A Comparative Study of Various Convolutional Neural Network Architectures for Eye Tracking System

Jennifer, Joevian Krislynd, Steven Aprianto, and Derwin Suhartono
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: {jennifer017, joevian.krislynd, steven.aprianto}@binus.ac.id, dsuhartono@binus.edu

***Abstract*—Eye tracking system is a technology where it can detect, trace, and analyze the whole eyes movements. Eye tracking system has been used in many fields and sectors so the number of studies for this system has been increased significantly. Eye tracking method has been switched to machine learning. One of the most used algorithms for this system is Convolutional Neural Network (CNN) which has a lot of architectures, such as LeNet-5, AlexNet, VGG16, GazeNet, and ResNet-18. This paper aims to find the most appropriate Convolutional Neural Network's Architecture for building an eye tracking system by comparing chosen architectures that have been mentioned above. The study showed that ResNet-18 has the highest result as it gets 90.53% accuracy while the other results are lower than it.**

***Index Terms*—eye tracking system, convolutional neural network, convolutional neural network's architecture**

I. Introduction

As its name, eye tracking is where eyes can be detected, traced, and analyzed with the help of technology. It has been used in several fields such as psychology, marketing, academic, medical, research, human computer interaction, design, and many more. With eye tracking technology, people can get 'much honest' information from the subject. The reason is that it can show human behavior without relying on human memory so that subject may not be aware of their eyes reflecting behavior. Besides that, eyes also provide rich information about a human where it can give cognitive processes and emotions from human. Some common indicators that are caught by eye tracking technology are pupil dilation, eye movement, gaze, fixation, and blinks count [1].

Eye tracking has become more accessible lately and its visibility in the market has been rising steadily. More research has been done as time goes by so that more datasets and source codes number is getting huge. Those researchers who are invested in eye tracking technology are mostly collaborating with each other to standardize and improve the methodologies in eye tracking. As can be seen in Fig. 1, in 2020, the global eye tracking market was hitting the number USD 368 million and is expected to hit USD 1.75 billion by 2025 [2].

Manuscript received March 21, 2022; revised August 12, 2022.

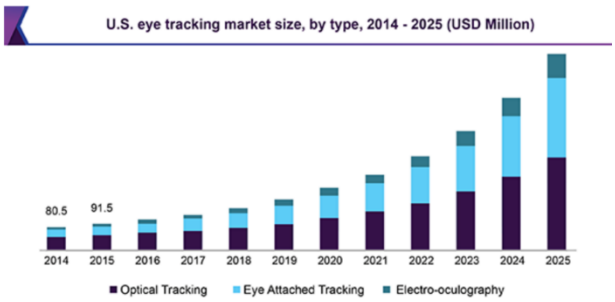

Figure 1. Eye tracking market size prediction.

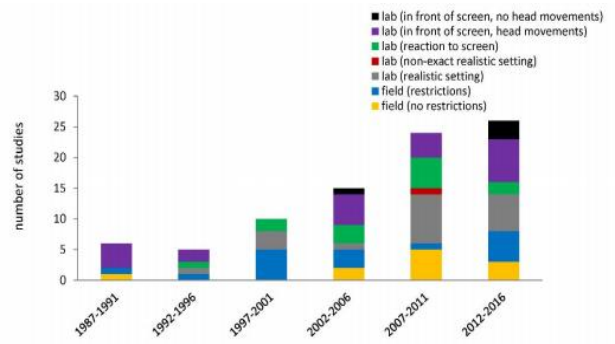

Figure 2. Numbers of eye tracking studies in laboratory.

Following the previous data, the number of research studies including eye tracking technology had also raised from year to year. Fig. 2 shows the data about counts of research studies that have been done from 1987 until 2016 where the data is limited to studies that have been done in laboratory only [3].

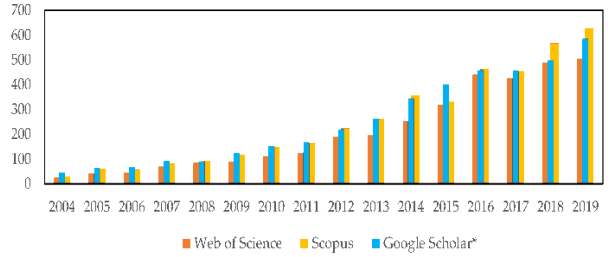

Figure 3. Search results for "eye-tracking" in paper titles.

While the above data shows eye tracking studies in laboratory only, Fig. 3 shows the numbers of eye tracking

doi: 10.18178/joig.10.4.178-183

studies by showing the numbers of papers in three website (Web Science, Scopus, and Google Scholar) that use "eye-tracking" as part of their title [4]. By adding "eye-tracking" to the title, it can be confirmed that those papers at least use eye tracking technology to support their paper, no matter what the main topic is.

Some data that have been shown are proved by the number of implementations that can be done by eye tracking technology. There are various applications of eye tracking technology, in example attentional bias task. It is where the researcher will analyze the preferential allocation of cognitive resources to the detection of stimuli. Usually, it is used to detect psychological tendencies such as depression, posttraumatic stress disorder, anxiety, and addictive disorder [5]. Another example of implementation for eye tracking technology is eye tracking mouse [6] where someone can control the mouse – as in computer device – by their eye movement. Besides those, there are also other implementations like assistive technology, interaction monitoring, facial expression recognition, auxiliary driving, and medical diagnosis. Eye movement analysis also can be used to discriminate patients with depressive disorders from controls, as well as patients with bipolar disorder from patients with unipolar depression by investigating mood regulation and psychomotor disturbances in depressive disorder [7]. Another study [8] examined whether physiological reactivity to depression-relevant stimuli, measured via pupil dilation, serves as a biomarker of depression risk among children of depressed mothers.

Even from the plenty amount of implementation results, eye tracking itself has several issues such as lighting camera quality, and subject angle. Those issues sometimes can reduce the performance of eye tracking technology since the machine will not process the input properly if there are some detentions. By low or too bright lightning, the machine could probably inaccurately recognize subject's eyes, while improper angle could bring to miscalculation of the indicators. Bad camera quality covers all the issues where it could reduce the detection and calculation accuracy.

Based on the problem, this paper will conduct a comparative study against various Convolutional Neural Network (CNN) architectures. Some architectures that will be used are LeNet-5, AlexNet, VGG16, GazeNet, and RestNet-18. The main goal from this research is to find the most ideal architecture for eye tracking system.

II. Literature Review

A. Eye Detection

Eye detection is an important aspect in various applications such as eye detection/recognition, human computer interface, or driver behaviour analysis. Human eye's locations are essential information for many applications, including psychological analysis, facial expression recognition, auxiliary driving, and medical diagnosis [9].

A lot of data can be extracted from eye detection, such as eye gaze, fixation (aggregation of gaze points) [10], pupil dilation, number of blinks, etc. Meanwhile, interesting things that have become the focus of many studies are saccades [10] which are changes from one fixation to another and glissade which refers to slower eye movements [11].

B. CNN Architecture

There are 5 different architectures of CNN that are compared namely, LeNet-5 [12], AlexNet [13], VGG16 [14], ResNet-18 [15], and GazeNet [16]. These architectures are being used without any modification or arrangement of their hidden layer. The composition of hidden layers which are the default form, for each architecture is being explained below:

- LeNet-5

One of the simplest CNN architectures with 2 convolutional layers, a pooling layer or also known as sub-sampling layer, and 3 fully connected layers

- AlexNet

This architecture is a more complex form of LeNet5 with addition of hidden layers and parameter being used. The composition is 5 convolutional layers, 3 fully connected layers, with 3 pooling layers

- VGG16

Much bigger composition and complexity than AlexNet, this architecture has 16 hidden layers: 13 convolutional layers, and 3 fully connected layers. Though the filter being used by VGG16 is smaller than AlexNet's

- ResNet-18

This variant of Residual Net (ResNet) might be the simplest form of ResNet that is still being used nowadays, applies 18 hidden layers: 17 convolutional layers and a fully connected layer (plus an additional softmax layer)

- GazeNet

Make use of parallel fully connected layer (two branches of 3 fully connected layer). The number of convolutional layers for this architecture is also 3

C. Related Work

At first the most used method to build an eye tracking system was by implementing any changes to pixel collected by the camera then make use of threshold to get position of pupil per time. That sequential information then analysed with an event detection algorithm such as, dispersion-based algorithm and velocity-based algorithm. Output of analyzation task is statistical mapping between calculated values and time. There was research in 2013 [6] that aimed to build a low cost HCI device as an Assistive Technology (AS) that targeted for people with disabilities especially for those who are restricted in motion and has speech impairment to help patients with motor and speech impairment communicate and perform basic task. The proposal was an Eye Tracking Mouse (ETM) system controlled by wearable glasses that is embedded with webcam and processing software. Eye Tracking System were built using a proposed method called Eye Tracking with Adapted Segmentation Threshold (ETAST) that combined with Starburst algorithm to determine pupil location. Further next the generated data are processed

and transformed in some way so that it becomes a signal that triggers mouse motion.

Research in 2018 [17] discussed on how eye tracking technology can enhance the pilot-aircraft interaction. Elaborating several eye movement characteristics, such as dwell times, time without fixation, and blinks, those indicators were then integrated according to different flight phases or aircraft configurations. From the analyzing result, it shows how cockpit has been divided into several Area of Interest (AOI) and it also considered how non-monitored AOI should be also included in Flight Eye-Tracking Assistance (FETA), not only the one that being monitored. At the following year (2019), there was a study [18] that combines eye fixation on AOI matrix and emotional response from 60 participants intended to investigated correlation between familiarity and visual signal of consumer response. The result was that eye tracking task, facial expression, and self-reported response were prominent things in doing the analysis of variance.

Eye movement technology opens a new way for the study of automatic detection of depression. A study in 2020 claimed that eye movement signals acquired by eye tracking technology (in this case using Tobii T120 with 120 Hz sampling frequency), are all direct reflections of the brain's information processing demand and can quantitatively characterize emotional perception [19]. This study gathered two various data: eye movement behavioural signals which are the position and time recorded during the stimuli task; and pupil dilation. All data were collected from 96 participants which 48 of them has been claimed were on depressed situation. Similar use of this multimodal research has been proposed with different eye tracking systems, SMI RED eye tracker that was developed by Prosper tech company in Germany. This study [20] claimed that to detect depression, combining mental health self-examination data and eye tracking data to extract multi-modal features is an ideal way to build a robust depression detection system.

Massive growth of Machine Learning and Deep Learning has affected the method for producing and building eye tracking system. The old method is continuing being used today but is not as popular as the newer one. Popular Machine Learning technique that is being used is Convolutional Neural Network (CNN). Back then in 2016, there is research about using CNN for eye tracking which it is configured for robust pupil detection. CNN algorithm was used two times where both stages evaluate regions with CNN. The difference is one image is downscaled and the other one is pure grayscale. This research was done by using over 79,000 images which the 41,000 are complementary to existing image from the literature. The result is it showed how it outperformed state-of-the-art approaches by a large margin while avoiding high computational costs [8].

At that time huge number of studies in eye tracking field using CNN has significantly increased. Various CNN configuration, for example in architecture has been proposed to train a reliable eye tracking system. A study held in 2018 [19] proposed the use of Cascaded Convolutional Neural Network (Cascaded CNN). This research empowers GI4E public dataset is used, which contains 103 test subjects with 1236 visible facial images and the BioID dataset which consists of 1,521 frontal face images with significant variation in illumination and head pose. At first, the local extreme points and gradient values in the full facial image are calculated outputting several candidate eye regions. Generated eye region is evaluated by the first set of CNN to determine the eye region and eye class (right or left). Furthermore, the second CNN was used to find the centre of the eye. The proposed method from this research resulted 85.6% performance.

Latest proposed method in eye tracking field was Cascaded Faster R-CNN with Gabor Filters and Naïve Bayes (FRCNN-GNB) [21] combining Recurrent CNN setup with Gabor filters and Naïve Bayes model. This method used pre-trained Res-Net 50 networks to classify CASIA Iris-Distance database V4 as trained data. To be specific this research made use of 2,567 ultraviolet images from 142 subjects. The baseline was Faster R-CNN setup. The study surpassed and enhanced its baseline on the precision score from 0.96 to 0.99.

Since eye detection technology has been replaced from traditional (making use of threshold) to machine learning based method, a comparison study of different machine learning eye detector has been chosen as the topic of this paper. To be specific, this paper will be comparing the Convolutional Neural Network's Architecture.

III. Methodology

A. Workflow

The dataset in this paper is the MPIIGaze dataset [22]. Fig. 4 shows how the dataset is being processed. Carry out data preparation such as normalizing the dataset to get the optimal form of the data. Process the data using the CNN algorithm, on the grounds that CNN has the capability to process data in the form of images. After going through some processing, results such as gaze estimation and eye placement are obtained.

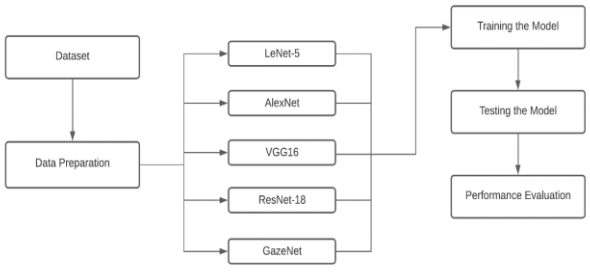

Figure 4. Workflow.

B. Dataset

The dataset for this paper is obtained from an existed dataset that has been collected before and the data type is image in .mat format while the original one is image in .jpg format. The whole dataset itself becomes a file

in .h5 format. The total data for this dataset is 213,659 images from total 15 participants. The number of images collected by each participant varied from 34,745 to 1,498 images. Fig. 5 shows one of the original images on the dataset.

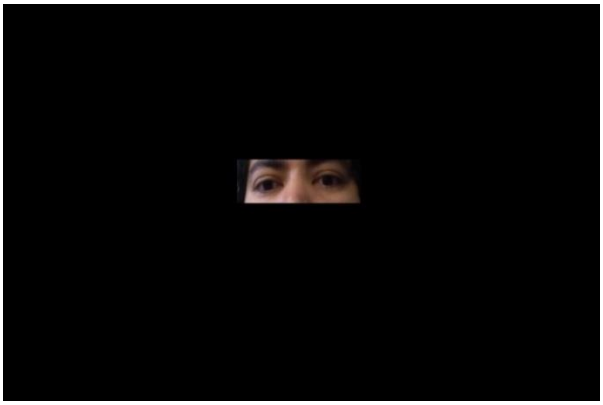

Figure 5. MPIIGaze dataset image participant 00 day 01.

C. Dataset Preparation

To set the effective dataset, the dataset should be prepared. In this paper, the dataset is normalized to separate between left and right eyes and to get the subject's gaze target. At first, image from the dataset is loaded, then the centre of eyes, head pose, and gaze target from the image are collected. After that, both left and right eye images are normalized while being adapted based on the head pose and gaze target. Adjustment is done by converting head pose and gaze target to the polar coordinate system. Each head poses and gaze target is different between left and right eye. The result of normalization is in .mat file and the whole data are combined into one file with .h5 format. After being normalized, dataset is split into two, which are training dataset (80% or 170,927 images) and testing dataset (20% or 42,732 images). Training dataset is used for training the model while testing dataset is used for showing the result of the model.

D. Algorithm

The algorithm that is used is Convolutional Neural Network (CNN) with 5 different architectures that have been explained in the previous section. The study did not make any enhancement nor modification from all the default setup of those 5 architectures. The Rectified Linear Unit (ReLU) activation function was applied to all these architectures. Adam was chosen as the optimizer while the learning rate is set to 0.01.

E. Data Analysis

Analysis plan for this study is by measuring each chosen CNN Architecture performance in gaze measurement and determining eye position with accuracy metric as the core standard. Quality and performance of any system, including eye tracking system, are depended on how accurate delivered data is from the system [23]. Accuracy on an eye tracking system is measured by how similar the position of the real situation, which are eye boundary and gaze direction, with the one being generated by the system. The formula for calculating the accuracy of eye tracking system is explained below [24].

$$Accuracy = \sqrt{((TargetX - GazePointX)^2 + (TargetY - GazePointY)^2)}$$

IV. Result and Discussion

Comparing five architectures of CNN algorithm which are LeNet-5, AlexNet, VGG16, GazeNet, and ResNet-18, the result shows the difference between the accuracy from each architecture. The study measurement is based on CNN models' accuracy per epoch which indicates how correct the boundary of an eye predicted is by the system as well as the gaze direction. Training phase is set to be done in 25 epochs with no validation. The dataset is split into 80% for training (170,927 images) and the rest (42,732 images) is used for testing.

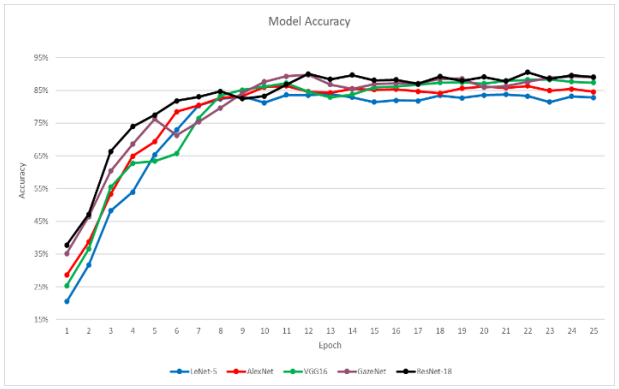

Figure 6. Accuracy of epoch for five different CNN architectures during training phase.

As reflected in Fig. 6, all the models reach less than 40% at the first epoch. The result consistently gets better as the epoch increase. When it comes to epoch 5 until 7, the models start to become more stagnant than before. The rest of the epochs' accuracy is decreasing and increasing dynamically but can still maintain to be stable. The graph shows that LeNet-5 model perform quite stable accuracy performance even though it has the lowest accuracy score. On the other hand, GazeNet results better accuracy but less stable than the other architecture.

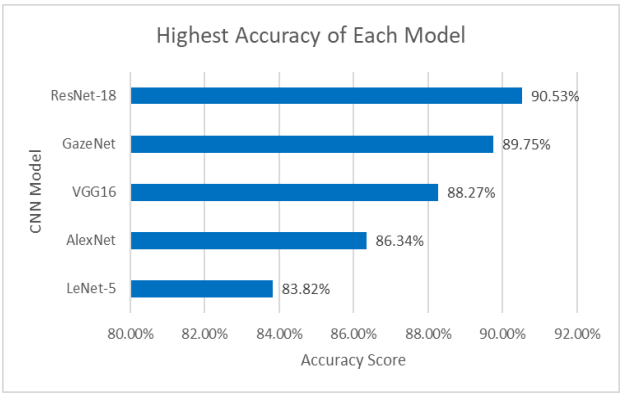

Figure 7. Maximum accuracy score performed by each model.

TABLE I. ACCURACY COMPARISON OF FIVE LATEST EPOCH

Epoch	*LeNet-5*	*AlexNet*	*VGG-16*	*GazeNet*	*ResNet-18*
21	83.76%	85.79%	87.94%	86.24%	87.75%
22	83.21%	86.34%	88.23%	87.59%	90.53%
23	81.43%	84.91%	88.27%	88.90%	88.53%
24	83.19%	85.43%	87.65%	89.25%	89.60%
25	82.77%	84.56%	87.34%	89.12%	89.03%
Avg	82.87%	85.41%	87.89%	88.22%	89.09%

Table I present accuracy of each model in the last five epochs while the graph shows maximum accuracy score from each model. Assuming the last five epochs represent the most stable condition between the whole phases, the average of these data is computed to find out stability of the model while searching for the highest score. The highest average score as pictured in Fig. 7 is performed by ResNet-18, which also gains the highest accuracy score (90.53% accuracy in epoch 22) of the whole experiment.

V. CONCLUSION

Eye tracking system has been implemented in many sectors and mostly use CNN algorithm as the main algorithm. Various CNN architecture has been researched and utilized to build a robust eye tracking system. The conclusions from the study are mentioned below:

- LeNet-5, AlexNet, VGG16, GazeNet, and ResNet-18 are some of CNN architectures and become the subject of the research
- All the architectures mentioned on the point above show high accuracy performance as the model for eye tracking systems which are above 83.5%
- ResNet-18 has the highest average accuracy compared to the other architectures
- The result is based on overall and last five epoch
- For the future goals, more variables are intended to be included such as activation function, learning rate, optimizer, etc

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Jennifer conceived the idea and participated in technical details. Joevian Krislynd developed the theory and performed the computational experiment. Steven Aprianto processed the experimental data and take a lead in writing the manuscript. Derwin Suhartono was advising and supervising the whole research. All authors had approved the final version.

REFERENCES

[1] A. George and A. Routray, "Real-Time eye gaze direction classification using convolutional neural network," in *Proc. International Conference on Signal Processing and Communications*, 2016, pp. 1-5.

[2] O. Heslinga. (March 25, 2021). Future trends part 1: Eye tracking. IMotions. [Online]. Available: https://imotions.com/blog/future-trends-part-1-eye-tracking/

[3] S. Hüttermann, B. Noël, and D. Memmert, "Eye tracking in high-performance sports: Evaluation of its application in expert athletes," *International Journal of Computer Science in Sport*, vol. 17, no. 2, pp. 182-203, 2018.

[4] A. Strzelecki, "Eye-Tracking studies of web search engines: A systematic literature review," *Information*, vol. 11, no. 6, p. 300, 2020.

[5] I. W. Skinner, M. Hübscher, G. L. Moseley, H. Lee, B. M. Wand, A. C. Traeger, and J. H. McAuley, "The reliability of eyetracking to assess attentional bias to threatening words in healthy individuals," *Behavior Research Methods*, vol. 50, no. 5, pp. 1778-1792, 2018.

[6] R. G. Lupu, F. Ungureanu, and V. Siriteanu, "Eye tracking mouse for human computer interaction," in *Proc. E-Health and Bioengineering Conference*, 2013, pp. 1-4.

[7] N. Carvalho, E. Laurent, N. Noiret, G. Chopard, E. Haffen, D. Bennabi, and P. Vandel, "Eye movement in unipolar and bipolar depression: A systematic review of the literature," *Frontiers in Psychology*, vol. 6, 2015.

[8] K. L. Burkhouse, G. J. Siegle, M. L. Woody, A. Y. Kudinova, and B. E. Gibb, "Pupillary reactivity to sad stimuli as a biomarker of depression risk: Evidence from a prospective study of children," *Journal of Abnormal Psychology*, vol. 124, no. 3, pp. 498-506. 2015.

[9] L. Bin and F. Hong, "Real time eye detector with cascaded convolutional neural networks," *Hindawi*, vol. 2018, article ID 1439312, 2018.

[10] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-Art of visualization for eye tracking data," *EuroVis – STARs*, 2014.

[11] N. Marcus, H. Ignace, and H. Kenneth, "Post-Saccadic oscillations in eye movement data recorded with pupil-based eye trackers reflect motion of the pupil inside the iris," *Vision Research*, vol. 92, pp. 59-66, 2013.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based learning applied to document recognition," *Proc. the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[15] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks," *Journal of Medical Systems*, vol. 44, no. 2, pp. 1-16, 2020.

[16] B. Mahanama, Y. Jayawardana, and S. Jayarathna, "Gaze-Net: Appearance-Based gaze estimation using capsule networks," in *Proc. the 11th Augmented Human International Conference*, May 2020, pp. 1-4.

[17] C. Lounis, V. Peysakhovich, and M. Causse, "Intelligent cockpit: Eye tracking integration to enhance the pilot-aircraft interaction," in *Proc. the ACM Symposium on Eye Tracking Research & Applications*, June 2018, pp. 1-3.

[18] N. M. Gunaratne, S. Fuentes, T. M. Gunaratne, D. D. Torrico, H. Ashman, C. Francis, C. C. Viejo, and F. R. Dunshea, "Consumer acceptability, eye fixation, and physiological responses: A study of novel and familiar chocolate packaging designs using eye-tracking devices," *Foods*, vol. 8, no. 7, p. 253, 2019.

[19] L. Mi, C. Lei, Z. Qian, L. Peng, L. Sa, L. Richeng, F. Lei, W. Gang, H. Bin, and L. Shengfu, "Method of depression classification based on behavioral and physiological signals of eye movement," *Complexity*, vol. 2020, article ID 4174857, 2020.

[20] H. Wang, Y. Zhou, F. Yu, L. Zhao, C. Wang, and Y. Ren, "Fusional recognition for depressive tendency with multi-modal feature," *IEEE Access*, vol. 7, pp. 38702-38713, 2019.

[21] A. K. Nsaif, S. H. M. Ali, K. N. Jassim, A. K. Nseaf, R. Sulaiman, A. Al-Qaraghuli, O. Wahdan, and N. A. Nayan, "FRCNN-GNB: Cascade faster R-CNN with Gabor filters and naïve bayes for enhanced eye detection," *IEEE Access*, vol. 9, pp. 15708-15719, 2021.

[22] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4511-4520.

[23] A. M. Feit, S. Williams, A. Toledo, A. Paradiso, H. Kulkarni, S. Kane, and M. R. Morris, "Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design," in *Proc. the Chi Conference on Human Factors in Computing Systems*, May 2017, pp. 1118-1130.
[24] J. Johnsson and R. Matos, *Accuracy and Precision Test Method for Remote Eye Trackers*, Sweden: Tobii Technology, 2011.

Jennifer was born in Jakarta on 22nd August 2001. She graduated from Sang Timur High School and currently she is a student in Bina Nusantara University majoring in Computer Science at BINUS Kemanggisan, Jakarta, Indonesia.

Joevian Krislynd was born in Jakarta on 15th February 2001. He graduated from Budi Mulia High School and currently he is a student in Bina Nusantara University majoring in Computer Science at BINUS Kemanggisan, Jakarta, Indonesia.

Steven Aprianto was born in Tangerang on 14th April 2001. He graduated from Perguruan Buddhi High School and currently he is a student in Bina Nusantara University majoring in Computer Science at BINUS Kemanggisan, Jakarta, Indonesia.

Derwin Suhartono is faculty member of Bina Nusantara University, Indonesia. He got his PhD degree in computer science from Universitas Indonesia in 2018. His research fields are natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals.