Advancing Human-Computer Interaction: A Deep Learning Approach to Eye Tracking

Saliha Benkerzaz[1]*, Abdeslem Dennai [2], Youssef Elmir [3]

[1] Faculty of Exact Sciences, Smart Grids Renewable Energies Laboratory, Tahri Mohammed University, Bechar, 08000, Algeria.

[2] Faculty of Exact Sciences, Smart Grids Renewable Energies Laboratory, Tahri Mohammed University, Bechar, 08000, Algeria.
[3] Laboratoire LITAN, École supérieure en Sciences et Technologies de l'Informatique et du Numérique, RN 75, Amizour, 06300, Bejaia, Algérie.

E-mail corresponding: salihabenkerzaz@gmail.com

Received: 05/05/2023; Accepted: 04/09/2023; Published: 15/09/2023

ABSTRACT

The accelerated progress of human-computer interface (HCI) technologies has generated considerable interest in creating efficient and accurate eye-tracking algorithms. This paper introduces a novel approach utilizing deep learning techniques for real-time eye tracking. Our study's proposed Convolutional Neural Network (CNN) trains on an Eye-Chimera database. The results were notable, as the model exhibited a commendable average accuracy rate of 97.67%. The algorithm under consideration underwent a comprehensive evaluation, which yielded data suggesting a meager error rate of 0.023. This result serves as evidence of the system's resilience and reliability. Moreover, this technology's commendable precision and minimal margin of error make it an invaluable instrument for augmenting user experiences and facilitating accessibility within virtual reality, gaming, and assistive technologies. Consequently, its potential for widespread implementation holds promise for significantly increasing many human-computer interaction applications, yielding good outcomes.

Keywords: Eye tracking, eye detection, convolutional neural network (CNN), human-computer interface (HCI).

Tob Regul Sci. ™ 2023;9(1): 5229 - 5239
DOI: doi.org/10.18001/TRS.9.1.363

I. Introduction

Computer vision is among the main areas that have contributed to the development of artificial intelligence and created many challenges. Among these challenges is tracking eye movements.

Eye movement is a means by which a person can express his needs and desires. With the help of the eye, the world around us can perceive by taking pictures and sending them to the brain for processing.

Eye tracking is a technique that tracks the positions and movements of the eyes to analyze behavior and visual attention [1]. It involves measuring the motion of an eye to the head or the point of gaze (where one is looking).

Checking eye movement contributed to the development of human-computer interaction [2]. This technology permitted the creation of intelligent interfaces that keep in touch with the computer for those with disabilities [3], especially mobility or speech disabilities.

Eye-tracking applications based on computer vision allow setting the camera on one eye or both eyes. So, two areas in this field are distinguished, the first is to locate the eye's location in the image, and the second is to track eye movement and estimate its direction through the data obtained and analyzed. Thus, used directly in applications or tracked via frames in the video in real-time [4]. Eye tracking tools are also used in studies on the visual system [5, 6], psychology [7], marketing, and product design [8] to quantify eye locations and eye movement. Recent studies have shown that deep learning in computer vision successfully resolves various issues, like using deep convolutional neural networks that effectively solve complex problems, which have also demonstrated notable success in image classification-related tasks [9, 10], speech recognition, and language processing.

Eye-tracking models have emerged as powerful tools in human-computer interaction in recent years. These models utilize advanced technology to capture and analyze the intricate patterns of eye movements during various tasks. Eye-tracking models provide insights into cognitive processes, visual attention allocation, and decision-making mechanisms by tracing gaze points and fixations. This article presents a convolutional neural networks model, which analyzes the image and trains it to identify the direction. Our model aims to provide a deeper understanding of How to determine the direction more precisely where we can use this result to manage an intelligent interface.

The second section of our paper provides an overview of the related work, while the third section defines the materials and procedures employed in our study. The fourth section of the paper presents the study's results, followed by a detailed discussion. The last section serves as the conclusion, summarizing the main points of the research and suggesting different possibilities for future work.

Related Work

Choi et al. [11] Introduced a methodology based on a five-layered convolutional neural network (CNN) to describe the gaze zones of drivers and predict their head pose. The researchers used a

dataset of male and female drivers and eyeglass wearers to identify nine distinct areas of visual focus. The accuracy of this model exceeded 95%.

In their study, Vora et al.[12] devised a system for detecting gaze using a Convolutional Neural Network (CNN). This technique comprises two main components: a pre-processing unit and a fine-tuning unit. The pre-existing models, namely AlexNet and VGG 16, underwent independent fine-tuning processes, which yielded accuracy rates of 93.36% and 88.91%, respectively.

The CNN-based model created by Naqvi et al. [13] utilizes a near-infrared camera to effectively collect frontal view photos of drivers, ensuring that their vision remains unobstructed.

Marko et al. [14] introduce a novel architecture for eye movement classification, including deep convolutional neural networks to extract visual features and recurrent layers to capture temporal information. The architectural model demonstrated a validation accuracy of 92% and an accuracy of 88% on real-time tests, employing a conventional laptop and web camera.

Guo et al. [15] proposed the use of the tolerant and talented (TAT) training strategy for convolutional neural networks (CNNs) as a means to mitigate the problem of overfitting. The approach employs cosine similarity pruning and aligned orthogonal initialization, resulting in a 90% accuracy when evaluated on the Gaze Capture dataset.

Andronicus et al.[16] present a novel approach for enhancing gaze estimation using a Convolutional Neural Network (CNN). The proposed strategy incorporates both full-face and 39-point facial landmark images. The experimental results demonstrate a level of accuracy of over 70% and a notable performance improvement when utilizing a set of 39-point facial markers.

The study by Fangyu et al. [17] uses deep learning methodologies to ascertain the presence of Alzheimer's disease (AD) by analyzing eye-tracking patterns. An investigation is undertaken wherein a 3D visual task is performed to gather heat maps of visual attention. These heatmaps are subsequently utilized to construct a multilayered comparison convolution neural network (MC-CNN). The MC-CNN recorded an accuracy of up to 83%. It demonstrates constant validity in accurately categorizing individuals with Alzheimer's disease (AD) and those without the condition using eye-tracking data.

Anurag et al. [18] presented an artificial intelligence (AI) model, which has been trained using eye tracking data obtained from individuals lacking expertise in the field and demonstrates the ability to forecast eye fixations on optical coherence tomography (OCT) reports with an average accuracy rate of 88%. This development holds the potential to assist in training AI systems that incorporate eye movement information, enhance the interpretability of convolutional neural network (CNN) models, and contribute to advancements in medical education.**Table 1** summarizes the pertinent research that is taken into account for this study.

Table.1 Overview of the related work and research conducted in the field.

Paper	Dataset	Model	Accuracy
Choi et al. (2016)[11]	Own dataset	five-layered CNN-based methodology	95%.
Vora et al.(2017)[12]	Own dataset	VGG16 and AlexNet	93.36%
Naqvi et al. (2018)[13]	Own dataset	VGG16	92.8%
Marko et al. (2018)[14]	Own dataset	proposed CNN model	92%
Guo et al.(2019) [15]	Gaze Capture	TAT for CNN	90%
Andronicus et al.(2022)[16]	GazeCapture TabletGaze	proposed CNN model	70%
Fangyu et al. (2023)[17]	eye-tracking data	MC-CNN	83%
Anurag et al.(2023)[18]	line-drawing dataset	proposed CNN model	88%
Our model	Eye-Chimera database	proposed CNN model	98%

In this section, a thorough review undertaken to assess eight pertinent studies in the field. Each work was rated based on its level of accuracy and the specific database utilized. These publications jointly give an essential foundation for understanding the scope of our research field. Significantly, we have organized our research results in a comprehensive tabular format, encompassing the recorded accuracy percentages. It is essential to mention that our study has provided an innovative convolutional neural network model, which has been enhanced with unique features. We will elaborate on these aspects in the following parts. Significantly, our model achieved the best accuracy percentage compared to all the research examined, highlighting our method's potential and innovative nature. As the study advances, we delve deeper into the complexities of our model and its significant results, demonstrating its valuable contributions to the respective field.

I. Materials And Methods

A. Eye-Chimera Database

Eye-Chimera database (**Eye** part from the **C**ognitive process **I**nference by the **M**utual use of the **E**ye and exp**R**ession **A**nalysis) is a comprehensive collection of 1170 front-face images. These images have been categorized based on the seven directions of the gaze. The database also identifies five points per eye, including the center point and four points defining the bounding box [19]. The construction process involved the participation of a group of 40 individuals, all falling within the age range of 20 to 30. Their actions were meticulously documented by a predetermined pattern, utilizing Canon 600D cameras with a resolution of 640×480 and Panasonic HDC-TM 60 cameras with a resolution of 1920×1080. Adequate lighting conditions were also employed during the recording process. This regulation pertains to the intricacies of the seven EAC scenarios, encompassing the neutral and optically defocused stance [20].

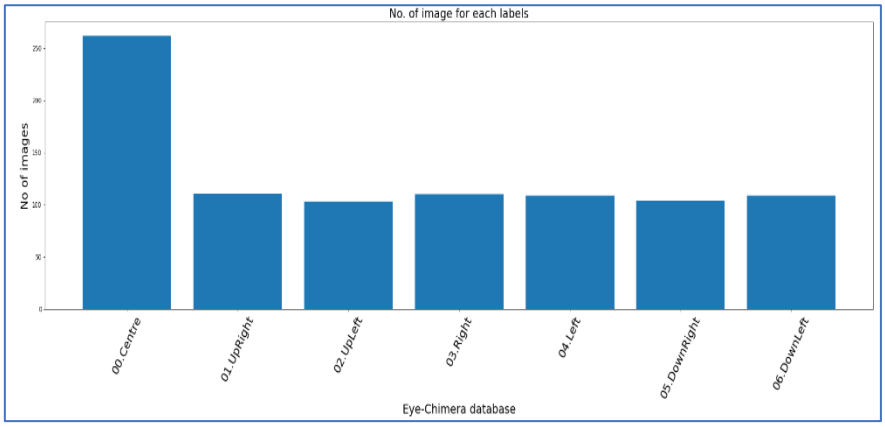

Figure 2- Eye-Chimera database

B. Methodology

Several types of neural networks exist, among which convolutional neural networks are widely employed in computer vision for various tasks. One such job is gaze estimation, which finds application in eye-tracking systems. Eye tracking, however, includes monitoring an individual's eye movements and can serve as a method of interaction with convolutional neural network (CNN)-based applications or as a data resource for enhancing the efficacy of computer vision models. The interconnection between these two entities is rooted in their mutually supportive functions within contexts of human gaze and visual perception.

This paper involved the development of a convolutional neural network (CNN) model tailored explicitly for eye image classification.

1. Data Preprocessin

In the data preprocessing phase shown in Figure 1, the script effectively prepares the images and their accompanying labels to be utilized in the deep learning model. The algorithm sequentially processes each image within the 'train generator' by iterating through the batches of data. Initially, the picture receives a conversion of its color representation from RGB to BGR with OpenCV, followed by a subsequent conversion to grayscale. Afterward, the script enlarges the grayscale image to a standardized dimension of 128x128 pixels and performs pixel value normalization to ensure uniformity. By employing cascading classifiers, the system can detect the existence of eyes and mouths inside the processed image. When the system detects the presence of both eyes and mouths, it extracts and resizes the region containing the eyeballs to dimensions of 128x28 pixels. Once the one-hot encoding is transformed into class indices, the processed eye regions are appended to the 'XTrain' list, while the matching labels are incorporated into the 'YTrain' list. The careful preparation of the data guarantees that it is suitably organized and prepared for subsequent model training and analysis.

2. CNN Architecture

The model's architecture, depicted in Figure 1, comprises several layers collaborating to acquire knowledge and identify patterns within the input images. The input images, which have a height of 28 pixels and a width of 128 pixels, undergo a sequence of convolutional layers. Distinct numbers of filters and kernel sizes characterize each layer. Following the convolutional layers are max-pooling layers, which help extract crucial information by diminishing the spatial dimensions of the feature maps. To mitigate the issue of overfitting, it is common practice to incorporate dropout layers into the neural network architecture. These dropout layers are positioned deliberately after each max-pooling layer, with a dropout rate of 0.3. The retrieved features are further processed, and the fully linked layers make predictions located at the end of the network. The model is finalized by incorporating a softmax activation function at the output layer, which effectively categorizes input images into one of seven unique classes. The architectural design of this system optimizes it to achieve exceptional performance in the domain of image classification, as evidenced by the experimental findings obtained in our study.

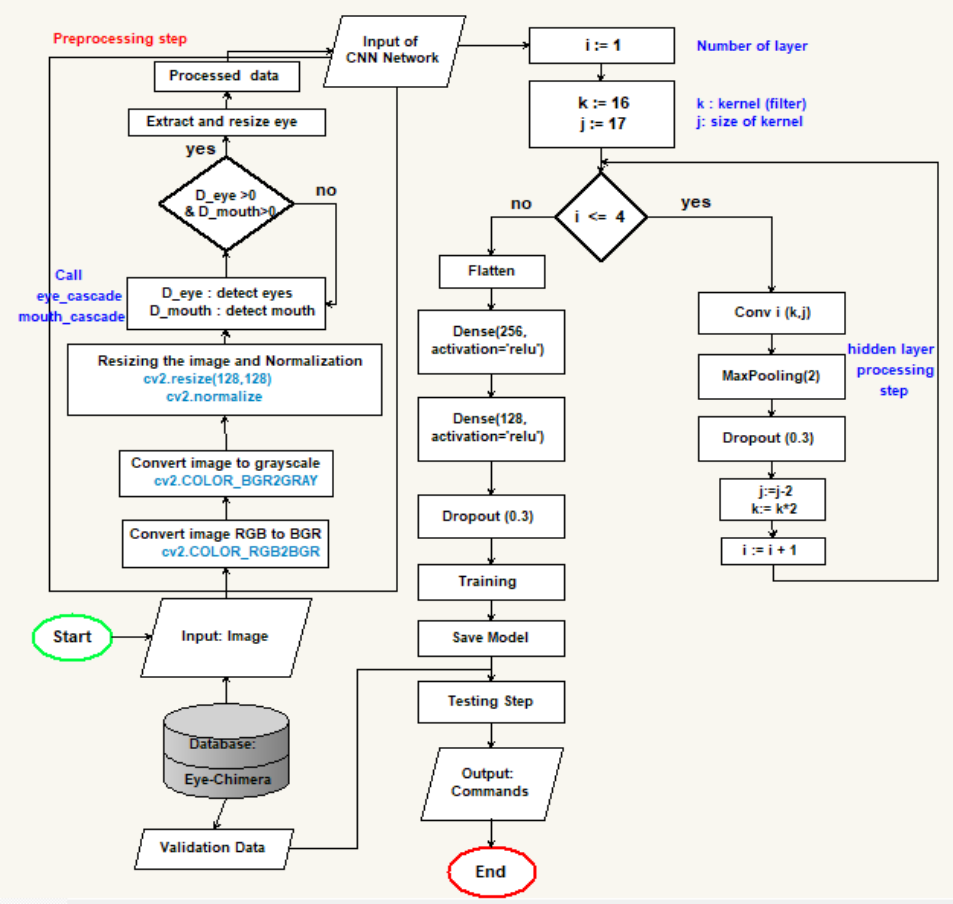

Figure 1- CNN model proposed architecture

II. Result And Discussio

The proposed approach consists of two steps, presented and discussed in this part, along with its results. During the first stage of data preprocessing, we identify the regions of the eyes that provide the primary information upon which our proposed model is constructed. The second step involves the suggested neural network's architecture, which consists of four hidden layers with the characteristics described in the previous session. The results obtained are shown in the following figures.

The assessment of our model's performance demonstrates exceptional results, indicating its efficacy in addressing the intricate issue of image classification. The model demonstrated exceptional performance during the training process, as shown in Table 2, by achieving a significantly low loss value of 0.0605. This result indicates the model's effectiveness in reducing errors in its predictions. In addition, the model demonstrated a notable accuracy rate of 97.67%, signifying its proficiency in generating accurate predictions on the validation or test dataset. The estimated error rate of 2.33% further supports the model's competency. This low percentage indicates a minimal difference between the predicted outcomes and the actual values, providing additional evidence for the model's accuracy.

Table.2 The model's experimental results (accuracy, loss value, and error rate).

	Dataset	Accuracy	Loss value	Error rate
CNN proposed	Eye-Chimera	97.67%	0.0605	2.33%

Figure 2 shows that the training and validation accuracy form a similar trajectory throughout the ascent; simultaneously, the training and validation loss curves demonstrate a matching drop, which suggests that our model is exhibiting satisfactory performance. The similarity between the training and validation curves indicates that the model shows proficient learning behavior concerning the training data and can apply its acquired learning to novel validation data. In essence, the synchronized movement of the curves indicates that the model is neither excessively fitting the training data (overfitting) nor failing to capture the underlying patterns (underfitting). Instead, it follows a well-balanced and resilient learning trajectory, which serves as a positive indication of its quality and ability to generalize.

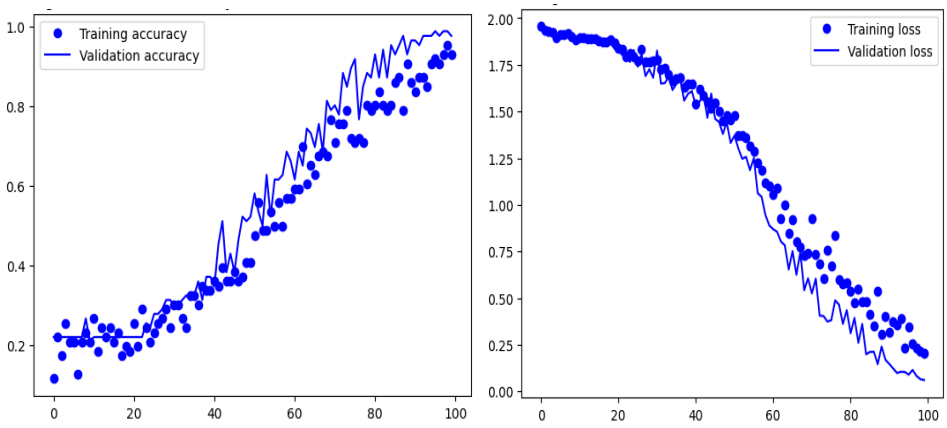

Figure 2- The training and validation process for the Convolutional Neural Network (CNN) model using the Eye-Chimera database.

Figure 3 gives a detailed examination of the data using the classification report and the confusion matrix. These details allow for a more comprehensive evaluation of each class individually, revealing a nuanced perspective. The precision values, ranging from 0.86 to 1.00, indicate the model's ability to accurately identify instances of each class. On the other hand, the recall values, particularly the value of 0.85 for class 1, demonstrate the model's effectiveness in correctly recognizing the actual instances within each category. The F1-score demonstrates a harmonious balance between these metrics and presents a cohesive equilibrium between both measurements.

Regarding the overall performance, the model demonstrated a high accuracy rate of 98%, thereby validating its proficiency in accurately categorizing a wide range of occurrences. Furthermore, the macro and weighted average values, roughly 0.98, demonstrate a robust overall performance, considering the disparities in class distribution. In summary, the extensive findings validate the effectiveness of our model, highlighting its appropriateness for the designated picture

classification objective and providing persuasive proof of its accomplishments in our study. These results serve as a monument to the efficacy of deep learning in addressing practical problems.

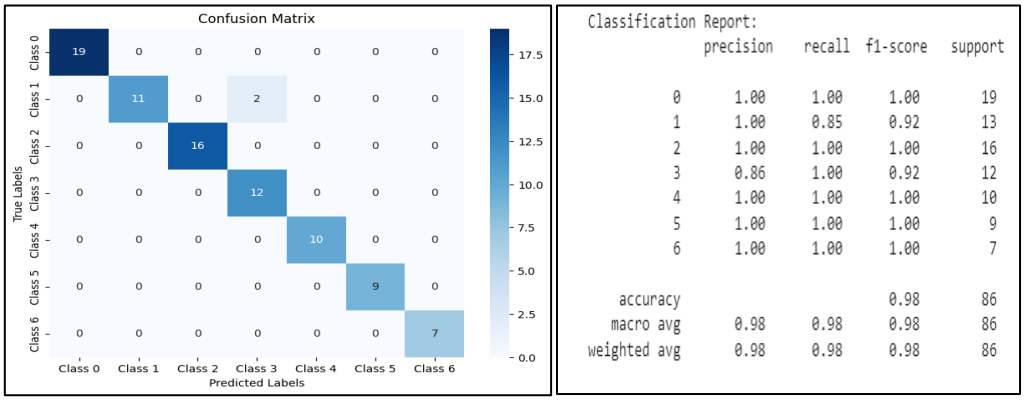

Figure 3- The classification report and confusion matrix for the proposed Convolutional Neural Network (CNN) model.

III. Conclusion

This article introduces a convolutional neural network model that analyzes images and trains them to accurately identify eye direction, aiming to improve intelligent interface management. By conducting a comprehensive analysis of pertinent related work and introducing our innovative convolutional neural network model, we have effectively showcased the capacity for substantial progress in the specified research domain. The model demonstrated exceptional accuracy, surpassing previous studies, and included novel aspects. The results above exhibit potential for various applications or implications.

As we contemplate the future, we must persist in exploring the potentialities unveiled by this research. Subsequent investigations could further explore eye direction detection by detecting sound, lip movement, or hand movement. Moreover, it expands the utilization of our model to encompass the development of intelligent interfaces to control devices, utilizing this framework to construct robotic systems. As we contemplate the advancements achieved thus far and the path ahead, we must acknowledge that each significant discovery propels us toward a more promising and enlightened future.

References

[1] Klaib, A. F., Alsrehin, N. O., Melhem, W. Y., Bashtawi, H. O., & Magableh, A. A. (2021). Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Systems with Applications*, *166*, 114037.

[2] Goldberg, J., Stimson, M., Lewnstein, M., Scott, N., and Wichansky. (2002) Eye Tracking in Web Search Tasks: Design Implications. In Eye Tracking Research & Applications (ETRA) Symposium. New Orleans, LA

[3] Majaranta, P., and Raiha, K.-J. (2002) Twenty Years of Eye Typing:Systems and Design Issues. In Eye Tracking Research &Applications (ETRA) Symposium,. LA.

[4] Al-Rahayfeh, A., & Faezipour, M. (2013). Eye tracking and head movement detection: A state-of-art survey. *IEEE journal of translational engineering in health and medicine*, *1*, 2100212-2100212.

[5] Greene, H. H. and Rayner, K. (2001) Eye Movements and Familiarity Effects in Visual Search. Vision Research, 41(27): 3763-3773.

[6] Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. Journal of memory and language, 59(4), 457-474.

[7] Rayner, K. (1998) Eye Movements in Reading and Information Processing: 20 Years of Research. Psychological Bulletin, 124(3): 372-422.

[8] Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., and Duffy, S. A.(2001) Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements. Journal of Experimental Psychology: Applied, 7(3): 219-226.

[9] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211- 252.

[10] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303- 338.

[11] H. Choi, Y. G. Kim, and T. B. H. Tran, ''Real-time categorization of driver's gaze zone and head pose -using the convolutional neural network,'' in Proc. HCI Korea, Jan. 2016, pp. 417–422.

[12] Vora, S., Rangesh, A., & Trivedi, M. M. (2017, June). On generalizing driver gaze zone estimation using convolutional neural networks. In 2017 IEEE Intelligent Vehicles Symposium (IV) (pp. 849-854). IEEE.

[13] R. Naqvi, M. Arsalan, G. Batchuluun, H. Yoon, and K. Park, ''Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor,'' Sensors, vol. 18, no. 2, p. 456, Feb. 2018.

[14] Arsenovic, M., Sladojevic, S., Stefanovic, D., & Anderla, A. (2018, March). Deep neural network ensemble architecture for eye movements classification. In 2018 17th International Symposium Infoteh-Jahorina (Infoteh) (pp. 1-4). IEEE.

[15] Guo, T., Liu, Y., Zhang, H., Liu, X., Kwak, Y., In Yoo, B.,.. & Choi, C. (2019). A generalized and robust method towards practical gaze estimation on smart phone. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 0-0).

[16] Akinyelu, A. A., & Blignaut, P. (2022). Convolutional neural network-based technique for gaze estimation on mobile devices. *Frontiers in Artificial Intelligence*, *4*, 796825.

[17] Zuo, F., Jing, P., Sun, J., Ji, Y., & Liu, Y. (2023). Deep Learning-based Eye-Tracking Analysis for Diagnosis of Alzheimer's Disease Using 3D Comprehensive Visual Stimuli. *arXiv preprint arXiv:2303.06868*.

[18] Sharma, A., Tian, Y., Kaushal, S., Zukerman, R., Chen, R. W., Liebmann, J. M., ... & Thakoor, K. (2023). A Foundational CNN Model for Predicting Eye Fixations on OCT Reports. Investigative Ophthalmology & Visual Science, 64(8), 363-363.

[19] Vrânceanu, R., Florea, C., Florea, L., & Vertan, C. (2013). NLP EAC recognition by component separation in the eye region. In Computer Analysis of Images and Patterns: 15th International Conference, CAIP 2013, York, UK, August 27-29, 2013, Proceedings, Part II 15 (pp. 225-232). Springer Berlin Heidelberg.

[20] Florea, L., Florea, C., Vrânceanu, R., & Vertan, C. (2013, September). Can Your Eyes Tell Me How You Think? A Gaze Directed Estimation of the Mental Activity. In BMVC.