International Journal of Artificial Intelligence Trends (IJAIT)
Vol. 3, Issue 10; No. 51, October 2024, pp. 519-528

DEVELOPMENT OF A DEEP LEARNING BASED HUMAN COMPUTER INTERACTIVE SYSTEM

[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

Article Info
Received: 10/10/ 2024
Revised: 21/10/2024
Accepted 22/10/2024

Corresponding Author's Email: probity2005@yahoomail.com

Corresponding Author's Tel:
+234 806 915 5380

ABSTRACT
Quality education is a fundamental right, but individuals with impairments, such as vision loss, often face significant challenges in accessing equitable learning environments. This study focuses on designing an intuitive system that leverages deep learning technologies to enhance human-computer interaction for visually impaired users, with a primary focus on improving accessibility in educational settings. By utilizing You Only Look Once-Version 5 (YOLOv5), a robust object detection algorithm that can accurately identify individuals in real-time, facilitating fast login processes based on facial biometrics was presented. In addition to the authentication system, an improved human-computer interaction platform that adapts to the specific needs of visually impaired users was modelled. This system integrates voice commands and speech-to-text capabilities, empowering student to navigate and interact independently with computer interface while sitting for examination. To further enhance this system, an adaptive environmental noise cancellation algorithm was proposed, utilizing the Least Mean Squares (LMS) filter to reduce background noise. The system's performance was evaluated through real-world tests scenario, validating its effectiveness in providing a user-friendly, accessible solution for visually impaired individuals in educational settings. The results of the trained YOLOV-5 model reported precision score of 0.91, recall of 0.90 and F1-score of 0.98. Comparative analysis with other state of the art algorithms reported our model as the most reliable due to the integration of the adaptive filter.

Keywords: Human Computer Interaction; YOLOv5; Least Mean Square; Facial Biometrics; Speech-to-Text

1. INTRODUCTION

"Embracing the power of innovation, we envision a future where every student, regardless of their visual abilities, can confidently embark on the journey of learning." While every child has the right to education, people with impaired vision have consistently struggled to exercise this right to acquire knowledge. Impaired vision is a state where one is unable to see completely or where one can see partially. In other cases, they are people whose vision cannot be corrected by contact lenses, surgery, or medications (Lv et al., 2022). Persons with this challenge encounter substantial difficulty when trying to exercise their right to acquire reading, and it has remained a major issue that has continuously attracted research attention.

Over the years, vast research on alternative visual aid systems, augmented communication systems, user-centered designs, and human-computer interactive designs has been presented to help improve the standard of living for the visually impaired (Oliviera, 2021). While these systems apply to diverse user requirements, Human-Computer Interaction (HCI) designs have continued to gain momentum in application areas like education to facilitate academic experiences for the visually impaired (Mosquera et al., 2021). HCI, according to Lv et al. (2022), is a field of study that focuses on the specialized technologies that facilitate interaction between human and computer systems. This HCI system has been applied to also improve the quality of life for physically challenged people and allow them to interact with the computer system through input commands from voice, sign, and gaze.

Over the years, HCI has been enhanced through the application of gesture recognition, lip recognition, speech translation, speech synthesis, and speech recognition and has been widely applied for the optimization of HCI (Wang et al., 2020; Prathiba and Kumari, 2021).In the context of education for impaired vision persons, the application of HCI has the potential to improve critical academic activities such as examinations. While the learning process has been standardized for the visually impaired, the need for a reliable system that facilitates their learning performance evaluation for the blind has remained a major challenge.

According to Fu and Lv (2020), traditional approaches to the academic examination of the blind encompass a variety of strategies, such as Braille system oral examination, scribes, specialized formats, tactile diagrams, and accessible technology. This traditional approach, while commendable, is now poised for transformation through the integration of deep learning technologies (Jarosz et al., 2021; Nayak et al., 2021).Deep learning (DL) is a type of machine learning algorithm, specifically an artificial neural network, that can solve complex sequential, time series, or image classification problems. DL is a neural network-inspired model applicable for the optimization of HCI to improve application and user diversity.

Today, many studies have applied DL for improved HCI. For instance, Abhijna et al. (2022) applied a deep neural network to develop a mobile application system that improved administrative activities for the visually impaired. Marvin (2020) applied deep learning to improve image recognition for the visually impaired. Hande and Bilawar (2022) proposed a deep voice assistance system for social media interaction using a deep learning technique, while Bhukhya et al. (2023) applied deep learning to the modelling of a visual assistance system that facilitates autonomous navigation for the blind, while Oey (2023) developed a natural language understanding-based chatbot interactive design using a using a deep learning technique. While these studies all contributed successfully to the application of deep learning to improved human-computer interactive systems, there is limited application of this technology for improved education performance for the blind. Secondly, the system currently in existence lacks reliable authentication means, which is a critical issue because Shi et al. (2023) argued that a system without credible means of authentication can give room for issues such as malpractice and manipulation of results and affect the overall exam integrity. Finally, the issue of environmental noise interference affects the performance of the existing system, making it not function effectively (Bahraini and Sadigha, 2024; Sudhakar et al., 2023).

Deep learning offers the opportunity to revolutionize the way visually impaired individuals navigate educational assessments (Lai et al., 2020) through secured user access control, interactivity, automation, and smart feedback capabilities with noise cancellation filters (Guan et al., 2023). Therefore, this study proposes a deep learning-based system to enhance human-computer interaction for visually impaired users. The system will utilize a biometric authentication approach through real-time facial recognition to regulate user access control, and then a deep learning-based voice management system with an adaptive filter will facilitate user-system interactivity through an audio feedback mechanism. The proposed system will be tailored towards an automated academic examination exercise, utilizing advanced speech recognition technologies to interpret and present exam content in audio formats, which enables the students to comprehend and respond to questions effectively. In addition, a scoreboard will also be integrated with the system to record the performance of the individual during the examination process and provide immediate scores once the exam is completed.

2. RESEARCH METHODOLOGY

The methodology used for the study is the object-oriented analysis and design methodology. In realizing the methodology, a biometric user authentication system was developed through facial recognition to grant user authentication to the system, then an improved human computer interactive system which utilized deep learning models was developed for test to speech and speech to text applications. To address the impact of environmental noise on the system, an adaptive filter specifically tailored towards Least Mean Square (LMS) filter was proposed. The performance of the new system was evaluated and validated through practical experimentation.

3. THE PROPOSED DEEP LEARNING BASED HUMAN COMPUTER INTERACTIVE SYSTEM

The proposed system utilizes deep learning to enhance human computer interaction for visually impaired users with application tailored towards academic examination process. The major component of the system are the user authentication process which requires biometric technology, specifically facial recognition to access users, then the next components utilized a database management system for questions scheduling, answers, user information and results computation, then the deep leaning based voice manager as reported in the Figure 1;

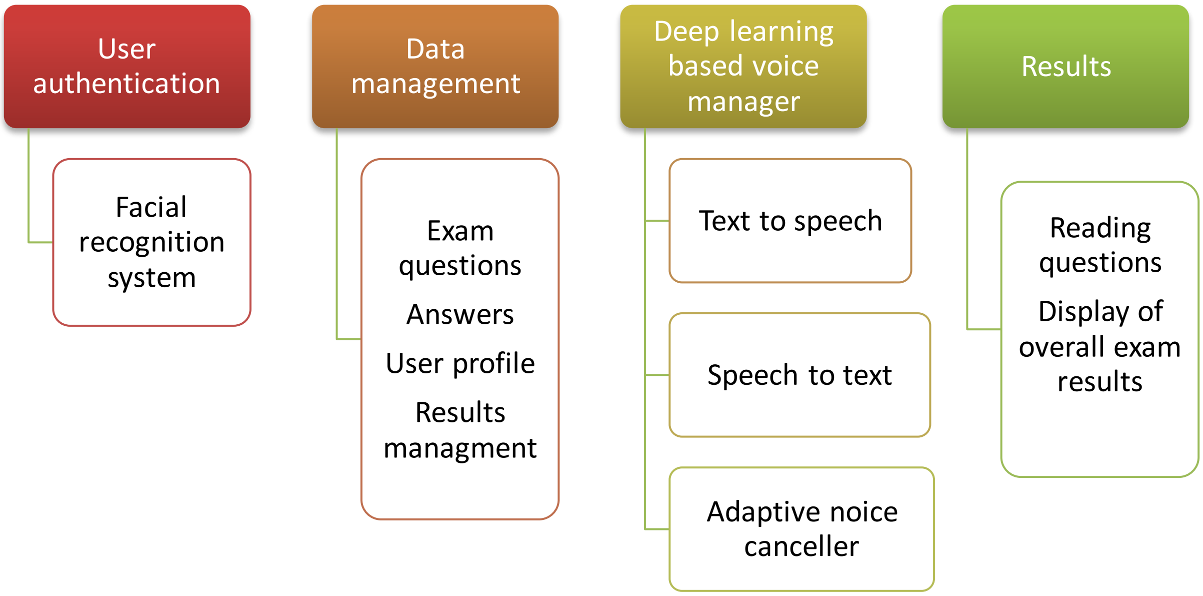

Figure 1: Proposed deep learning based human computer interactive system framework

The Figure 1 showcased the four main section of the deep learning based human computer interactive system. The user authentication system was improved utilizing facial recognition system. In this case, a deep learning-based model will be trained to verify faces of users and then grant automatic secured access to the system. The next section is a database management system where questions, answers, user information are managed and records of exams. Deep learning-based voice manager takes on the responsibility of text to speech and speech to text manager. This system was also equipped with adaptive noise cancer to address the impact of environmental noise on the system. Then finally is the results section which manages the output of the system such as audio and results. The activity diagram of the authentication process with acial recognition is reported in figure 2. The figure showcased the activity diagram analysing the flow chart of the authentication process. The operation utilized the system camera to capture data of user (face to be specific), then feed to YOLOV for facial recognition and verification of users. The verified user information is applied for automatic login to the system. The complete activity modelling of the system was reported in Figure 3. The Figure 3 showcased the operation of the proposed system, with the integrated biometric authentication system. When the user is authenticated through facial verification process, the system feedback to the user through a welcome address and then read out steps to conduct the examination. After which the user interacts with the system through voice command to carry out the examination process, utilizing the database as the question-and-answer source, while automatically scoring and managing the results. The impact of environmental noise on the input signal was mitigated using adaptive noise cancellation techniques, specifically least mean square algorithm.

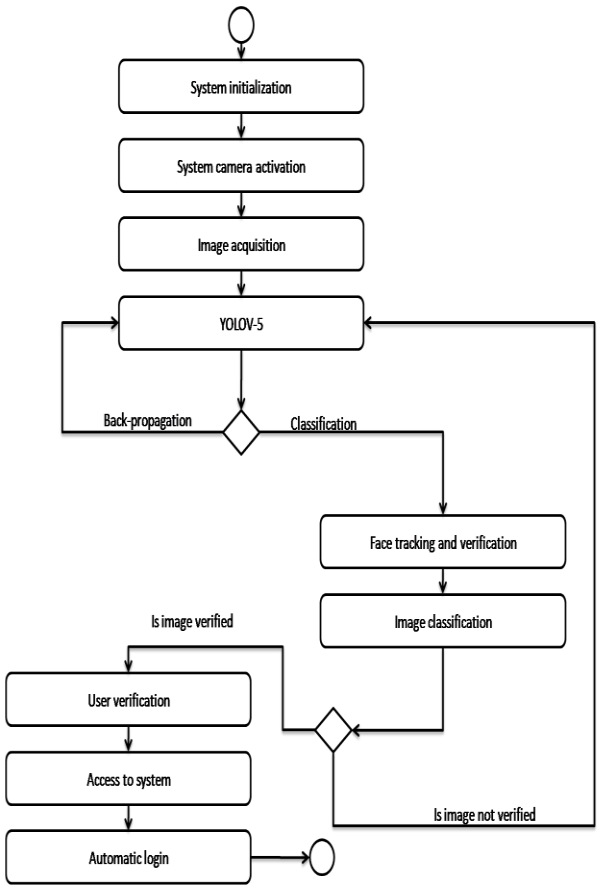

Figure 2: Activity diagram of authentication process

a. Data Collection

The human face data used for this study 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 were processed using data RobowFlow data augmentation tools and then applied for the training of the deep learning model.

4. MODELLINGYOLOV-5 ALGORITHM

The YOLOV-5 model is composed of three main components which are the backbone, neck and head. The input to these three YOLOV-5 components first augments the data and format into 640 x 640, then color channel of 3 which is the standard for RGB images is also used to form the new image size of 6406403. At the backbone, feature extraction is performed using focus, C3 and SPP modules. The necks aggregate these features using PAN and FPN and then feed to the head for prediction. The Focus module is applied to down-sampled the input image

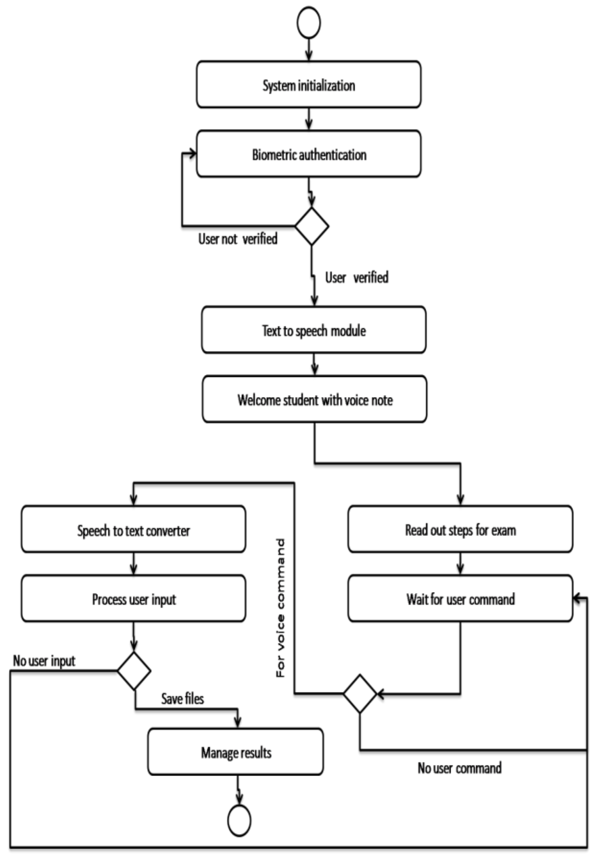

Figure 3: Activity diagram of the deep learning based human computer interactive system

(for instance as the input image is dimension 6406403), the focus down-sample the image into 32032012, using a convolutional filter of 3 2, while the value of 12 represents the color channel. The Figure 4 presented the block diagram of the focus module discussed. SPP module was also applied for the pooling of the of feature maps. The SPP is an additional module which uses spatial pyramid pooling fusion to provide multi-scale feature representation, thus allowing the capturing of multiple feature scales in diverse sizes. The SPP also make use of the spatial pyramid convolutional module to process the extracted features, thus contributing to the overall optimization of object detection performance. The Figure 5 presented the block diagram of the SPP module. The C3 module which is also a critical part of the backbone is composed of three CNN, concat, three bottle-neck and Cross State Partial (CSP). Each of the CNN was applied for the extraction of feature maps. The CSP improves the model representation and facilitates the learning of complex feature maps, while bottleneck reduces the colour channel number and works hand in hand with the CSP as bottleneck-CSP. These features are concatenated, combining the extracted from the CSP and CNN. Collectively the C3 was applied to facilitates learning of complex feature maps and improve the overall performance of the YOLOV-5. The Figure 6 presented the block diagram of the focus module with the interconnected bottleneck sub-module. Overall, these components of focus, C3, SPP and convolutional layer formulated the building block of the YOLOV-5 model as presented in the flow chart of Figure 7;

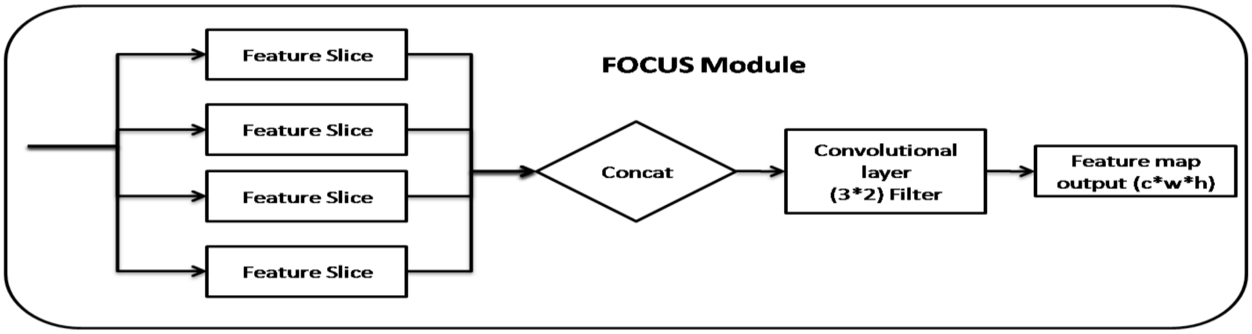

Figure 4: The Focus block module

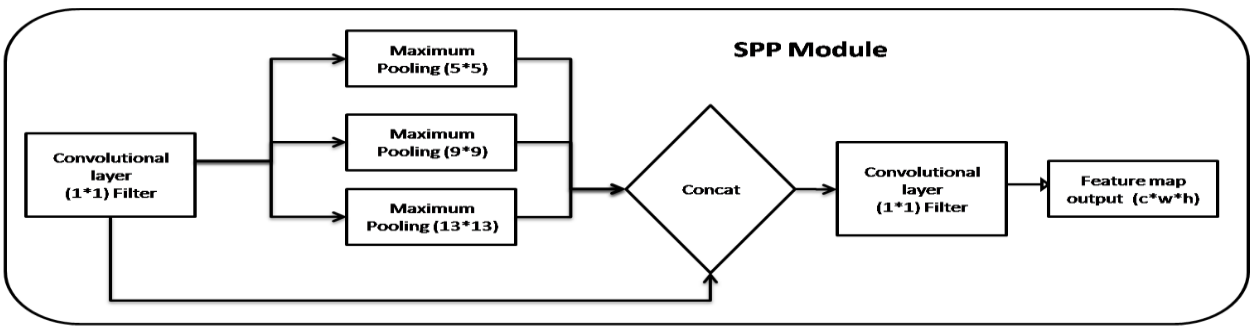

Figure 5: SPP module block diagram

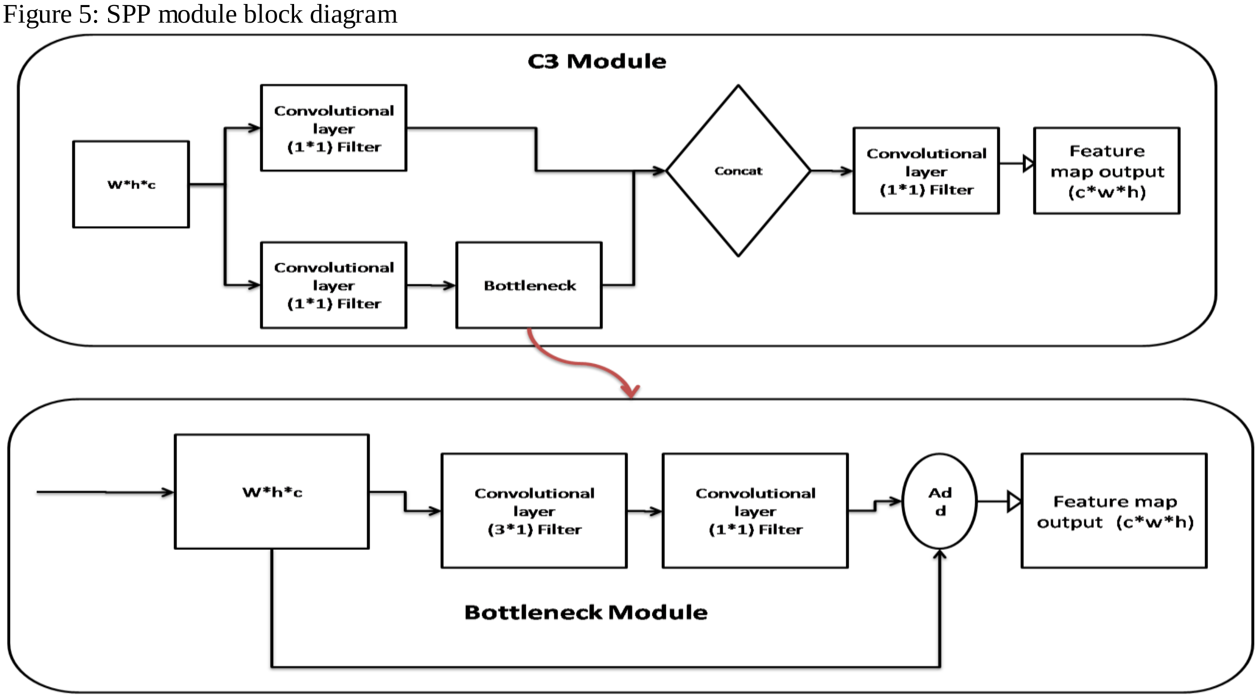

Figure 6: The Block diagram of the C3 Module

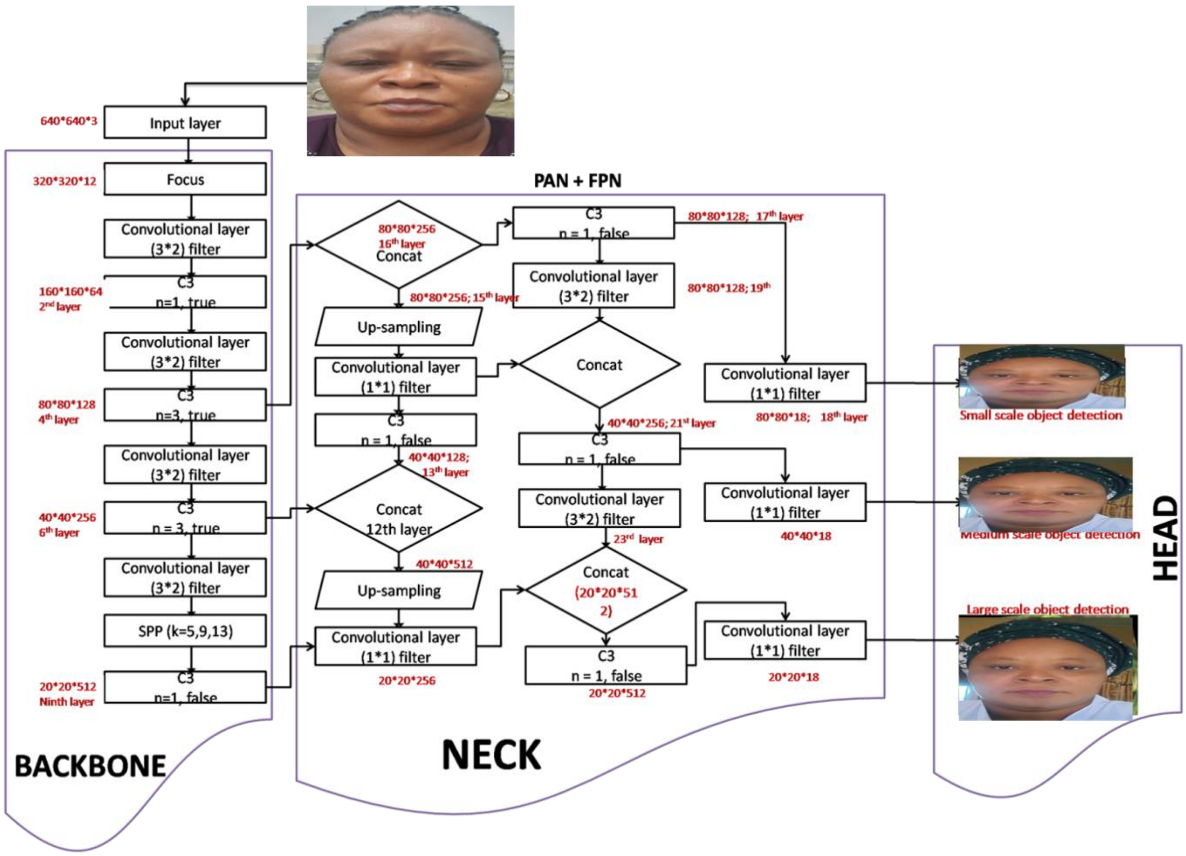

Figure 7: Flow chart of the benchmark YOLOV-5

Suppose the image data was imported to the YOLOV-5 model in Figure 7 and then by the input layer, the backbone first applies the focus to down sample the image at double intervals, by directly outputting the image into 32032012 and performing convolution with a filter kernel of 32. The filter determines the quantity of information extracted from each element of the image in the height of 2 and weight of 3 considered per convolution within the receptive fields and output a feature map. These feature maps are passed through another convolutional layer of 32 kernel filter of the first layer and C3 module of the second layer to output feature maps of 160160 and a colour channel of 64.The feature map is processed through a third convolutional layer with filter size of 32 and C3 in the fourth layer to output a feature size of 8080 and colour channel value of 128. The process continued through the fifth layer with a convolution of 32 and C3 in the sixth layer to output feature map of 4040256. In the seventh layer this feature map is scanned with 32 filter convolutions, and pass through the eight-layer SPP and C3 in the ninth to output the final feature map extracted from the input image as 2020512.

In the neck where these features are aggregated, a 11 convolutional filter is applied to generate the next convolutional layer of 2020256 at the tenth layer. These features are up-sampled in the 11^{th} layer and then output a 4040 512 feature map was spliced in the concat module in the 12^{th} layer slices the features maps and then output in the size of 4040512, which was features are passed through C3 module in the 13^{th} layer and then scan using a 1 1 convolutional filter to output a feature map of 4040 128 colour channels. These features are up-sampled through data augmentation in the 15^{th} layer to produce an output of 8080 and colour channel of 256. These features are passed through a concat module in the 16^{th} layer, which slices the feature maps and pas through C3 module in the layer 17^{th} to output a feature map of 8080128, then apply convolutional filter size of 11 to output a convolutional layer of 808018, applied in the head for the prediction of small student face images.

The same features output of the C3 in the layer 17^{th}arealso passed through a filter of 32 convolutional filter at the 19th layer and then output 4040128 feature map. In the 20th layer, the concat module slices the features maps and output a 4040256. These features are passed though the C3 module in the layer 21stwhich applied 11 convolutional filter scan to output a 404018 convolutional layer which is applied in the head for the prediction of medium size student face image image data.

Additional Output of the C3 in the layer 21 was scanned using 23 convolutional filters in the 23rd layer and output a 2020256 convolutional layer which is passed through the concat layer in the 24th layer responsible for feature slicing which outputs the images feature vectors into a convolutional layer size of 2020512. These features passes through C3 module in the 25th layer which outputs a 2020512 then a 11 convolutional filter is applied to output a 202018 convolutional layer applied in the head for large scale student face image detection. The Figure 8 presents the flowchart of the YOLOV-5 model with the integrated sub-section of the major component.

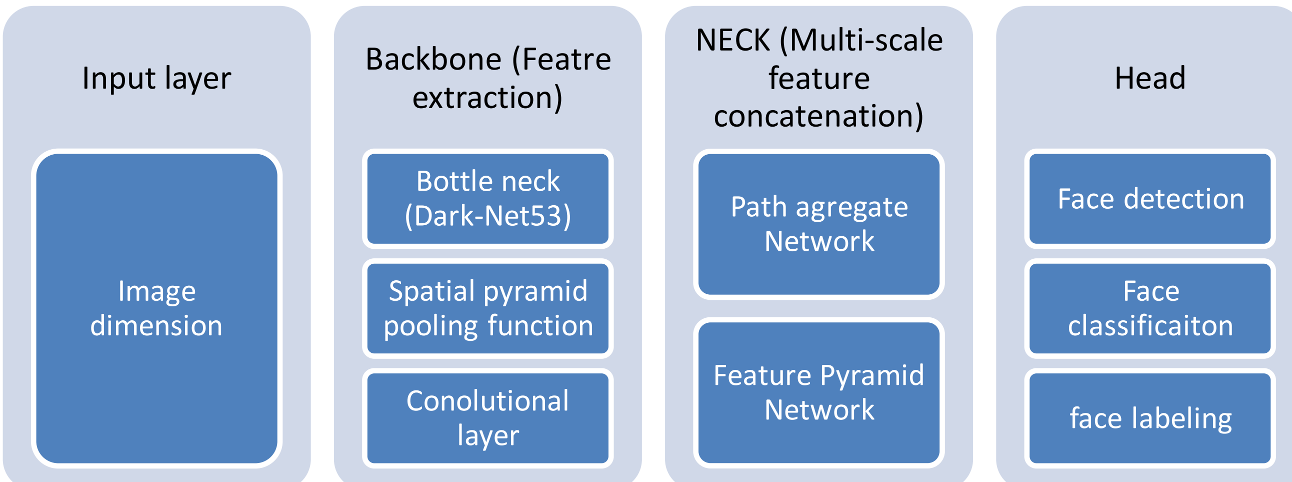

Figure 8: Block diagram of the YOLOV-5 Architecture

5. TRAINING THE YOLOV-5 MODEL

The benchmark YOLOV-5 model was trained in Roboflow platform. The training was performed using the student face image dataset, SGD algorithm and weight decay regularization algorithm. To start the training, first the data was split into training, test and validation sets respectively in the ratio of (80:10:10). The training data was imported first to the YOLOV-5 model in Figure 8. At the data importation, the input layer dimensioned the data into 6406403 and then applies the backbone for the feature extraction process using the focus, C3 and SPP to extract the feature maps. The focus module in figure 3.5 slices the image size into 3203203, then the C3 applied bottleneck, convolutional filter and concatenation process to extract the features maps. The SPP apply maximum pooling techniques to pool diverse images sizes of and concatenate to the head where the training process is performed. To begin training, the YOLOV-5 initializes the hyper-parameters, which is then applied to adjusts the hyper-parameters, while monitoring the gradient loss. During this process, the neurons with high weights are penalized based on the weight decay regularization process to ensure generalization of the models. As the training proceeds, the test and validation sets were applied to evaluate the model considering metrics such as mean precision, recall and accuracy, at iterative epoch. This process continued until consistency was recorded in the output of the evaluation metrics, which implied that the YOLOV-5 neurons have converged, and then the training automatically stops. The parameters for training the model are presented in Table 1.

Table 1: Training parameters

Items	Specification
Momentum	0.937
Epoch; batch size	150; 16
Weight decay	0.0005
Input size	640x640
Learning rate	0.01
Optimizer	SGD
Warm-up bias and momentum	0.8; 0.1
Warm-up epoch	3

6. 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.

7. RESULTS AND DISCUSSION

The evaluation of the YOLOV-5 model trained considers parameters such as Precision (P), Recall (R), Average Precision (AP_i), and mean average precision (mAP). The Equations of the parameters are;

$$P = \frac{TP}{TP+FP} \qquad 1$$

$$R = \frac{TP}{TP+FN} \qquad 2$$

$$AP_i = \int_0^1 P(R)d(R) \qquad 3$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad 4$$

Where TP is true positive student face image detection, FP is negative samples which are positively detected, FN are positive samples which are not detected.

Then the results of the YOLO-5 implementation to generate the facial recognition model is presented. the model was evaluated considering Equation 1 to 4 and the results are reported in Figure 9.

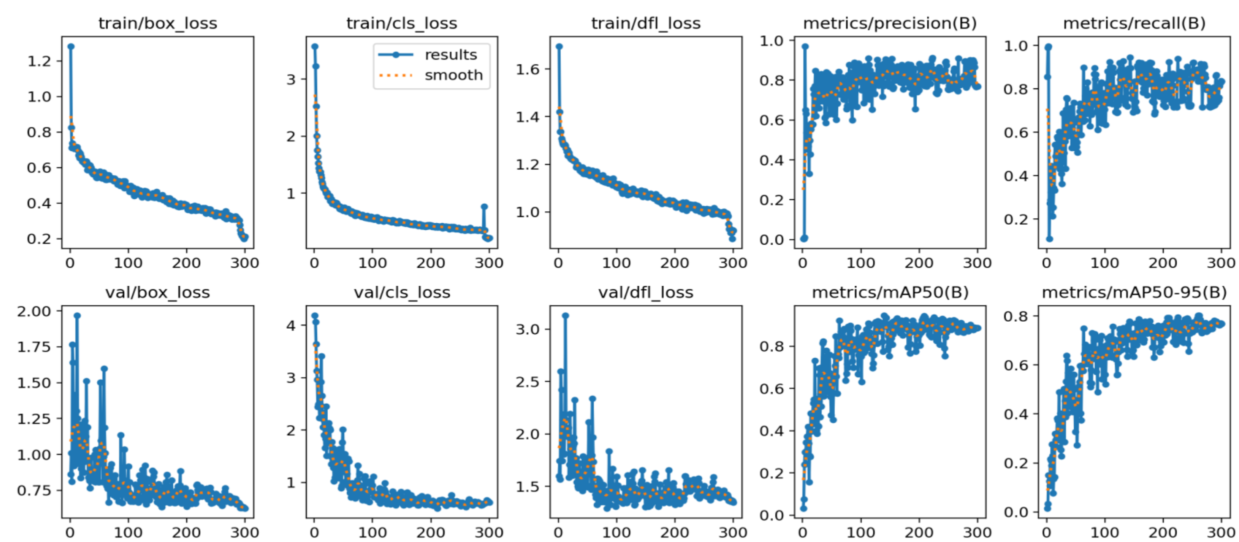

Figure 9: Training results of the YOLOV-5 Model

The training and validation metrics presented in the YOLOv5 results for the biometric face classification system provide a comprehensive overview of the model's performance, highlighting both its strengths and potential areas for improvement as shown in the Figure 9. The train/box_loss graph at top left indicates a steady decline throughout the training process, which suggests that the model is effectively learning to minimize the error in predicting bounding boxes around student faces. This improvement is crucial as accurate bounding box predictions directly impact the system's ability to localize faces accurately, which is foundational for subsequent classification tasks. Similarly, the train/cls_loss demonstrates a consistent decrease, reflecting the model's capacity to improve its classification accuracy over time. A well-performing classification component is essential for distinguishing between different students, ensuring that the authentication process is both reliable and secure. In contrast, the train/dfl_loss remains relatively stable with a slight decline, indicating that while the model is learning to handle hard-to-classify samples effectively, there are instances that still pose challenges. This is an important consideration for the model's robustness, as it implies the need for further refinement to ensure that all difficult cases are addressed adequately. Moving to the metrics for precision and recall, both metrics show a gradual increase, signifying that as training progresses, the model becomes more adept at identifying relevant instances and reducing false positives. High precision and recall are critical for a face recognition system, as they enhance the reliability of the authentication process, minimizing the risk of unauthorized access. The validation metrics further reinforce these findings. The val/box_loss and val/cls_loss trends mirror the training losses, indicating that the model is generalizing well to unseen data. This generalization is vital for real-world applications, where the model must perform consistently across diverse conditions and varied user profiles. However, careful monitoring of these metrics is necessary to prevent over-fitting, particularly if training losses continue to decline while validation losses begin to rise.

The mean Average Precision (mAP) scores, as shown in the metrics/mAP50(B) and metrics/mAP50-95(B) plots, serve as robust indicators of the model's performance. The increase in mAP values signifies that the YOLOv5 model is not only achieving high accuracy in face detection and classification but also maintaining that accuracy across a range of intersection-over-union (IoU) thresholds. This capability is crucial for a biometric authentication system, where precise identification is necessary for successful logins. Overall, the results indicate that the YOLOv5 model exhibits strong performance in the biometric face classification task, with promising implications for enhancing security and efficiency in student login systems.

8. CONCLUSION

This study successfully developed a deep learning-based human-computer interaction system tailored for visually impaired users, significantly improving their access to digital platforms and educational resources. By integrating biometric user authentication via YOLOv5 facial recognition, the system ensures secure, personalized, and seamless login experiences, reducing the need for manual intervention. The accuracy and reliability of the facial recognition system offer a practical solution to authentication challenges faced by users with visual impairments. Additionally, the integration of advanced speech-to-text and text-to-speech functionalities, complemented by an adaptive noise cancellation algorithm using the LMS filter, enhances the system's responsiveness to voice commands. This adaptability to noisy environments ensures clarity in voice recognition and interaction, providing visually impaired users with a more intuitive and user-friendly experience. The system demonstrates the importance of addressing environmental challenges, such as background noise, which can impede effective communication in traditional systems. Ultimately, this study highlights the potential of combining deep learning models and adaptive filtering techniques to create more accessible and inclusive technological solutions. The system's performance, validated through rigorous testing, shows that it can significantly contribute to improving educational accessibility for visually impaired individuals. Moving forward, the continued development and refinement of such systems can pave the way for more inclusive human-computer interactions, empowering individuals with disabilities to achieve greater independence and participation in digital learning environments.

REFERENCES

Abhijna U., Dharani K., kavya K., Roja G., & Juliet K., (2022) Virtual Assistance for Visually Impaired. IJARSCT ISSN (Online) 2581-9429 International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) DOI: 10.48175/IJARSCT-5087

Bahraini T., & Sadigh A., (2024) Proposing a robust RLS based sub-band adaptive filtering for audio noise cancellation. Applied Acoustics 216 (2024) 109755. https://doi.org/10.1016/j.apacoust.2023.109755

Bhukhya C., Bhumireddy K., Palakonalu H., Singh S., Bansod S., Pal P., & Kumar Y., (2023) Virtual Assistant and Navigation for Visually Impaired using Deep Neural Network and Image Processing. DOI: https://doi.org/10.21203/rs.3.rs-2867156/v1

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.

Guan J., Xiao F., Liu Y., Zhu Q., & Wang W., (2023) Anomalous Sound Detection Using Audio Representation with Machine ID Based Contrastive Learning Pretraining. arXiv:2304.03588v2 [cs.SD] 10 Apr 2023

Hande S., & Bilawar P., (2022) Digital Voice Assistant For Visually Impaired Users. International Journal of Advance and Applied Research www.ijaar.co.in

Jarosz M., Nawrocki P., Sniezynski B., & Indurkhya B., (2021) Multi-Platform Intelligent System for Multimodal Human-Computer Interaction. Comput. Inform. **2021**, 40, 83–103.

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

Marvin E., (2020) Digital Assistant for the Visually Impaired. Department of Computer System Engineering, Universitas Prasetiya Mulya, BSD, Indonesia. 978-1-7281-4985-1/20/$31.00 ©2020 IEEE https://www.afb.org/aw/18/2/15244

Mosquera-DeLaCruz J., Loaiza-Correa H., Nope-Rodríguez S., & Restrepo-Girón A., (2021) Human-computer multimodal interface to internet navigation. Disabil. Rehabil. Assist. Technol. **2021**, 16, 807–820.

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

Oey A., (2023) Practical Considerations for Implementing Adaptive Acoustic Noise Cancellation in Commercial Earbuds. Journal of Electronic & Information Systems. 5(2): 25-34. DOI: https://doi.org/10.30564/jeis.v5i2.5998

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.

Prathiba T., & Kumari R.S., (2021) Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer interaction. J. Ambient. Intell. Humaniz. Comput. **2021**, 12, 6215–6229.

Shi C., Huang M., Liu C., & Li H., (2023) Active noise control with selective perceptual equalization to shape the residual sound. Applied Acoustics 208 (2023) 109376. https://doi.org/10.1016/j.apacoust.2023.109376

Sudhakar M., Charan M., Pranai G., Harika L., & Yamini P., (2023) Audio signal noise cancellation with adaptive filter techniques. Materials Today: Proceedings 80 (2023) 2956–2963. https://doi.org/10.1016/j.matpr.2021.07.080

Wang L., (2020) Towards Human-Centered AI-Powered Assistants for the Visually Impaired. Master of Applied Science in Systems Design Engineering. University of Waterloo. Waterloo, Ontario, Canada