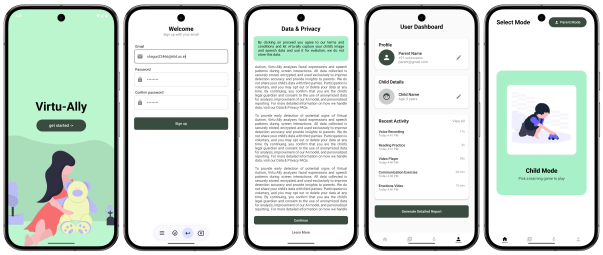

Figure 5.4.4: Hi-fi Prototype

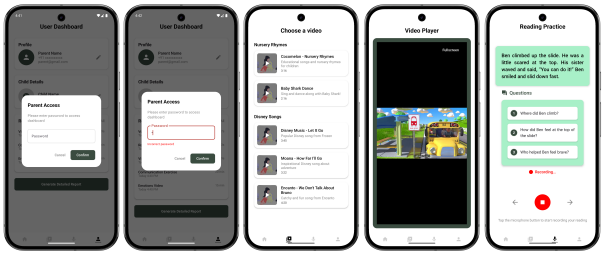

Figure 5.4.5: Hi-fi Prototype

5.5 Application Architecture and Navigation

The application architecture employs a **navigation-driven compositional structure** using the **Navigation Component for Jetpack Compose** (`androidx.navigation`). This implementation follows the **MVVM (Model-View-ViewModel)** architecture pattern, ensuring a clean separation of concerns and enhanced testability. Utilizing Jetpack Compose's **declarative UI paradigm**, we created a cohesive and enjoyable user experience across multiple screens. The navigation system leverages `rememberNavController()` and **state hoisting principles** for maintaining UI consistency.

5.6 User Onboarding and Mode Selection

The application starts with an **onboarding flow** that includes the following screens:

- **WelcomeScreen**
- **SignUpScreen**
- **LoginScreen**

This flow guides users through a contextually appropriate **data privacy consent** process, which strengthens user trust in the product. After authentication, the system presents a **mode selection interface** that allows different user roles (Child and Parent). The UI is built with **Material Design components** and follows the **theming system** for accessibility and consistent visual feedback.

5.7 Behavioral Monitoring

A key innovation in our implementation is the seamless integration of **behavioral monitoring capabilities** within the **video playback functionality**. While children engage with content on the **VideoPlayerScreen** component, the application captures visual frame data in the background using a **specialized frame capture system**. This system is built with **Android's MediaCodec and Surface APIs** and executed within `Dispatchers.IO` coroutine context for optimal performance.

These captured images are processed through our **DenseNet121-based CNN**, designed to detect early behavioral indicators of **virtual autism** in toddlers. This functionality utilizes **TensorFlow Lite for Android** for efficient **on-device inference**. This passive data collection approach follows **calm technology principles**, allowing for meaningful assessment without disrupting the child's natural viewing experience or adding cognitive load for parents.

For the **speech pattern analysis**, the application displays a series of age-appropriate phrases and questions on the screen for the child to read aloud. The child's spoken responses are recorded in real-time and saved as `.mp3` files. These audio files are then transcribed using **Whisper AI**, ensuring accurate and efficient conversion of speech to text. The transcriptions, along with the audio data, are processed by **Gemini 2.0 Vision**, which analyzes both the **verbal content** and **underlying speech patterns**. Based on this analysis, the system determines whether the observed behavior aligns with signs of Virtual Autism, offering another reliable layer of behavioral insight.

5.8 User Experience and Responsive Design

The **video playback functionality** supports both **portrait and landscape orientations**, leveraging **Android's Configuration API** and **ActivityInfo** for effective orientation management. This ensures usability across different viewing contexts, aligning with **universal design principles**. Furthermore, the **user dashboard** follows **information architecture principles** to organize child data, activity tracking, and reports in a cognitively accessible manner. The dashboard optimizes performance by employing **lazy loading** via **Compose's LazyColumn**.

5.9 Technology Stack and Security Considerations

Our implementation uses the following:

- **Kotlin Coroutines** for asynchronous operations
- **Flow** for reactive programming patterns
- **Hilt** for dependency injection
- **Android Security Best Practices** for handling sensitive camera data

These ensure a modular, maintainable, and secure codebase.

5.10 Iterative User Testing and Improvements

Through **iterative user testing** using the mid-fidelity prototype, we gathered valuable insights into **interaction patterns**, **cognitive load**, and **emotional responses**. Based on the feedback received, we implemented several **key changes** to enhance usability and ethical design. These include:

- **Incorporating a detailed privacy notice** to ensure transparency and build user trust.
- **Introducing password protection for accessing Parent Mode** to prevent accidental access by children.
- **Adding speech analysis as a second modality** alongside facial recognition, creating a more comprehensive and accurate detection system.

This multimodal approach to **behavioral detection**—combining engaging video content with **unobtrusive computer vision and audio analysis**—exemplifies the value of **embodied interaction** in advancing **Human-Centered AI (HCI)** research. These improvements not only make the system more robust but also align with ethical design principles focused on child safety, privacy, and early intervention.

5.11 HCI Principles

The design and implementation of Virtu-Ally takes into account Human-Computer Interaction (HCI) principles to ensure usability, trust, and ethical engagement—particularly for its sensitive user group: toddlers and parents.

- **Child-Safe Application:** The application follows usability principles and restricts sensitive information from children using password protection on the parent mode. This ensures that the child gets a gamified experience. This reduces the cognitive load on the toddler.
- **Data/Privacy Disclosure:** Users are explicitly informed of the data collection and user consent is obtained. This aligns with values of transparency and user control. Our unique TLDR box summarises the user permissions being obtained.
- **Gamified UI for toddlers:** The application is stylized with a game like interface for toddlers while it is saliently being used as a non-invasive diagnostic tool. It follows principles of calm-technology and focuses on keeping the app decluttered, usable and easily learnable for toddlers and parents alike.
- **User Customisation:** Once securely logged into the parent mode, parents can modify personal details of the child, the app follows transparency by informing the parent of what personal data has been recorded through recent activity.
- **Preventing Errors:** The applications has failsafes implemented against faulty input and handles all user inputs robustly with proper feedback on why the error occured.
- **Learnability:** The application ensure that toddlers and parents are able to learn and navigate the application through intuitive affordances. The nav bar helps with the same.
- **Consistency:** The application uses a consistent and child friendly design scheme using material UI design assets.
- **Usability:** Focus on child icon through visual cues like colour and size helps in directing attention of the child towards relevant section.
- **Trust:** The evaluation follow human-first approach focussing on making the users trust the report. A friendly environment is encouraged.
- **Explanability:** The evaluation results are reported to the user through a simple, interpretable report that explains the evaluation results and gives insights to interpret them.
- **Inclusivity and Empathy:** The reports ensure that it does not put create a hostile environment and thus uses terms like typical/atypical behaviour rather than the medical terminology.
- **Medical Follow-up:** Parents are encouraged to follow up the with a medical professional post the virtually evaluations, this is will ensure timely medical intervention if needed.

5.12 Image Modality

The implementation uses a DenseNet121 model with slight modifications to enhance stability and prevent overfitting. [9]

5.12.1 ***Dataset.*** The model is trained on a dataset of 2940 images labeled as autistic or not autistic. The dataset is split into:

- **Train:** 2058 images (1029 autistic, 1029 non-autistic)
- **Validation:** 441 images (221 autistic, 220 non-autistic)
- **Test:** 441 images (220 autistic, 221 non-autistic)

Figure 7.1: Example Images from the Dataset

5.12.2 ***Preprocessing.*** The images undergo several preprocessing steps to enhance the model's performance. First, the pixel values are normalized to bring them within the range of [0,1] making it easier for the model to learn and improving training efficiency. To add variety and prevent overfitting, the training images are randomly rotated, shifted and flipped horizontally. This helps the model generalize better by simulating real-world variations in image alignment. If any pixels are missing due to these transformations, they are filled with the nearest pixel values to preserve image consistency. The validation and test images are only rescaled without any modifications to ensure they reflect the original data distribution. These preprocessing steps strengthen the model by making it more resilient to slight changes in image positioning and orientation.

5.12.3 ***Model Architecture.*** The model utilizes the pre-trained **DenseNet121** architecture [8], which is known for its efficient feature propagation and reuse due to densely connected layers. DenseNet121 is pre-trained on the ImageNet dataset and used as a base model for feature extraction. The input shape of the model is set to (224, 224, 3) to accommodate the dimensions of the image dataset.

After the base model, a **Global Average Pooling (GAP)** layer is applied. This layer reduces the spatial dimensions of the feature maps by computing the average value for each feature map. GAP reduces the overall number of parameters, prevents overfitting and helps retain the most essential features making the model more robust and generalizable.

Next, a **Batch Normalization** layer is added to normalize the activations of the previous layer. This stabilization technique helps

standardize the outputs of the GAP layer which improves convergence speed and enhances the model's overall performance.

Following this, a **Dense layer** with 256 units and a ReLU activation function is included. The dense layer learns high-level, abstract representations of the image features extracted by the DenseNet121 base model. The ReLU activation introduces non-linearity enabling the model to capture more complex patterns and relationships in the data.

To prevent overfitting, a **Dropout** layer with a rate of 0.6 is applied. During training, this layer randomly deactivates 60% of the neurons forcing the model to learn more robust and general features rather than memorizing specific patterns thereby enhancing generalization.

The final layer is a **Dense output layer** with a single neuron and a sigmoid activation function. The sigmoid activation outputs a probability score between 0 and 1 making it suitable for binary classification. In this case, values closer to 1 indicate **Virtual Autism** while values closer to 0 indicate a negative diagnosis.

The model is compiled using the **Adam optimizer** with a learning rate of 0.0001 and 30 epochs which adapts the learning rate during training to enhance convergence. The loss function used is **binary cross-entropy**, which is appropriate for binary classification tasks as it measures the difference between the predicted and actual labels.

The model architecture can be seen in the image below:

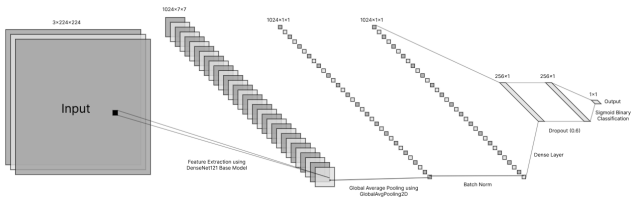

Figure 7.2: CNN Model Architecture

5.13 Speech Modality

The system uses Whisper model to transcribe a child's speech audio into text with detailed timestamps. This transcript is then analyzed using Google Gemini which provides a clinical assessment of the speech patterns for signs associated with virtual autism. [10]

5.13.1 ***Dataset.*** The Eigsti dataset, an openly available resource containing transcript files of children's speech is used. The dataset includes samples categorized into three groups:

- **TD (Typically Developing):** Children with no known developmental conditions.
- **DD (Developmental Delay):** Children showing general developmental delays, which may or may not include autism.
- **ASD (Autism Spectrum Disorder):** Children formally diagnosed with autism.

5.13.2 ***Preprocessing.*** The preprocessing begins with an MP3 file containing the child's audio. This audio is transcribed using OpenAI's Whisper model which produces accurate text along with detailed timestamps. The resulting transcript is then embedded within a structured prompt and passed to the large language model for further clinical analysis.

5.13.3 ***Model Architecture and Implementation.*** The transcribed text with timestamps is embedded into a carefully crafted prompt designed to simulate a clinical setting. The prompt used is as follows:

> You are a medical assistant. The following is a transcription of a child speaking. Please evaluate the language used, clarity, coherence, and structure. Identify any speech traits that may indicate characteristics associated with virtual autism. If the speech sounds typical, mention that clearly and report everything within 200 words.
> Here is the transcript:
> *<Transcript text>*
> Give a short, clinical evaluation with short reasoning based on the quotes above in simple language for the child's parent.

This prompt is then sent to the Gemini 2.0 Flash model via an API call to generate a clinical evaluation report. The prompt is effective for this use case because it clearly defines the role of the language model, includes a structured transcript with time context, and emphasizes accessible language for parents ensuring both clinical relevance and clarity.

The complete flow of the speech modality from audio to clinical report can be seen in the figure below:

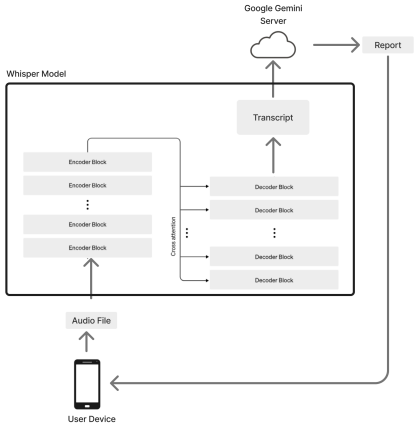

Figure 7.2: Speech Modality Flow

5.14 Explainability

To ensure transparency and build trust with users—especially in a sensitive context like child behavioral assessment—Virtu-Ally incorporates explainability mechanisms for both visual and audio-based predictions.

Facial Expression Explainability with Grad-CAM: For the facial expression analysis module, we use a DenseNet121-based

classifier trained to detect behavioral markers associated with Virtual Autism. To provide interpretability, we apply **Grad-CAM (Gradient-weighted Class Activation Mapping)** on the model outputs. Grad-CAM generates heatmaps that highlight the regions of the child's face that most influenced the model's classification. This helps parents and professionals visually validate which expressions—such as lack of eye contact, emotional flatness, or repetitive expressions—contributed to a prediction, reinforcing the model's decision-making process and increasing trust. In addition to visual explanations, the model also outputs a **confidence score** associated with each prediction. This score can be used as a supplementary metric by caregivers; for instance, if the model's confidence in an autistic classification is high, it is advisable to seek further evaluation from a qualified healthcare professional.

Speech Analysis Justification using Gemini 2.0 Flash: For the speech analysis component, we use **Whisper** to transcribe the audio input and then feed the transcriptions into the **Gemini 2.0 Flash** language model. The model is prompted to not only classify the behavior but also provide timestamp-based justifications for its decision. This allows us to highlight specific speech segments that exhibit signs of Virtual Autism or normal behavior—for instance, *monotone speech, lack of natural pauses, or limited vocabulary.* By providing this contextual reasoning, parents can understand **why** a speech segment was flagged, making the system more accountable and interpretable.

Together, these explainability tools support responsible AI usage by offering **transparent, human-interpretable feedback**, empowering caregivers to make informed decisions based on not just predictions, but also their underlying rationale.

6 Evaluation

6.1 Evaluation Metrics

The system uses the following metrics to assess its performance:

- **Accuracy:** Measures how often the model makes correct predictions.
- **Loss:** Binary cross-entropy indicating how close predicted probabilities are to true labels.
- **F1 Score:** Balances precision and recall, especially useful in handling class imbalance.
- **ROC AUC:** Reflects the model's ability to distinguish between classes across thresholds.

These metrics are suitable for virtual autism detection as they collectively capture overall accuracy, confidence, robustness to imbalance, and discrimination power.

6.2 Performance results

The model was trained for 30 epochs using the following hyperparameters: Adam optimizer with a learning rate of 1×10^{-4}, binary cross-entropy loss function and a dropout rate of 0.6 to reduce overfitting. The model achieves high accuracy with validation and test accuracies approximately reaching 80% indicating good performance and reliable generalization. Further model improvements and fine-tuning could enhance its accuracy even more. The results are summarized below along with the plots of accuracy and loss per epoch and ROC AUC curve:

Metric	Accuracy	Loss	F1 Score	ROC AUC
Train	97.86%	0.0614	0.472	0.4896
Validation	81.63%	0.7947	0.4858	0.5199
Test	80.27%	0.4806	0.7754	0.8567

Table 1: Model Performance across Datasets

The test results show that the model works well for autism detection using speech. It has a good accuracy of 80.27%, a high F1 Score of 0.7754, and a strong ROC AUC of 0.8567. These numbers mean the model is reliable at identifying speech traits linked to autism and can be helpful in early detection.

6.3 Explainability results

To improve transparency and trust in our facial expression evaluation pipeline, we utilized **Grad-CAM (Gradient-weighted Class Activation Mapping)** on the DenseNet model. As shown in Figure 7.2, the overlay heatmaps indicate that the model attends to facial areas such as the **eyes, mouth, and overall facial symmetry**, which are clinically relevant to autism diagnosis. For example, in autistic predictions, the attention maps are often focused on atypical facial expressions or disengaged eye contact. Conversely, in non-autistic predictions, the heatmaps highlight relaxed, symmetrical facial features.
This interpretability mechanism enables both developers and clinicians to understand whether the model's decisions align with observable human features, thus mitigating the "black-box" nature of machine learning.

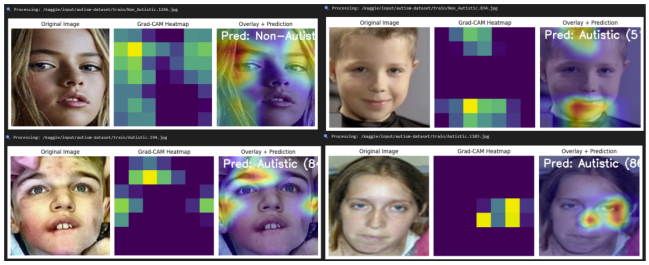

Figure 7.2: Grad-CAM Visualizations

For the speech modality, explainability was enabled by transcribing children's spoken responses into text and analyzing them through **Gemini 2.0 Flash**.
As shown in Figure 7.2, the model identifies key features such as:

- **Vocabulary and Clarity**: Use of appropriate, clear language indicates typical development.
- **Coherence and Structure**: Logical and relevant responses demonstrate conversational understanding.
- **Speech Traits**: Absence of echolalia, atypical prosody, or repetitive language lowered the likelihood of a virtual autism diagnosis.

The model also provides **explicit reasoning** by stating which common indicators (e.g., echolalia, repetitive speech) are absent or present in the audio. This approach not only offers diagnostic clues but also ensures that predictions are anchored in concrete linguistic behaviors, supporting clinicians and parents in understanding the rationale behind AI inferences.

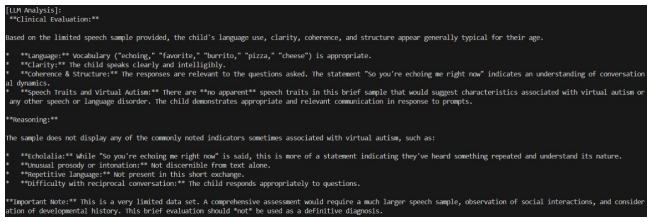

Figure 7.2: Gemini 2.0 Flash Clinical Reasoning.

6.4 User Testing Results

We tested the application with a small subset of 6 children to assess its effectiveness and overall usability. While we cannot share specific outcomes due to privacy concerns, the feedback from parents was positive. They appreciated how the application kept their children engaged through interactive videos and speech based activities, making the experience enjoyable and non-intimidating. Additionally, the clinical reports generated were clear, concise, and easy to understand. Parents found the insights helpful in deciding whether to seek external support, as the reports provided proper explanations for each observation in a simple, parent-friendly manner. The transcripts of the interviews with the parents of the toddlers have been added on the drive link submitted.

7 Discussion

This project marks a meaningful step toward making early behavioral screening for Virtual Autism more accessible and user-friendly. By combining facial expression analysis, speech evaluation, and explainability tools, we've built a system that not only performs well but is also designed with empathy and real-world use in mind. Our goal has always been to support, not replace, clinical evaluation and we've kept that at the heart of every decision. We conducted multiple rounds of testing and iteration, which helped us identify and address real concerns from potential users—such as the need for privacy, simpler navigation, and explainable results. Whether it was through adding a password-protected parent mode or using clear, intuitive visual feedback with Grad-CAM, every feature was carefully designed with the end user in mind.

By combining facial expression analysis, speech evaluation, and explainability tools like heatmaps and LLM-generated justifications, we've built a system that's not just functional but also trustworthy.

8 Future Work

In future iterations, we aim to expand the multimodal pipeline by incorporating **eye gaze tracking** as an additional modality. Eye gaze patterns are a key behavioral marker often used in clinical assessments of autism and could provide deeper insights when fused with existing facial and speech features. This could be achieved through lightweight, privacy-preserving computer vision techniques using front-facing cameras, adding another layer of behavioral nuance. Additionally, further research will explore real-time feedback, personalization based on cultural and linguistic diversity, and deeper integration with health services for smooth handoffs to professionals, we aim to add the functionality to locate nearby healthcare services that specialize in Virtual Autism and ASD care.

References

F. Thabtah and S. Shahamiri, "Autism AI: a New Autism Screening System Based on Artificial Intelligence | Cognitive Computation." Accessed: Feb. 02, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s12559-020-09743-3.

W. Almusawi, "(PDF) 'Virtual autism' or Autism? How can we prevent misdiagnosis?" Accessed: Feb. 02, 2025. [Online]. Available: https://www.researchgate.net/publication/384569527_Virtual_autism_or_Autism_How_can_we_prevent_misdiagnosis.

K. Mamun, S. Bardhan, Md. Ullah, E. Anagnostou, J. Brian, and S. Akhter, "Smart autism — A mobile, interactive and integrated framework for screening and confirmation of autism | IEEE Conference Publication | IEEE Xplore." Accessed: Feb. 02, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7592093.

"Virtual Autism Evaluation." Accessed: Feb. 02, 2025. [Online]. Available: https://www.yellowbusaba.com/post/virtual-autism-evaluation.

M. C. Yadav, B. Venkatachalam, Bhavani Venkatachalam, A. Parmar, "'Virtual Autism': Effects of Excessive Screen Exposure on Communication in Young Children - A Preliminary Study," *Austin Journal of Autism Related Disabilities*, vol. 10, no. 1, Jun. 2024, doi: 10.26420/austinjautismrelatdisabil.2024.1070.

X. Liao, Y. Zhang, J. Sun, H. He, and S. Yu, "Application of Machine Learning Techniques to Detect the Children with Autism Spectrum Disorder," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/6907465.

X. Cao, W. Ye, E. Sizikova, X. Bai, M. Coffee, and H. Zeng, "Vitasd: Robust Vision Transformer Baselines for Autism Spectrum Disorder Facial Diagnosis | IEEE Conference Publication | IEEE Xplore," Accessed: Feb. 20, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10094684

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Jan. 28, 2018, arXiv: arXiv:1608.06993. doi: 10.48550/arXiv.1608.06993. [Online]. Available: https://arxiv.org/abs/1608.06993

C. Senol, "Autism_Image_Data," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/cihan063/autism-image-data

I.-M. Eigsti, L. Bennetto, and M. Dadlani, "Beyond pragmatics: Morphosyntactic development in autism," *Journal of Autism and Developmental Disorders*, vol. 37, pp. 1007–1023, 2007. doi: 10.1007/s10803-006-0239-2. https://asd.talkbank.org/access/English/Eigsti.html