Volume 3, Issue X, October 2024, No. 47, pp. 613-627
Submitted 10/10/2024; Final peer review 20/10/2024
Online Publication 22/10/2024
Available Online at http://www.ijortacs.com

DEVELOPMENT OF A DEEP LEARNING TECHNOLOGY TO IMPROVE HUMAN-COMPUTER INTERACTION FOR VISUALLY IMPAIRED

[1*]Iluno Amalachukwu C., [2]Ogochukwu C. Okeke, [3]Ike Mgbeafuluike
[1,2,3]Department of Computer Science, Faculty of Physical Sciences,
Chukwuemeka Odumegwu Ojukwu University, Anambra State
Email: probity2005@yahoomail.com
Tel: +234 806 915 5380

Abstract

Accessing equitable learning settings can be extremely difficult for people with disabilities, such as vision loss, despite the fact that receiving a high-quality education is a fundamental right. The main goal of this study is to improve accessibility in educational settings by developing an intuitive system that combines deep learning technology to improve human-computer interaction for visually impaired users. The study's primary goal was to create a biometric user identification system that uses facial recognition to give students safe, customised access. A robust object identification system that can reliably identify people in real-time was demonstrated using You Only Look Once (YOLOv5), enabling quick login procedures based on face biometrics. A better human-computer interaction platform that caters to the unique requirements of visually impaired users was also modelled in addition to the authentication system. By combining speech-to-text and voice commands, this technology enables students to freely explore and interact with a computer interface while taking tests. An adaptive ambient noise cancellation technique that uses the Least Mean Squares (LMS) filter to cut down on background noise was developed as a way to improve this system even further. For real-time human-computer interaction, the LMS filter continually adjusts to shifting noise levels, enhancing voice recognition performance and intelligibility in loud settings. Lastly, the study combines these models into a single human-computer interactive system that combines adaptive noise cancellation, voice recognition, and biometric identification. It then deploys the system for analysis. The system's efficacy in offering a user-friendly, accessible solution for visually impaired people in educational settings was confirmed by evaluating its performance using real-world test scenarios. The study was recommended for deployment at school and colleges where students with impaired vision are admitted to help improve their quality of education.

Keywords: Human Computer Interaction; Deep Learning; YOLOv5; Least Mean Square; Biometric User Identification; Voice Recognition

1. INTRODUCTION

Examinations present significant challenges for individuals with impaired vision and those who are completely blind, requiring careful consideration of accessibility, accommodations, and inclusive assessment practices (Ganesan et al., 2022). For individuals with impaired vision, the

challenges primarily revolve around accessing examination materials, navigating the testing environment, and effectively demonstrating their knowledge and skills. Visual impairments can vary in degree, ranging from partial sight to low vision, posing unique needs that must be addressed to ensure fair and equitable assessment opportunities (Tubo et al., 2020).

According to Oliviera (2021), one of the primary challenges for individuals with impaired vision during examinations is the accessibility of test materials. Printed texts, diagrams, charts, and other visual content can be inaccessible without appropriate accommodations such as enlarged print, high contrast materials, or electronic formats compatible with screen readers and magnification software. Furthermore, Tubo et al. (2020) posited that the layout and formatting of examination papers can significantly impact readability and comprehension for individuals with visual impairments, necessitating clear organization, concise language, and intuitive navigation features.

Human-computer interactive systems represent a dynamic interface between users and technology, facilitating seamless communication and interaction across various digital platforms (Nayak et al., 2021). At their core, these systems aim to optimize user experience by integrating intuitive interfaces, responsive feedback mechanisms, and personalized interactions tailored to individual preferences and needs. Through the convergence of human cognition and computational capabilities, interactive systems empower users to navigate complex tasks, access information, and accomplish goals with efficiency and ease (Wang et al., 2020). From user-friendly software applications and responsive websites to interactive kiosks and smart devices, human-computer interactive systems span diverse domains and contexts, reshaping how individuals engage with technology in both professional and personal spheres.

One key aspect of human-computer interactive systems lies in their interface design, which plays a crucial role in shaping user interactions and perceptions. Effective interface design involves understanding user behaviours, cognitive processes, and ergonomic principles to create intuitive layouts, clear navigation paths, and visually engaging elements that enhance usability and accessibility (Fu and Lv, 2020). By employing user-centred design methodologies and conducting iterative usability testing, designers can iteratively refine interfaces to align with user expectations, preferences, and task requirements (Lai et al., 2020). Moreover, Wang et al. (2020) suggested that human-computer interactive systems leverage a diverse range of input modalities, including touch-screens, voice recognition, gestures, and motion sensors, to accommodate varied user preferences and abilities.

According to Lv et al. (2022), deep learning represents a subset of machine learning algorithms inspired by the structure and function of the human brain's neural networks. At its core, deep learning involves training artificial neural networks with large datasets to recognize patterns, make predictions, and perform tasks without explicit programming instructions (Lai et al., 2020). One of the defining characteristics of deep learning is its hierarchical architecture composed of multiple layers of interconnected nodes, or neurons, which process and transform input data through successive layers of abstraction. This hierarchical representation enables deep neural networks to learn complex features and representations directly from raw data, allowing for more

sophisticated and nuanced decision-making capabilities compared to traditional machine learning approaches.

Deep learning has revolutionized Natural Language Processing (NLP) applications, particularly in the realms of Text-To-Speech (TTS) and Speech-To-Text (STT) systems (Nguyen et al., 2021). These applications leverage deep learning models to understand and generate human language with remarkable accuracy and naturalness, significantly enhancing user experiences across various domains (Koizumi et al., 2019). Text-To-Speech (TTS) systems utilize deep learning techniques, particularly Recurrent Neural Networks (RNNs) and convolutional neural networks (CNNs), to convert written text into spoken language. Deep learning models learn to map text input into corresponding acoustic features, such as phonemes, intonation, and prosody, capturing the nuances of natural speech patterns (Nogales et al., 2023). WaveNet, developed by DeepMind, and Tacotron architectures are notable examples of deep learning-based TTS systems known for their ability to generate high-fidelity and expressive speech synthesis (Zhou et al., 2019). These systems have found widespread applications in virtual assistants, audiobook narration, and accessibility tools for visually impaired individuals, and personalized customer service interactions, enhancing the accessibility and usability of digital content. On the other hand, speech-to-text (STT) systems employ deep learning algorithms, particularly recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer architectures, to transcribe spoken language into textual representations (Nogales et al., 2023). Deep learning models trained on vast corpora of speech data learn to extract meaningful features from audio signals, discerning phonetic patterns, language syntax, and semantic context to accurately transcribe spoken utterances. State-of-the-art STT systems, such as Google's Deep Speech and Baidu's Deep Speech 2, have achieved remarkable performance gains in terms of transcription accuracy and real-time processing capabilities (Guan et al., 2023). These systems are widely deployed in voice-activated virtual assistants, dictation software, call center automation, and transcription services, streamlining communication processes and enabling seamless interaction between humans and machines (Nguyen et al., 2021).The aim of this study is to design an intuitive system using deep learning to enhance human computer interaction for visually impaired users.

2. METHODOLOGY AND DESIGN METHODS

The research methodology used for the study is the object oriented analysis and design methodology. The research methods used for the system development are biometric user authentication system, face detection and tracking, real-time facial recognition, deep learning-based voice manager, adaptive filtering technique, data management, result computation module, and deep learning based human computer interactive system. These methods are discussed as follow;

2.1 Biometric user authentication system

This method utilized the face of the student to grant access to the examination dashboard. The system compares the captured face data against stored class album data and then use to grant authentication for the examination. The reason for the adoption of this method was due to its

reliability, credibility and ability to provide system integrity, and access control, by allowing only registered and qualified students to per-take in the exam process. The biometric authentication process is performed with facial recognition technique.

2.2 Face detection and tracking

This method utilized computer vision algorithms, specifically one stage object detector to track and detect face of the student once positioned for the examination. The process involves identifying and tracking a face within clustered images or video scene. The method searched for HAAR features such as eyes, nose, mouth and track their movement in real-time. This process pipelines the next step which is the face recognition process.

2.3 Real-time facial recognition

The facial recognition process utilized a trained object detector with faces of all the students in the class album to generate a face recognition model. The method involves data collection, deep learning model, training, and then generation of the model for the facial recognition process.

i. Data collection

The human face data used for this research was collected from Kaggle repository as the primary source of data collection. The sample size of the dataset images is 180 face data. The secondary source of data collection is a self-volunteered dataset created to test the model and validate the results. The dataset was processed using data Roboflow data augmentation tools and then applied for the training of the deep learning model.

ii. Deep learning model

The deep learning model used or the work is the You Can Only Look Once (YOLOV) algorithm. Specifically, YOLOV-5 is a one stage object detection model which has capability for real-time image classification. The YOLOV-5 is made of the backbone, neck and head respectively. The backbone contained the Darknet-53 as the pre-trained model for image extraction using series of convolutional neural network and spatial pyramid pooling. The extracted image features are feed to the neck for fusion using concatenation, and attention networks. These features are trained in the head to predict the face of the student using bounding box and confidence score for the classification.

iii. Real-time face recognition model

The real-time face recognition model is the output of the trained YOLOV-5. During the training process, regularization and optimization back-propagation algorithms are applied to generate a regularized model for real-time classification of images. With this model, when the student sits, the cameras collected the images and then match with the trained YOLOV-5 model to recognize the student and then grant authentication.

2.4 Deep Learning-Based Voice Manager

The deep learning-based voice managers are already trained deep learning model for text to speech and speech to text applications. This deep learning-based model utilized recurrent neural network and convolutional neural network configuration to train a model capable of performing natural language processing applications. This method enables the voice interaction process of the system with users. The user commands the system with voice, which is then converted to text

for a particular function such as reading the exam procedures, questions, answers, results. These functions are feedback to user through audio format.

2.5 Adaptive filtering technique

This method was introduced to address the issues of adaptive filter due to environmental noise. Since the system will be utilized in an exam hall with many other students, and all interacting with their system through voice, the impact of environmental noise generated from the environment has to managed, hence presenting the application of adaptive filter using least mean square algorithm. This filter utilized gradient descent-based algorithm to update filter coefficients based on instances of error from the input signal.

2.6 Data management

Data management involves the storage, retrieval, organization, and manipulation of data within the system database using MySQL. This includes activities such as data indexing, query optimization, question scheduling, answer scheduling, score sheet and recovery. Effective data management is crucial for ensuring data integrity, availability, and security of the examination reports. The database development tool used is Robowflow. The Figure 1 presents the database development environment. This was achieved loading the data into the Robowflow environment and then using the annotation tool to label the images according to student name and registration number.

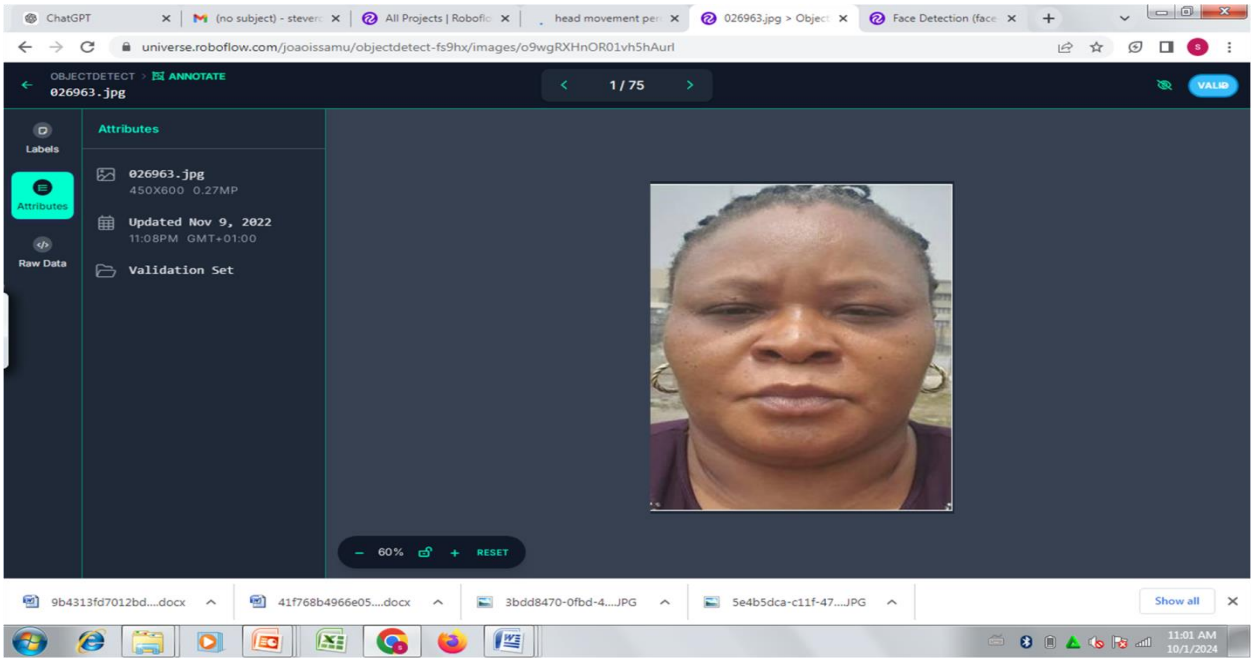

Figure 1a: The data development tool interface

2.7 Database Design and Structure

Table 1: Data description

Data Attribute	Description	Type	Example Values
Image_ID	Unique identifier for each face image	Categorical	IMG001, IMG002, IMG003
Student_ID	Unique identifier for each	Categorical	P001, P002, P003

	Student		
Image_Resolution	Dimensions of the image (width x height)	Numerical	128x128, 256x256, 512x512
Age	Age of the Student in the image	Numerical	18, 25, 40
Gender	Gender of the Student in the image	Categorical	Male, Female
Pose	Head pose or angle at which the face was captured	Categorical	Front, Left, Right, Upward, Downward
Facial_Expression	Facial expression captured in the image	Categorical	Neutral, Happy, Sad, Angry, Surprised
Illumination	Lighting conditions during image capture	Categorical	Low, Normal, High
Occlusion	Whether the face is partially covered (glasses, mask)	Categorical	Yes, No
Image_Format	Format of the image file	Categorical	JPEG, PNG, BMP
Face_Bounding_Box	Coordinates for the region containing the face (x, y, w, h)	Numerical	(50, 30, 100, 120)
Key_Facial_Points	Coordinates of key facial landmarks (eyes, nose, mouth)	Numerical Array	[(34, 65), (40, 70), (55, 80), …]
Embedding_Vector	Feature vector representing the face (after extraction)	Numerical Array	[0.25, 0.75, -0.34, 0.10, …]
Time_of_Capture	Timestamp when the image was taken	DateTime	2024-10-01 12:45:30
Camera_Angle	Angle of the camera relative to the face	Categorical	0°, 45°, 90°
Distance_to_Camera	Distance between the subject and the camera (in meters)	Numerical	1.0, 1.5, 2.0

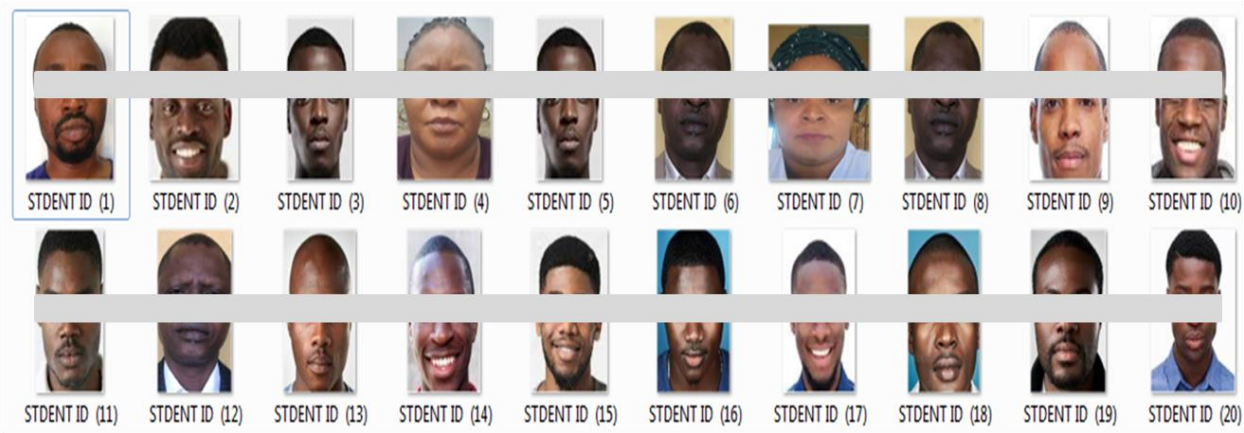

Figure 1b: Database design

3. DEEP LEARNING BASED HUMAN COMPUTER INTERACTIVE SYSTEM

This method leverages YOLOV-5 model to enable user authentication through facial recognition in real-time, then deep learning-based voice manager was applied to process user voice input

while carrying out the exam process. The system facilitates examination for the visually impaired and the results, communicated to the user as audio.

3.1 The system Block diagram

The system block diagram showcased the interconnectivity and relationships between each module of the deep learning based human computer interactive system as shown in Figure 2.

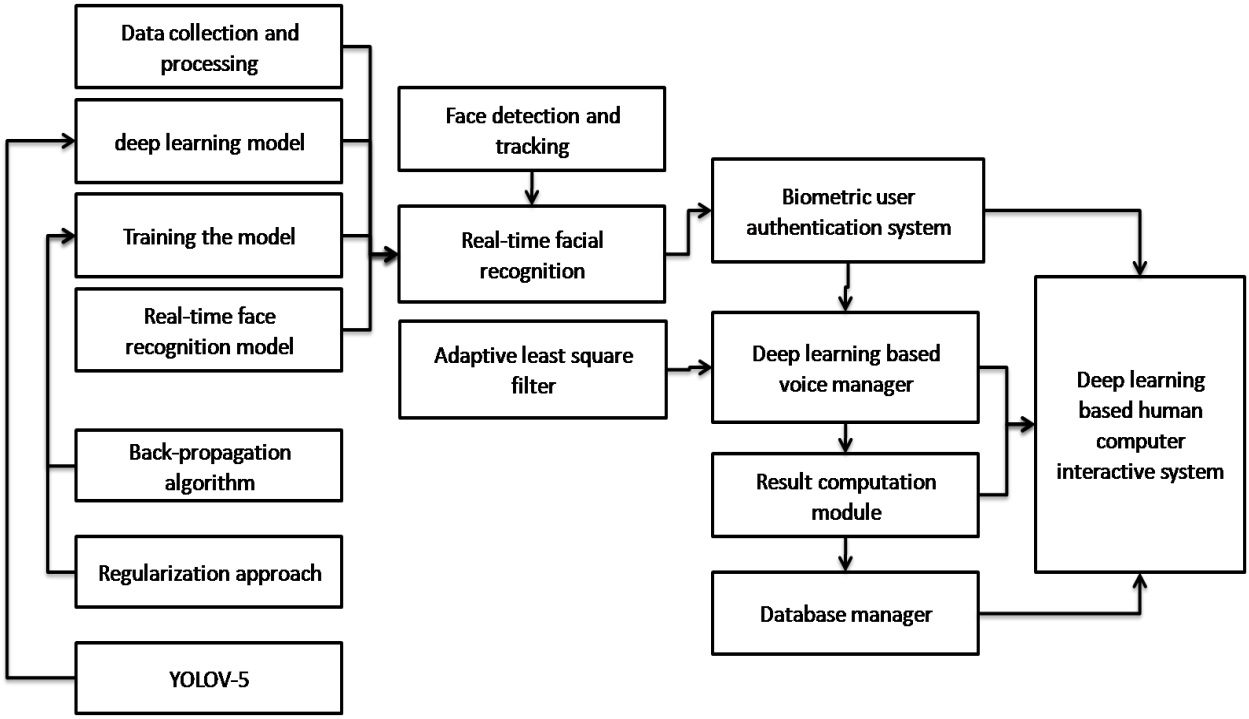

Figure 2: Block diagram of the deep learning based human computer interactive system

The Figure 2 reported the system block diagram which started with the biometric authentication developed with real time facial recognition system. The biometric system was made possible using methods such as data collection, processing through augmentation and then training of deep learning model (YOLOV-5), using back-propagation and regularization approach to generate the real-time face recognition model. Deep learning-based voice manager is another module in the system whose responsibility is to convert input voice to text and also output text to speech, to facilitate human computer interaction. In addition, the adaptive filter was integrated to the voice manager to address the issues of environmental noise using least square filter technique. Results computation module showcased the section responsible to determine the score of the examination process carried out by the user, then the database manager is the server where every information such as exam questions, user information, answer and results are stored. The high-level model of the system was reported in the Figure 3. The Figure 3 showcased the high-level illustration of the deep learning based human computer interactive system. The figure presented an interactive model of the use case, showing the user captured by the camera and then feed to the biometric authentication system. The laptop installed with the deep learning based human computer interactive system has a microphone which collected audio information from users and then process and give feedback through microphone

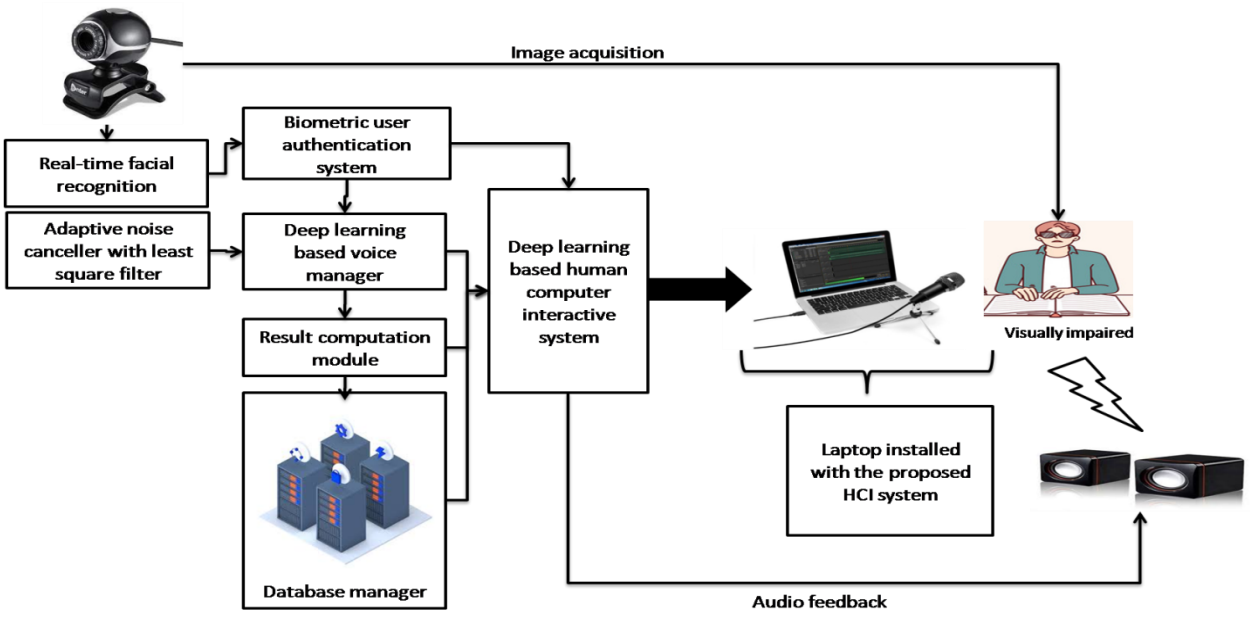

Figure 3: High level illustration of the proposed system

3.2 YOLOv5-based Facial Recognition for Student Authentication

A real-time YOLOv5-based facial recognition system integrates the powerful object detection capabilities of YOLOv5 with face recognition algorithms for swift and accurate identification. YOLOv5 is employed to detect and localize faces within video streams or images, leveraging its speed and efficiency to process frames in real time. After face detection, the system passes the cropped face images through a recognition module, often using pre-trained deep learning models that convert facial features into embedding vectors. These vectors are then compared with a database of known student embeddings to identify or verify individuals. The real-time nature of this system allows for immediate student identification in scenarios like attendance monitoring or access control. With YOLOv5's lightweight architecture and high detection accuracy, the system ensures fast processing, making it highly suitable for environments requiring quick identification with minimal latency, such as classrooms or examination halls.

3.3 Face recognition-based biometric login authentication

The face recognition-based biometric login authentication system using YOLOv5 combines object detection with facial recognition to enhance security and speed in authentication processes. YOLOv5 is utilized to detect faces in real time, leveraging its ability to quickly identify facial regions within video streams or images. Once a face is detected, the system extracts the relevant facial features, creating an embedding (a numerical representation of the face's unique characteristics). This embedding is then compared with a stored database of enrolled users' face embeddings to either authenticate or reject the login attempt. The process begins with YOLOv5 detecting and localizing the face, making it highly efficient for real-time applications like login systems. The system then processes the detected face through a facial recognition algorithm to match it against the user database. The advantage of YOLOv5 lies in its speed and precision, as it can process multiple faces in a single frame, ensuring quick and accurate detection even in challenging conditions like varying lighting, occlusions, or different

angles. This makes the YOLOv5-based biometric login system highly secure, reducing the risk of unauthorized access, while offering a seamless user experience with quick authentication, especially useful in environments such as secure logins for student exams, online platforms, or enterprise systems.

3.4 Google Speech-to-Text (Deep learning voice Manager)

For the blind student examination process, a speech-to-text system plays a crucial role in ensuring accessibility. The system allows blind students to answer exam questions by speaking their responses, which are then converted to text in real time. This is achieved by integrating a speech recognition engine, such as Google Speech-to-Text or a custom-trained model, into the exam process. The accuracy of the system in capturing the spoken words, punctuation, and formatting is essential for ensuring that the answers are recorded exactly as the student intended. Pre-processing techniques like noise reduction and voice normalization help to improve recognition accuracy, while the system must also support multiple languages or dialects based on the student's native language. The transcribed text is automatically saved, ensuring a seamless and efficient process for students with visual impairments.

3.5 Voice-Based Question-And-Answer System

A voice-based question-and-answer system enables users to interact with a system using spoken language rather than text input. In this setup, a user can ask questions verbally, and the system, leveraging speech recognition technologies, converts the spoken words into text. This text is then processed by a natural language understanding (NLU) model to interpret the query. The system retrieves the appropriate answer from a knowledge base or generates a response using pre-trained language models, which is then converted back into speech using text-to-speech (TTS) technology. This type of system is particularly useful for accessibility, allowing users with visual impairments or limited typing ability to engage seamlessly with digital platforms. Voice-based Q&A systems are increasingly applied in customer service, virtual assistants, educational tools, and exam processes, providing an intuitive and hands-free experience.

3.6 Adaptive filter

The Least Mean Squares (LMS) filter works by adaptively adjusting its filter coefficients to minimize the difference between a noisy input signal and a desired reference signal, effectively cancelling out the noise. It begins with small, random filter weights and iteratively updates these weights based on the error between the filter's output and the desired signal. The core of its functionality lies in the LMS update rule, where each coefficient is adjusted by a small fraction (controlled by a step size) of the product of the error and the input signal. As the input signal is processed, the error decreases, allowing the filter to learn and converge toward the optimal coefficients. This adaptive mechanism helps improve noise cancellation in real-time, making it ideal for speech enhancement in systems like voice-based human-computer interaction.

3.7 System Development Algorithms

A. Stepwise of Training YOLOv5 Algorithm

1. Start
2. Data Preparation: Collect and Annotate Data of student faces
3. Environment Setup: Clone YOLOv5 Repository

4. *Model Configuration: Configure YOLOV-5 hyper-parameters*
5. *Training the Model: Execute the training script of YOLOV-5 with the prepared dataset.*
6. *Monitor Training: Keep track of loss, learning rates, and validation results*
7. *Test the Model: Evaluate performance in terms of precision, and recall*
8. *Deployment: Export the Model to build the software for biometric student identification*
9. *End*

B. Stepwise of Deep Learning-Based Biometric Authentication System with YOLOv5

1. *Start*
2. *Data Collection: Gather a dataset of images for face recognition*
3. *Face Detection with YOLOv5:*
4. *Utilize the YOLOv5 model to detect faces in the input images.*
5. *Configure the model to output bounding boxes around detected faces.*
6. *Face Recognition:*
7. *Extract features from the detected faces using a face recognition algorithm*
8. *Store these features in a database along with the corresponding user identities.*
9. *User Authentication: During login, capture a live image of the user's face.*
10. *Use YOLOv5 to detect the face and extract features.*
11. *Compare the extracted features against the stored features in the database to verify the user's identity.*
12. *Output Result:*
13. *If the user is authenticated, grant access to the system*
14. *If authentication fails, prompt the user to try again or use an alternative method*
15. *End*

C. Stepwise of Face Recognition System

1. *Start*
2. *Data Acquisition: Collect images for training the face recognition model.*
3. *Face Detection: Use a face detection algorithm YOLOv5*
4. *Feature Extraction: Use YOLOV-5 bottleneck to generate face feature vector*
5. *Classifier: Trained YOLOV-5 model*
6. *Real-Time Recognition: Detect and recognize faces of student in the video stream*
7. *Output Handling: Display the recognized identity*
8. *Login: initialize login access for exam*
9. *End*

D. Least Mean Squares (LMS) filter algorithm for adaptive noise cancellation:

Step 1: Initialization:

Set the filter coefficients to small random values
Define the step size which determines the learning rate and convergence speed.
Select the filter length (number of coefficients).

Step 2: Input Signal:

Receive the input signal (noisy signal) at each time
Provide the desired signal (clean reference signal or expected output).

Step 3: Filter Output:

Calculate the filter output

Step 4: Error Calculation:

Coefficient Update:
Update the filter coefficients using the LMS update rule:

Step 5: Iteration:

For each new incoming sample of the input signal, repeat steps 3, 4, and 5.

The filter continuously adapts its coefficients to minimize the error.

Step 6: Convergence:

Over time, the filter coefficients converge, reducing the noise in the signal, leading to improved output quality (e.g., cleaner audio for speech-to-text systems).

4. SYSTEM IMPLEMENTATION

The system implementation for the biometric authentication and face recognition system involves a comprehensive setup that integrates various components to function seamlessly. Initially, the model development is executed using Google Colab, where the YOLOv5 algorithm is employed for face detection and recognition. This platform provides access to powerful GPU resources, significantly speeding up the training and testing phases of the model. The implementation includes installing the necessary libraries and frameworks, such as PyTorch and OpenCV, and utilizing the Ultralytics YOLOv5 repository to leverage pre-trained models and fine-tune them on the specific dataset of student faces. The dataset, consisting of images and corresponding labels, is meticulously organized to ensure optimal performance of the face recognition system. Following the model training, the implementation process extends to the development of the examination interface, which is designed using Python. This interface facilitates the submission of exam questions and answers, ensuring that students can engage with the system effectively. The implementation also incorporates speech-to-text and text-to-speech functionalities, enabling students to participate in exams using their voices, which is particularly beneficial for visually impaired individuals. Furthermore, the integration of a database management system allows for secure storage and retrieval of student profiles, authentication logs, and exam results. Overall, the successful implementation of the system not only enhances the educational experience for students but also establishes a reliable and efficient biometric authentication process.

5. RESULTS OF SYSTEM INTEGRATION OF THE HUMAN COMPUTER INTERACTION SYSTEM

This section presents the results of the integrated system for student login and examination process. The effectiveness of the model was demonstrated considering several student faces and the performance during examination as shown in Figure 5.

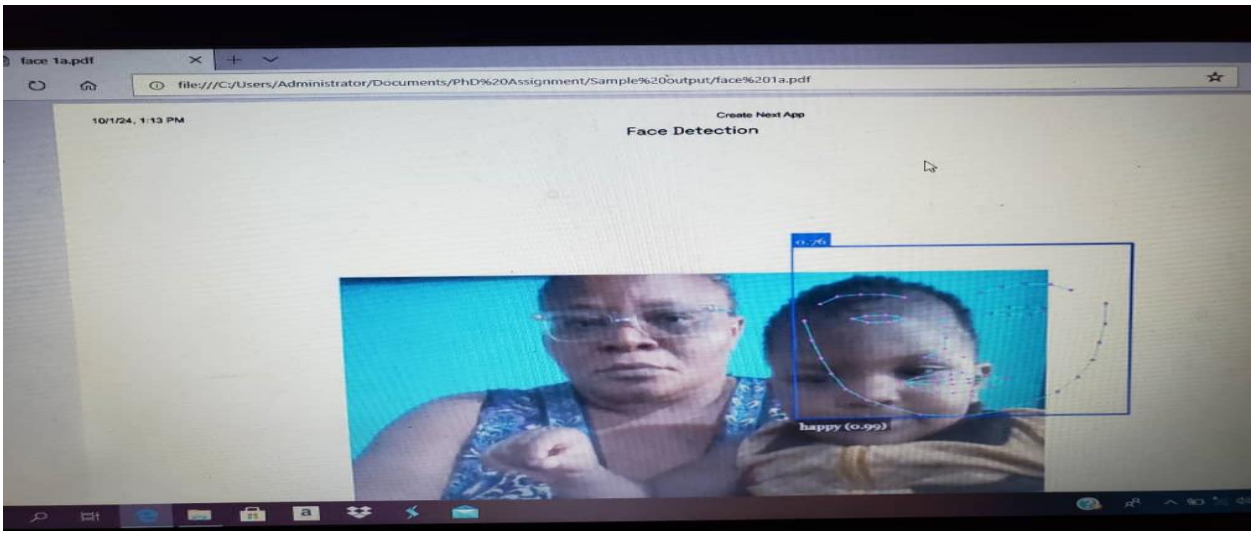

Figure 5: Result of face recognition

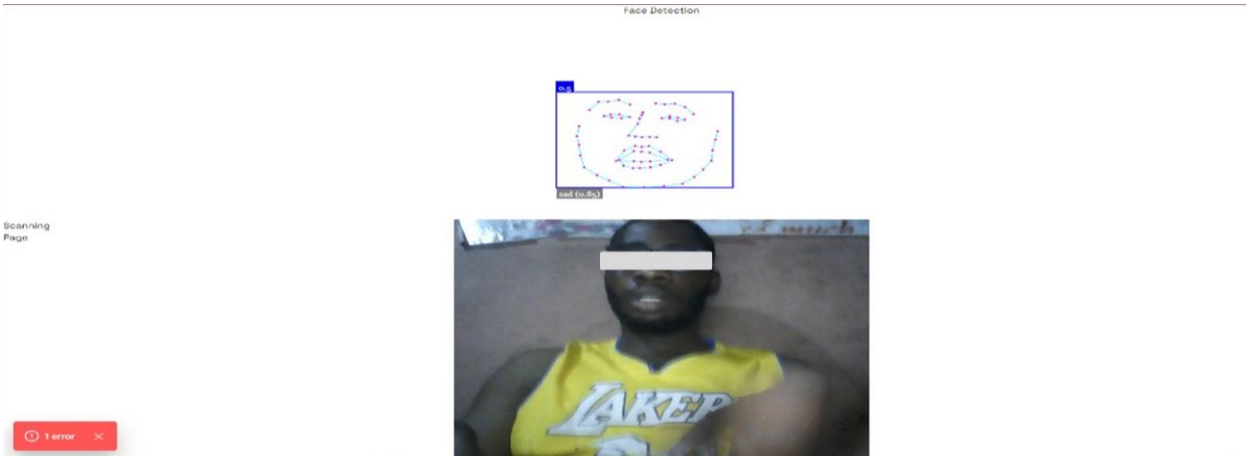

Figure 6: Result of Biometric student authentication

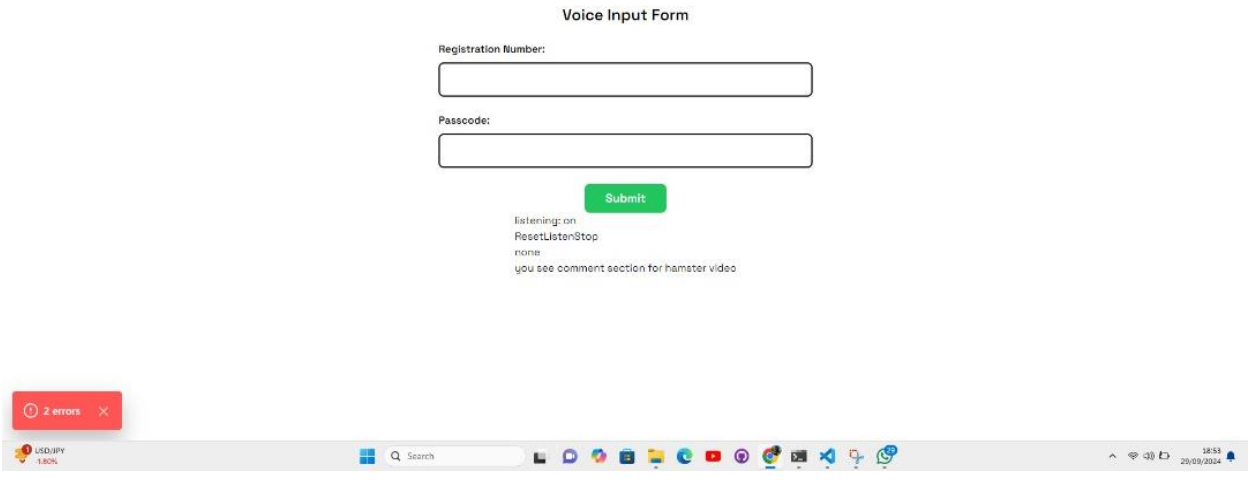

Figure 7: Voice input for registration number verification

Figure 8: Voice input interface for exam and score dashboard

Figure 9: Voice input for exam and score board

While the proposed biometric authentication and face recognition system offers significant benefits, several limitations must be acknowledged. Firstly, the system's performance is highly dependent on the quality of the training data; inadequate or biased datasets lead to inaccurate face recognition and authentication results. Additionally, the system struggle in poorly lit environments or when faces are partially obscured, affecting the reliability of face detection and recognition. Another limitation is the potential for false positives or negatives, where unauthorized users are granted access, or legitimate users are denied, which could undermine trust in the system. Furthermore, the need for a stable internet connection for cloud-based features, such as Google Colab, could pose accessibility issues in areas with limited connectivity. Lastly, concerns around privacy and data security must be addressed, as the storage of sensitive biometric data presents risks if not managed properly.

6. CONCLUSION

This study addresses the challenges faced by visually impaired individuals in accessing quality education by developing an advanced, deep learning-based human-computer interaction system. The primary goal was to enhance accessibility through a biometric user authentication system

that uses YOLOv5, a powerful object detection algorithm, for facial recognition. By incorporating this algorithm, the system ensures secure and personalized access for students based on facial biometrics, making login processes both accurate and efficient. The design prioritizes ease of use, ensuring visually impaired individuals can access digital platforms seamlessly without requiring manual assistance. In addition to authentication, the system integrates speech-to-text and text-to-speech functionalities to enable intuitive human-computer interaction. Voice commands and adaptive responses allow visually impaired users to navigate through tasks and applications. To further improve the system's performance, an adaptive noise cancellation algorithm using the Least Mean Squares (LMS) filter was introduced. This filter dynamically adjusts to fluctuating environmental noise levels, ensuring clarity during speech recognition tasks. As a result, the system is able to process voice commands more effectively in noisy environments, significantly enhancing user experience. Lastly, all components, including the biometric login, voice recognition, and noise cancellation, were integrated into a single, comprehensive human-computer interactive system. The system was rigorously tested to evaluate its performance, and the results demonstrate that it is highly effective in supporting visually impaired users. This innovative system holds great promise for improving accessibility in educational environments, offering a practical solution to challenges faced by those with visual impairments. By enabling more independent navigation and interaction with technology, this research contributes to the development of inclusive learning environments.

REFERENCES

Fu Q., &Lv J., (2020) Research on Application of Cognitive-Driven Human-Computer Interaction. Am. Sci. Res. J. Eng. Technol. Sci. **2020**, 64, 9–27.

Ganesan J., Azar A.T., Alsenan S., Kamal N.A., Qureshi B., &Hassanien A.E., (2022) Deep Learning Reader for Visually Impaired. Electronics **2022**, 11, 3335. https://doi.org/10.3390/electronics11203335

Guan J., Liu Y., Kong Q., Xiao F., Zhu Q., Tian J., & Wang W., (2023) Transformer-Based Autoencoder with ID Constraint for Unsupervised Anomalous Sound Detection. EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:42 https://doi.org/10.1186/s13636-023-00308-4

Koizumi Y., Saito S., Uematsu H., Kawachi Y., & Harada N., (2019) Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma. IEEE/ACM Transactions on Audio, Speech, And Language Processing, Vol. 27, No. 1, January 2019.

Lai H., Chen H., & Wu S., (2020) Different Contextual Window Sizes Based RNNs for Multimodal Emotion Detection in Interactive Conversations. IEEE Access **2020**, 8, 119516–119526.

Lv Z., Poiesi F., Dong Q., Lloret J., & Song H., (2022) Deep Learning for Intelligent Human–Computer Interaction. Appl. Sci. **2022**, 12, 11457. https://doi.org/10.3390/ app122211457

Nayak S., Nagesh B., & Routray A., (2021) A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences. Comput. Electr. Eng. **2021**, 93, 107280.

Nguyen M., Nguyen D., Pham C., Bui D., & Han H., (2021) Deep Convolutional Variational Autoencoder for Anomalous Sound Detection. 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE) | 978-1-7281-5471-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE48956.2021.9352085

Nogales A., Donaher S., & García-Tejedor A., (2023) A deep learning framework for audio restoration using Convolutional/ Deconvolutional Deep Autoencoders. Expert Systems With Applications 230 (2023) 120586https://doi.org/10.1016/j.eswa.2023.120586

Oliviera J., (2021) Using Interactive Agents To Provide Daily Living Assistance For Visually Impaired People. Pontifical Catholic University Of Rio Grande Do Sul School Of Technology.

Tubo F., Ikechukwu E., & Doris C., (2020) Development of An NLP-Driven Computer-Based Test Guide For Visually Impaired Students; https://arxiv.org/ftp/arxiv/papers/2401/2401.12375.pdf

Wang Z., Jiao R., & Jiang H., (2020) Emotion Recognition Using WT-SVM in Human-Computer Interaction. J. New Media **2020**, 2, 121–130.

Zhou Y., Bao C., & Cheng R., (2019) GSC Based Speech Enhancement with Generative Adversarial Network. Proceedings of APSIPA Annual Summit and Conference 2019APSIPA ASC