Precision Journal of Information Technology & Scientific Innovation
Volume 1, 2026, **pp.** 1-16, **Article ID:** PD2601102003
DOI: inprocess
OPEN ACCESS

Research Article

Real-Time Eye Movement Tracking and Prediction Using Machine Learning for Vision-Based Applications

Hadia Tazeem[1, *]

[1]*School of Information Technology, Deakin University, Melbourne, Australia*

Article History

Received: 25 September, 2025 Revised: 08 December, 2025 Accepted: 23 December, 2025 Published: 04 April, 2026

Abstract:

Introduction: Real-time gaze tracking is essential for human–computer interaction, automotive safety, assistive systems, and medical diagnostics. Traditional hardware-based trackers are costly and intrusive, while many machine learning (ML) models are too computationally demanding for real-time use.

Methods: This study develops and evaluates a lightweight deep learning framework for gaze direction prediction using three architectures: a baseline Convolutional Neural Network (CNN), a single-eye MobileNetV2, and a dual-eye MobileNetV2. Performance was assessed using accuracy, precision, recall, F1-score, AUC, latency, and frames per second (FPS) to determine deployment feasibility.

Results: This show that the CNN achieved 73.23% accuracy, an AUC of 0.81, and a latency of 0.56 ms per sample. The single-eye MobileNetV2 model reached 75.02% accuracy, with an AUC of 0.80, high recall for "Not Looking Ahead" (0.90), and low latency of 1.30 ms per sample, achieving a throughput of 772 FPS. The dual-eye MobileNetV2 model achieved the highest accuracy (86.35%) and AUC (0.93), with excellent precision and recall for both classes, but suffered from higher latency (5615 ms per sample) and a throughput of 0.18 FPS. The single-eye MobileNetV2 thus offers the best balance between predictive performance and real-time operation.

Conclusion: The findings demonstrate that lightweight and interpretable ML models can deliver accurate, low-latency gaze tracking suitable for AR/VR systems, driver monitoring, and healthcare applications.

Keywords: Eye tracking, gaze prediction, machine learning, real-time systems, mobile-net, vision applications.

1. INTRODUCTION

Human perception centres around eye movements, demonstrating the focus of attention and cognition. Real-time gaze monitoring offers a strong window into user intention, situational awareness, and decision-making. Gaze tracking can also help prevent accidents in safety-critical areas, such as automotive systems, by monitoring drivers' alertness. Meanwhile, healthcare provides an avenue to evaluate cognitive burden, early neuropsychiatric disorders or rehabilitation [1]. Gaze is a natural form of intuitive control in human-computer interaction (HCI) that minimises human-to-digital interface interaction [2]. Eye-tracking research has become popular across computer vision, neuroscience, and machine learning. Recent technological advances have brought to the fore the significance of gaze in immersive media, such as augmented reality (AR) and virtual reality (VR). Gaze leads to better interaction and optimises the design of the rendering and interface in these settings, offering efficient, low-latency input systems [2]. Research indicates that AR and VR technologies that utilise gaze information may provide interactive and personalised entertainment, education, and industrial experiences [3, 4]. Gaze-based driver monitoring can predict attention lapses and distractions, and build safer intelligent transport systems [5]. Gaze tracking has also found its way into healthcare, with wearables connected to machine learning serving as a source of real-time monitoring, diagnostics, and tasks that support rehabilitation [6]. Meanwhile, assistive technologies enable individuals with mobility impairments to control equipment or communicate through gaze, underscoring their social and accessibility benefits [7].

*Address correspondence to this author at School of Information Technology, Deakin University, Australia; E-mail: hadia.tazeem@gmail.com

Although these opportunities exist, solutions for gaze tracking currently face challenges with scalability and implementation. Hardware-based trackers, such as infrared or optical trackers, are expensive, obtrusive, and portable, making them unsuitable for many users, including perioperative theatres. This limits their use in consumer solutions like mobile AR/VR or in-vehicle systems [8]. Conversely, software-based computer vision models are cheaper but often lack robustness to changes in illumination, occlusion, or head pose. Moreover, most deep learning architectures that aim to mitigate these constraints are computationally intensive and exhibit high latency and low frame-per-second (FPS) performance. In real-time use, tens of milliseconds of latency is catastrophic for user experience, safety, and deployment trust [9].

1.1. Objective of Study

To fill this gap, three objectives are established in this study.

1. The study creates a lightweight machine learning pipeline that can run in real time, track gaze, and achieve minimal latency, while optimised for resource-constrained systems (AR/VR headsets and embedded automotive systems).
2. The study contrasts three model architectures: a basic convolutional neural network (CNN), a single-eye tracker based on MobileNetV2 and a dual-input MobileNetV2 that uses both eyes. This comparison balances accuracy and efficiency, recognising trade-offs between computation and predictive reliability.
3. The study interprets its results using Gradient-weighted Class Activation Mapping (Grad-CAM) rather than accuracy measures. It is essential to provide visual explanations of model predictions to build trust and acceptance in healthcare or driver monitoring contexts, where black-box systems raise concerns about accountability and bias.

The contribution of this work is threefold. In real-time gaze prediction, the study offers the first organised comparison between CNNs, MobileNet, and dual-eye MobileNet configurations. The study analysed traditional performance measures (accuracy and ROC-AUC) and deployment-relevant measures (latency and FPS), providing a comprehensive picture of the model's appropriateness for deployment. Furthermore, the study performs an explainability analysis using Grad-CAM to identify the areas that affect gaze prediction, thereby enabling informed decision-making. The synergistic blend of benchmarking, performance assessment, and explainability is intended to bridge the gap between research prototypes and actual implementation. The rest of the paper is organised as follows. Section 2 outlines the associated literature, including conventional hardware trackers, computer vision-based methods, and deep learning methods, and identifies the remaining gaps. Section 3 describes the methodology, including dataset characteristics, preprocessing, and architectures. Section 4 contains findings, such as quantitative and interpretability results and latency measurements. Section 5 discusses implications, limitations, and potential future research directions. Lastly, Section 6 concludes by outlining the contributions and the overall significance of lightweight, interpretable gaze tracking for machine vision implementation.

This study tests the hypothesis that lightweight CNN-based architectures, particularly MobileNetV2, can achieve gaze-tracking accuracy comparable to heavier deep learning models while maintaining real-time performance suitable for embedded and edge systems. Accordingly, the research aims to evaluate how model design influences the trade-off between predictive accuracy, latency, and interpretability in vision-based human–computer interaction applications.

2. LITERATURE REVIEW

2.1. Traditional Eye Tracking

Conventional eye-tracking techniques, such as infrared (IR) illumination systems, electroencephalography (EEG), and optical trackers, have traditionally relied on hardware. These systems scan eye characteristics, such as pupil position, corneal reflections, and glints, to determine gaze direction. IR-based systems are expensive and require specialised equipment, though they are constrained by the need for a controlled environment and high precision [10]. EEG-based systems can provide more information about the neural mechanisms underlying visual attention, but they are invasive, and the electrode arrangement is unsuitable for daily use. Recently, smartphone prototypes equipped with infrared illumination and convolutional neural networks (CNNs) have been proposed as cheaper alternatives, achieving higher accuracy than conventional natural-illumination systems [10]. Nevertheless, these solutions are still expensive to produce, cannot be scaled for large-scale implementation, and are frequently bulky in mobile or wearable applications, which narrows their use.

2.2. Computer Vision-Based Methods

Some scalability issues have been overcome by the advent of computer vision (CV)-based eye-tracking methods that use ordinary cameras without specialised hardware. Machine learning classifiers, including Haar cascades, Support Vector Machines (SVMs), and Random Forests, were early used to identify eye features and categorise gaze direction [11]. Although these approaches minimised the use of specialised sensors, they were prone to changes in illumination, head pose, and occlusion, which limited their robustness in uncontrolled settings. Solutions to such challenges include introducing object detection models, such as YOLO v4, into mobile eye-tracking analysis, enabling automatic labelling of gaze data in natural settings with greater precision and less manual work [12]. Although these advances have been made, CV-based algorithms are still struggling because they rely on handcrafted features that do not generalise well to real-world variability.

2.3. Deep Learning Approaches

Deep learning has changed gaze estimation by enabling end-to-end feature learning, eliminating the need for handcrafted features. CNNs have been the main approach to gaze classification systems that can learn spatial information from eye images under various conditions [13]. For example, CNN-based pupil detectors are highly resistant to noise and

blinking artefacts, providing more accurate gaze estimation with wearable devices [14]. Backbones trained with methods such as ResNet, VGG, and MobileNet are more efficient and precise thanks to transfer learning. They are therefore applicable to mobile devices and embedded systems [15]. MobileNet has gained widespread use because of its light architecture, which can achieve decent accuracy at minimal computational and memory costs.

The temporal modelling has also become popular, and recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) have been incorporated with CNNs to learn sequential gaze dynamics. The architectures enable models to consider saccades and smooth pursuits to enhance prediction accuracy in real-life tasks [16]. Equally, hybrid methods utilising CNNs with attention mechanisms have shown promise for managing variability in head pose and illumination [17]. These developments emphasise the capability of deep learning models to generalise to datasets and interaction conditions, but at high computational cost in many cases.

Recently, Transformer-based architectures have emerged as strong alternatives to convolutional models for gaze estimation and broader vision tasks. Vision Transformers (ViT) and Swin Transformers model long-range spatial dependencies through self-attention, enabling superior contextual reasoning about eye shape, iris position, and facial geometry compared to purely convolutional filters. Studies have demonstrated that Transformer encoders can outperform conventional CNNs in gaze tracking and attention estimation by effectively integrating global visual cues while remaining computationally efficient through hybrid CNN-Transformer pipelines [18]. In addition, the Anomaly Transformer framework has recently been adapted for visual time-series prediction and gaze-motion anomaly detection, providing enhanced temporal awareness and robustness under illumination or head-pose variations [19]. These developments highlight the growing trend toward attention-driven, transformer-based gaze estimation, which combines interpretability with scalable performance across diverse vision applications.

2.4. Real-Time Challenges

Even though they are accurate, deep learning-based systems exhibit specific issues when deployed to the real-time domain. One key bottleneck is the high computational cost, as the intricate CNN and RNN architectures require substantial processing power, resulting in latency issues. Experiments evaluating edge computing for eye-tracking show that cloud-based inference can be rapid, but communication latency prevents real-time applications. Conversely, in-device inference is limited by memory and energy [15]. There are also problems with datasets: gaze datasets are usually unbalanced across gaze classes, making it hard for models to generalise [20]. Annotation noise also makes training more complex, as errors in gaze labelling reduce the model's reliability. These restrictions limit the use of eye-tracking solutions in latency-critical applications, including driver monitoring, AR/VR interaction, and assistive technologies.

2.5. Research Gaps

Although deep learning has advanced gaze tracking to a high level, significant gaps remain. Compared to other models, lightweight, real-time models that can work on mobile or embedded platforms without compromising accuracy are lacking. More recent systems, such as smartphone-optimised CNNs, are showing promise but remain unable to trade off accuracy and latency under unconstrained conditions [21]. Moreover, little has been done to investigate the interpretability of gaze-tracking models. As transparency and confidence in AI systems are questioned, explainable methods (such as Grad-CAM) have not been widely used in gaze prediction studies. There is little research that carefully examines the area(s) of the eye or the face that lead to model predictions, making their use somewhat unacceptable in risky applications like healthcare or safety system development. To address these research gaps, it is necessary to develop lightweight, scalable models and incorporate interpretability frameworks to enable responsible deployment.

3. METHODOLOGY

3.1. Framework Overview

The suggested real-time gaze-tracking and prediction system given in Fig. (**1**), is based on an organised flow, ensuring computational efficiency and accuracy. The pipeline takes the raw gaze data as input, followed by preprocessing to improve data quality and consistency. Image preprocessing is then sent to one of the three model structures investigated in this paper: a baseline Convolutional Neural Network (CNN), a single-input MobileNetV2, and a dual-input MobileNetV2 that takes the right and left eyes as inputs. After feature extraction and classification, the model's gaze-direction predictions are also validated using various evaluation metrics. Techniques in visualisation, such as Grad-Cam, were also incorporated to enable interpretation and build confidence in the model, which could be used to make real-world decisions, since the methods identified parts of the eye that were critical to the model's decision. The most important aspects of this end-to-end pipeline are lightweight performance, accuracy, and explainability, which should be deployed in vision-based applications such as driver monitoring and augmented reality systems.

3.2. Dataset

The experiments were conducted on the Gaze Direction Detection dataset, which consists of paired images of the right and left eyes with binary labels indicating whether the subject is looking forward (Class 1) or not (Class 0) [22] as shown in Fig. (**2**). The initial data set comprised 53,000 samples, each consisting of an image of the left eye, a photo of the right eye, and a picture of the target. The population was skewed, with the majority of samples belonging to the "Not Looking Ahead" group. Specifically:

- Class 0 (Not Looking Ahead): 36,111 samples
- Class 1 (Looking Ahead): 16,889 samples

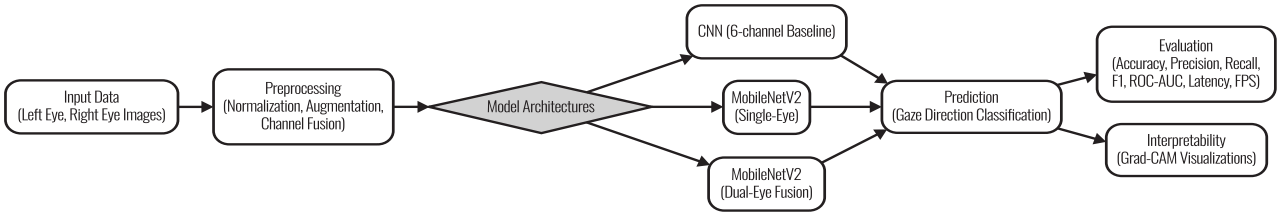

Fig. (1). Methodology framework.

Data Cleaning and Filtering:

A preliminary exploratory analysis was conducted to assess data integrity, and no similar or entirely black images were found. Nevertheless, the quality of the samples varied significantly, especially in sharpness and texture detail. To overcome such problems, Laplace variance and edge density measures were used to evaluate the quality of images. Consequently, 11,287 poor-quality images (left- and right-eye images) were removed from the dataset. The dataset was trimmed down to 41,713 usable pairs of images (right-eye and left-eye images) after cleaning with the following distribution of classes:

- Class 0 (Not Looking Ahead): 27,824 samples
- Class 1 (Looking Ahead): 13,889 samples

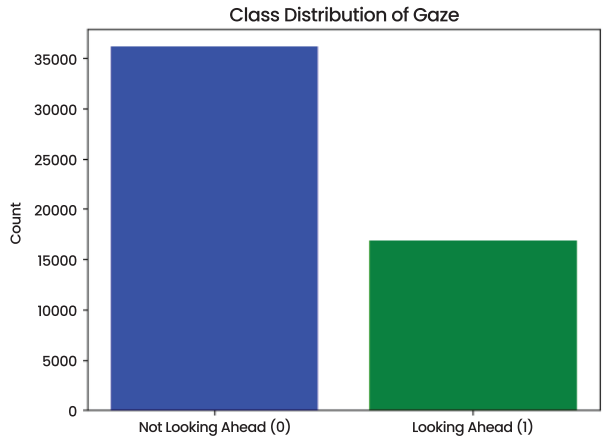

Fig. (2). Class distribution.

Balancing the Dataset:

Downsampling was used to balance the dataset and address class imbalance, ensuring equal representation of both classes. The resulting balanced dataset contained 27,064 samples, with 13,533 per class. This was balanced by eliminating the redundant samples in the Not Looking Ahead category, which gives the following balanced data:

- Class 0 (Not Looking Ahead): 13,533 samples
- Class 1 (Looking Ahead): 13,533 samples

Dataset Splits:

The balanced sample was further split into training, validation, and test samples in the subject-exclusive mode, ensuring that each subject's data did not appear in more than one sample to prevent identity leakage. The data was divided in the following way:

- Training Set: 18,946 samples (70%)
- Validation Set: 4,060 samples (15%)
- Test Set: 4,060 samples (15%)

The dataset was balanced across classes (13,533 samples per class). The splits were randomly selected within the subject-exclusive constraint to avoid potential data leakage between sets.

Data Augmentation:

Since real-world eye-tracking systems are variable (*i.e.,* illumination changes, head position, partial occlusion (with glasses or reflections)), several augmentation strategies were introduced during training. These included:

- Random brightness scaling
- Horizontal flips
- Random rotations (±15°)
- Contrast jittering

The objective of these augmentation methods was to subject the model to a range of visual tests and make it more resilient to natural variability. Irrespective of these precautions, domain shifts, such as controlled *versus* natural settings, remain an issue. Thus, the next task will be to investigate domain-adaptation and transfer-learning methods that exploit cross-dataset characteristics (*e.g.,* MPIIGaze to GazeCapture) to improve model accuracy across different conditions. Also, research on multi-domain data and adversarial adaptation frameworks will further enhance the model's resilience for embedded and mobile applications.

3.3. Preprocessing

Data preprocessing was used to prepare the eye images for deep learning models. The input images were scaled to the (0,1) pixel intensity range to standardise dynamic ranges and minimise sensitivity to lighting conditions. The training used augmentation strategies, such as random rotations, horizontal flips, and brightness changes, to enhance generalisation and resistance to environmental change. This ensured the models could adapt to different gaze directions and illumination conditions experienced in real-life applications.

For the CNN baseline, the right and left eye images were combined into a six-channel input by concatenating the respective RGB channels. This six-channel fusion enabled the CNN to learn joint spatial and contextual patterns across the two eyes. For the MobileNet-based architectures, preprocessing involved generating three-channel inputs for the left and right eyes. Such separation enabled the dual-input MobileNetV2 to be trained on eye-specific features and then to fuse features at the feature level. Meanwhile, the single-input version used only right-eye images to mimic lightweight deployment cases.

To further enhance cross-domain robustness, augmentation strategies were designed to mimic real-world variations, including head-pose shifts, glare, and partial occlusions caused by eyewear. Future improvements will explore domain-adaptation and transfer-learning methods that allow the model to maintain consistent accuracy across diverse environments, lighting conditions, and user profiles.

3.4. Model Architectures

Three architectures were investigated to assess the trade-off among model complexity, interpretability, and efficiency. The former was a six-channel base CNN that required six-channel inputs. It consisted of several convolutional layers with ReLU activation, separated by max pooling and batch normalisation to stabilise training. The feature maps were compiled using global average pooling, and dense layers were used for classification. Despite being relatively shallow, with about 111k trainable parameters, this architecture provided a point of comparison for computational efficiency.

The second network was a single-input MobileNetV2. Initialising the convolutional backbone with ImageNet-pretrained weights was used as transfer learning. The model took three-channel right-eye images as input, which were processed using depthwise separable convolutions, as in MobileNetV2. This design significantly reduced the number of parameters while still allowing the extraction of hierarchical features. The last layers were fine-tuned to train the model for the gaze prediction task without compromising MobileNet's efficiency.

The third and most advanced configuration was a dual-input MobileNetv2 configuration, in which the right and left eyes were processed independently by identical MobileNet backbones. Embeddings were extracted, concatenated, and fed to dense layers to produce the final prediction. This technique enabled this network to obtain eye-specific features and to leverage cross-eye fusion. This model was more computationally expensive but more functional, particularly in selecting fine asymmetries in the gaze direction between the two eyes.

3.5. Evaluation Metrics

Some of the performance metrics used in this research to compare gaze-tracking models are based on classification accuracy and real-time deployment efficiency. The metrics will play a critical role in determining the models' ability to correctly classify gaze directions in real-time applications such as driver monitoring systems and augmented reality (AR) systems. The most common performance measure is accuracy, which assesses the effectiveness of the classification process by determining the fraction of correct predictions among the total number of prediction samples. Precision is the ratio of true positives to all optimistic predictions, and it measures how well the model avoids false positives. Recall or sensitivity): this metric of the model indicates its capacity to detect all cases of interest by computing the ratio of true positives to the actual positives. F1-score provides a balanced measure by combining precision and recall, particularly when class imbalance is a problem. Area Under the Curve (AUC) is used to measure a model's ability to classify positive and negative classes; the higher the AUC, the more powerful the model. Precision-Recall AUC (PR-AUC) is applicable when data is imbalanced, as it assesses the trade-off between precision and recall. Also, the Brier score assesses the accuracy of probabilistic predictions; the lower the score, the better the calibration. False Positive Rates of 1 per cent and 5 per cent provide information on the model's performance within specific false positive rate ranges.

In real-time measurements, latency is the average time the model spends processing one sample, which is important when speed is needed, as in real-time applications. Frames Per Second (FPS) is used to determine the number of samples the model can be processed within a second and is an indication of the real time performance. To analyze the models, confusion matrices will be drawn to indicate the true positives, false positives, true negatives, and false negatives and ROC curves will be plotted along with PR curves to determine the discriminative strength and the precision-recall ratio of the model. Calibration curve is used to compare the predicted and actual outcomes to assess the calibration of the model. To statistically verify it, bootstrap confidence intervals of AUC determine AUC robustness, DeLong test compares values of AUC, and test McNemar compares errors of a classifier. Lastly, Grad-CAM heatmaps enhance the interpretability of models to identify which parts of the eye images are the most predictive.

3.6. Reproducibility & Training Protocol

This study carried out its experiments through the Kaggle cloud platform, with a single NVIDIA P100 graphics card with 16 GB of GPU memory. The available RAM was 29 GB and the disk space was 57.6 GB. The inference load percentage was approximately 150% which implied that the CPU was heavily loaded when running the gaze-tracking tasks. TensorFlow 2.13, Keras 3.0 and Python 3.11 were used to develop the models, which offers a solid platform to perform machine learning and deep learning procedures. As far as the training configuration is concerned, the batch size was configured to 1 (one sample at a time) to resemble the real-time analysis. The images had a resolution input of 224x224 pixels. Adam optimizer was applied with learning rate 1e-4 and models were trained with 50 epochs and early stopping and ReduceLROnPlateau was used in order to avoid overfitting. To guarantee reproducibility, all the experiments were performed with the same random seed,

so the results could be reproduced. The data was prepared well in order to have a clean and quality data to train and evaluate on. The training and evaluation operations are well elaborated so that anyone having access to the same platform, hardware, and code can recreate the findings. This study is reproducible by recording the system specifications and the very environment that was used, as a result the results can be applied to other similar hardware systems.

4. RESULTS & ANALYSIS

4.1. Convolutional Neural Network (CNN)

This part shows the performance analysis of the Convolutional Neural Network (CNN) model by different measurements, visualizations, and statistical tests. Two categories of the model, *i.e.,* looking ahead and not looking ahead are evaluated using various diagnostic tools. The Confusion Matrix (Fig. **3**) offers the understanding of how the model works in predicting the actual labels. The CNN model was able to find the correct samples of 1479 cases of not looking ahead and 1494 cases of looking ahead out of 4060 total samples. Nonetheless it had some misclassifications with 551 samples being incorrectly predicted as "Looking Ahead" as opposed to Not Looking Ahead, and 536 samples being incorrectly predicted as Not Looking Ahead as opposed to Looking Ahead. These misclassifications are not unusual to any machine learning model and are important to the learning of the limitations of the model.

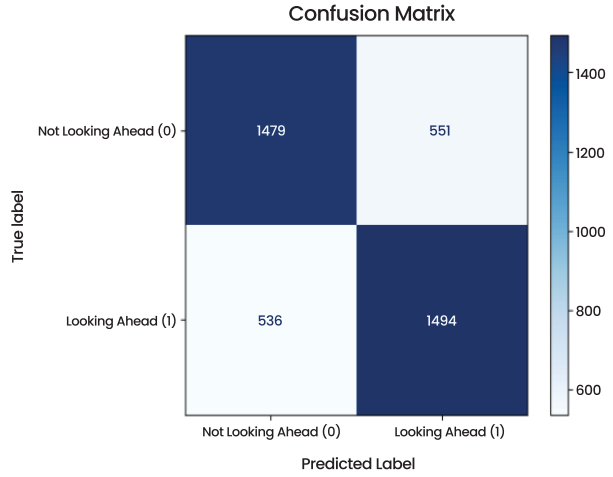

Fig. (3). Confusion matrix of CNN model.

The Receiver Operating Characteristic (ROC) Curve (Fig. **4**) shows that it performs very well with an Area Under the Curve (AUC) of 0.81. The shape of the curve indicates that the model is effective in terms of differentiating between the two classes and has high true positive rate (sensitivity) and low false positive rate, implying that the model has strong classification capabilities.

The Precision-Recall Curve (Fig. **5**) reveals that the model has an average precision (AP) of 0.80 indicating a strong capability to access relevant examples of both classes especially in a situation where there is class imbalance issue. Precision-recall curve also shows that the model is precise and recalls well, which can be a highly important component in the assessment of classifiers in imbalanced data.

Concerning the calibration, the Calibration Curve (Fig. **6**) shows that the model prediction of the probabilities is in good agreement with the real result, and the probability distribution of the model adheres to the graph of the perfect calibration. It implies that the probabilities given by this model are accurate and there is no serious biasness of the model with any of the classes.

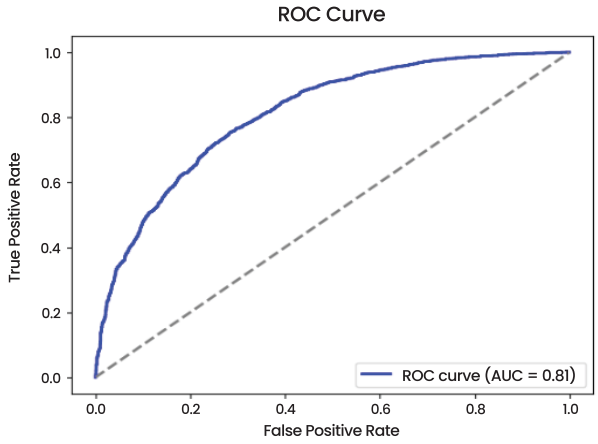

Fig. (4). Receiver operating characteristic (ROC) curve of CNN model.

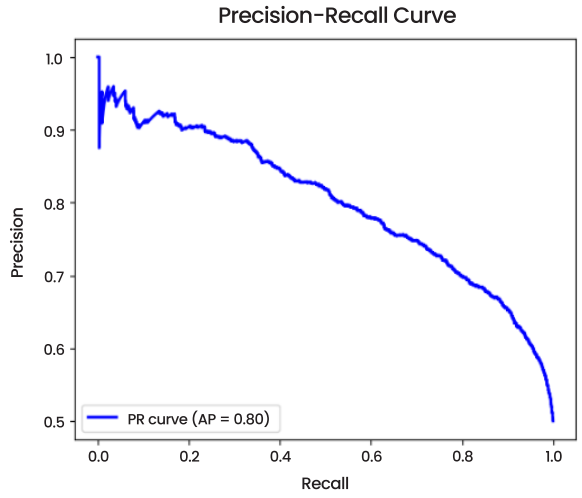

Fig. (5). Precision-recall curve of CNN model.

The Loss Over Epochs (Fig. **7**) curve indicates that training loss reduces substantially, which is one of the signs of successful learning. The validation loss, however, is rather fluctuating, which could indicate overfitting, because it deviates at a point once the number of epochs has reached a specific point. This deviation is a significant factor in model

optimization, which suggests that regularization methods or additional fine-tuning is necessary to avoid overfitting.

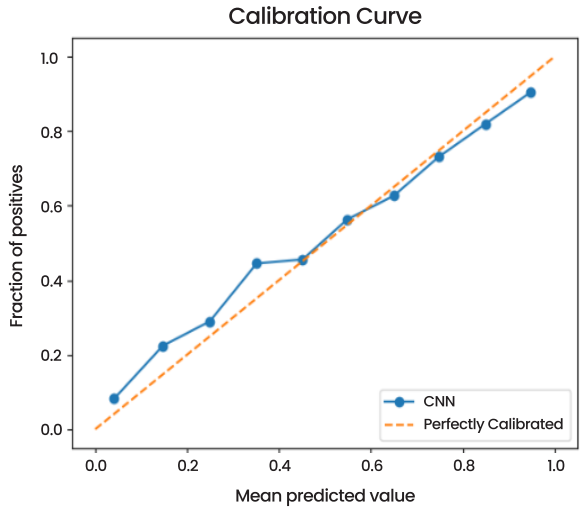

Fig. (6). Calibration curve of CNN model.

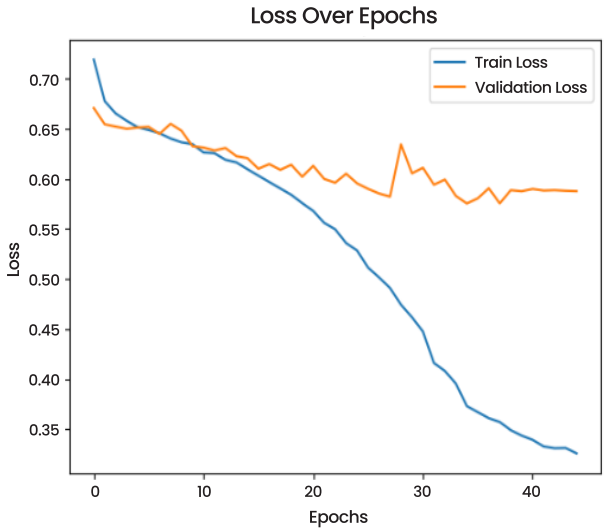

Fig. (7). Loss over epochs of CNN model.

In the Accuracy Over Epochs (Fig. **8**), the training and validation accuracy are seen to improve consistently. The accuracy of the training becomes gradually increasing whereas validation accuracy also increases but with slower pace, which makes it possible to suggest that the model has the generalizing potential but it can still be improved to decrease the difference between training and validation performance.

Model Metrics:

- **Precision and Recall**: The model achieved precision and recall scores of 0.73 for both "Looking Ahead" and "Not Looking Ahead" classes.
- **Test Loss and Accuracy**: The test loss was 0.5363, with a corresponding test accuracy of 0.7323.
- **AUC**: The model's AUC is 0.81, with a 95% confidence interval of [0.7872, 0.8129], indicating strong discrimination power.
- **Recall@FPR ≤ 1%**: The model maintained an excellent recall rate of 0.9990 at a false positive rate (FPR) of ≤1%.
- **McNemar's Test**: The p-value of McNemar's test was 1.0000, indicating no significant difference between the model's performance and the baseline.
- **Latency:** The CNN model achieved a latency of 0.56 ms per sample.
- **Throughput:** The model demonstrated a throughput of 1769.94 FPS.

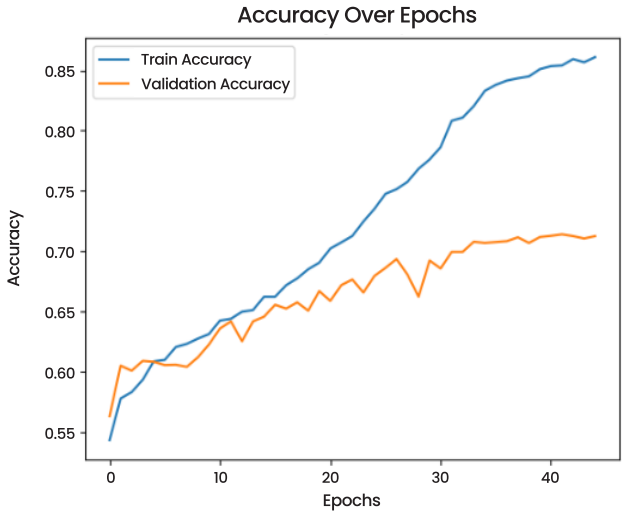

Fig. (8). Accuracy over epochs of CNN models.

To summarize, CNN model has a good performance in the classification of the two target categories with high accuracy, recall and precision. The evaluation measures have proven its usefulness and especially when distinguishing the classes with a strong AUC value and high recall rates yet low false positive rates. Although there are instances of overfitting as evidenced by the variability of the validation loss, the model has a high generalization capacity, which means that as the tuning and optimization process continues, the model can be improved. These findings form a good basis of using this model into practical use where binary classification is assured.

4.2. MobileNetV2 Model (Single Eye)

The MobileNetV2 model evaluation on the task of looking a step ahead *vs.* not. There are some critical observations that can be made concerning its performance in "Not Looking Ahead'. Fig. (**9**) below that is the Confusion Matrix indicating distributions of true and predicted labels of the model. The model had 7951 samples with a total of 4850 Not Looking Ahead and 1115 Looking Ahead. Nevertheless, the model

incorrectly classified 567 samples of Not Looking Ahead into the Looking Ahead category and 1419 samples of the Looking Ahead category into the Not Looking Ahead category. These misclassifications also reflect in the overall performance of the model which is also investigated using other metrics.

Fig. (**10**) shows that AUC of the model is 0.80 which is good performance in distinguishing between two categories. The curve indicates that the model has a high true positive rate with a relatively low false positive rate that is good in its predictive reliability.

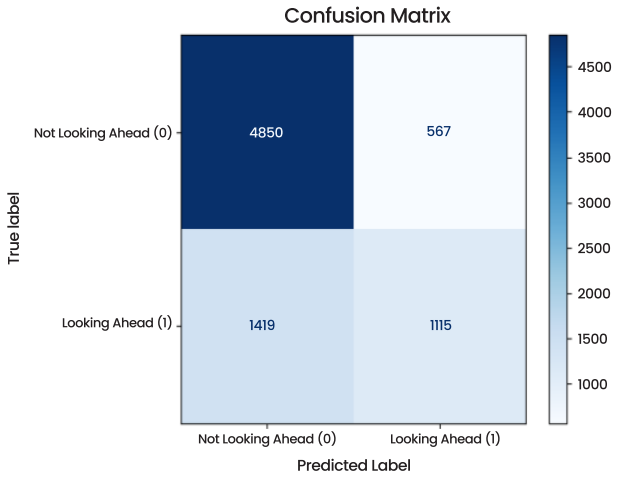

Fig. (9). Confusion matrix.

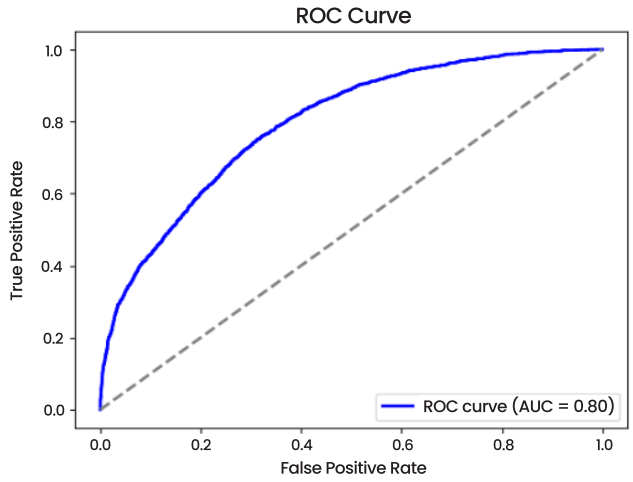

Fig. (10). ROC curve.

Precision-Recall Curve (Fig. **11**) scores an average precision (AP) of 0.66 which indicates that the model has some problem with preserving high precision especially the "Looking Ahead" class. This imprecision lays out areas of work to be done, in particular where the class imbalance may exist.

The Calibration Curve (Fig. **12**) indicates that the predicted probabilities of the model are reasonably well-calibrated. The calibration curve reveals that the predictions are very close to the perfectly calibrated line, which means that there is a strong predictive power of the model in terms of probabilities.

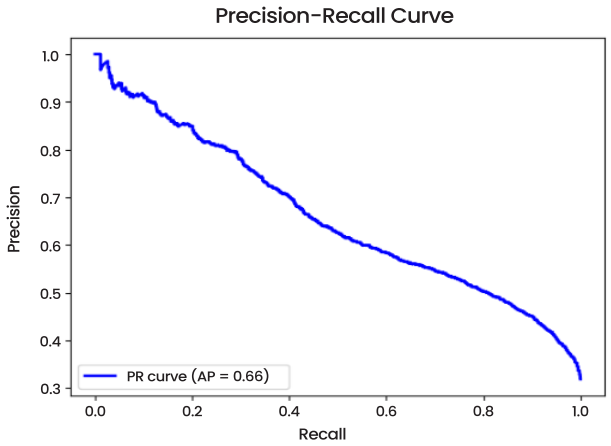

Fig. (11). Precision-recall curve.

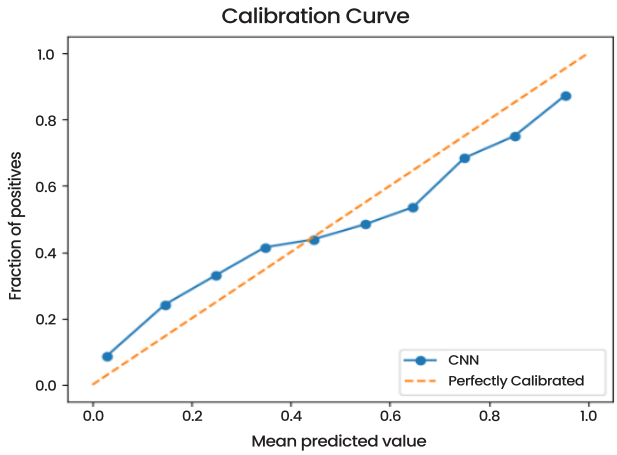

Fig. (12). Calibration curve.

The Loss-Over-Epochs (Fig. **13**) demonstrates that the loss during training was reducing linearly whereas that during validation varied considerably. This points to the possibility of overfitting since the model also did well on the training data than the validation set.

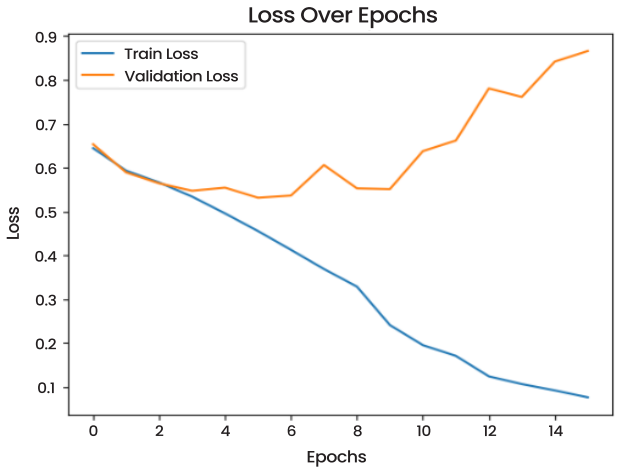

Fig. (13). Loss over epochs.

Accuracy Over Epochs (Fig. **14**) indicates that the training accuracy increased continuously, whereas validation accuracy reached the value of 0.80, which means that the model generalizes quite well and that it can be enhanced through additional tuning to be more robust.

Fig. (**15**) demonstrates a Grad-CAM heatmap and the original eye image. Grad-CAM (Gradient-weighted Class Activation Mapping) shows what the model uses in the image to make its predictions. Regions that have more importance in terms of significance to classification are usually represented in the heatmap, but in this example, the heatmap is not easily seen implying a lack of interpretability or interest.

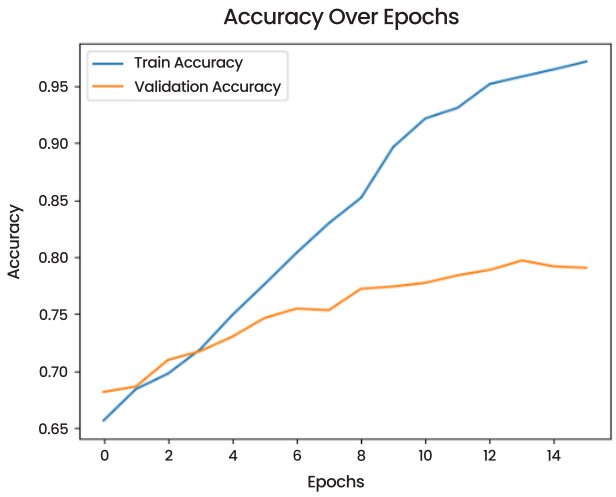

Fig. (14). Accuracy over epochs.

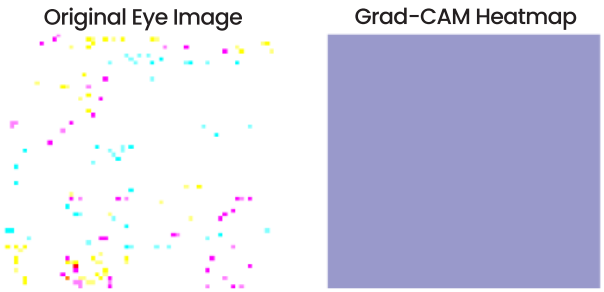

Fig. (15). Grad-CAM heatmap.

Model Metrics:

- **Precision and Recall**: Precision for "Not Looking Ahead" was 0.77, and recall was 0.90. For "Looking Ahead," precision was 0.66, and recall was 0.44.
- **Test Loss and Accuracy**: The test loss was 0.5260, and the test accuracy was 0.7502.
- **Recall at FPR ≤ 1%**: The model achieved an outstanding recall of 0.9996 at a false positive rate (FPR) ≤1%.
- **Brier Score**: 0.1690, indicating relatively good calibration of probability outputs.
- **Latency:** The MobileNetV2 (Single Eye) model achieved a latency of 1.30 ms per sample.
- **Throughput:** The model demonstrated a throughput of 772.13 FPS.

Finally, the MobileNetV2 model has a good classification accuracy of 75% but has difficulties with the accuracy of the "Looking Ahead" classification, which is possibly an indication of the unbalanced data or requires additional adjustment. However, it has good calibration and high recall which is encouraging particularly when there is high need of recall.

4.3. MobileNetV2 Model (Dual Eye)

The MobileNetV2 model (Dual Eye) shows a high level of performance in a range of metrics, which suggests that it is a robust model to use in the case of the "Looking Ahead" *vs.* Not Looking Ahead classification task. The Confusion Matrix (Fig. **16**) brings out the prediction distribution. The model was correct on 3935 instances of "Not Looking Ahead" and 1468 instances of "Looking Ahead" with 292 and 562 instances of misclassification, respectively, in the total sample of 6257. The distribution indicates that the model is highly biased to the correct classification of "Not Looking Ahead.

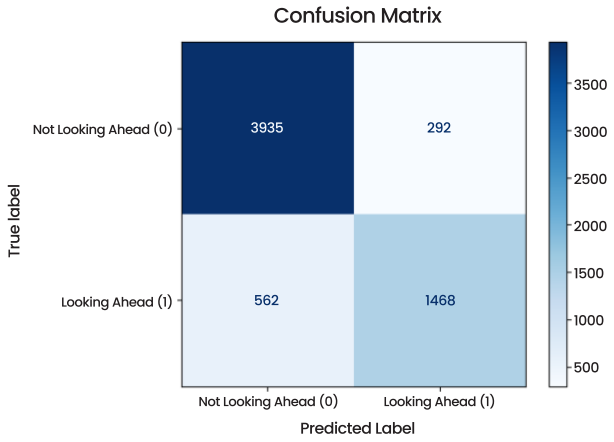

Fig. (16). Confusion matrix.

Fig. (**17**), the ROC Curve, demonstrates an impressive Area Under the Curve (AUC) value of 0.93 which correlates to excellent performance of the ROC in the differentiation of the two categories. The steep upward curve of the ROC indicates the high true positive rate and low false positive rate of the model that accentuate the fact that the model predicts correctly.

Precision-Recall Curve (Fig. **18**) indicates the average precision (AP) of 0.87 with a higher preciseness of the model in discovering true positives and more so in the "Looking Ahead" class. Although the recall slightly decreased, the model

still achieves a high precision recall ratio, which is essential in case of imbalanced datasets.

According to the Calibration Curve (Fig. **19**), the model can be used to estimate probabilities that are pretty close to real-life outcomes. The curve is much near the perfectly calibrated line, which indicates that the probability estimates of the model are accurate and valid.

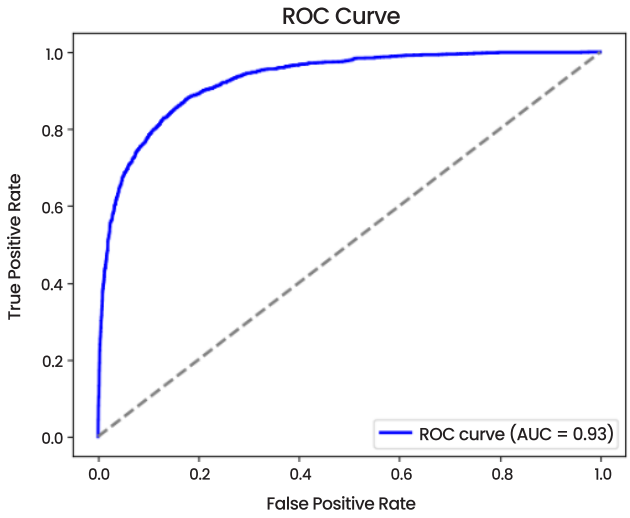

Fig. (17). ROC curve.

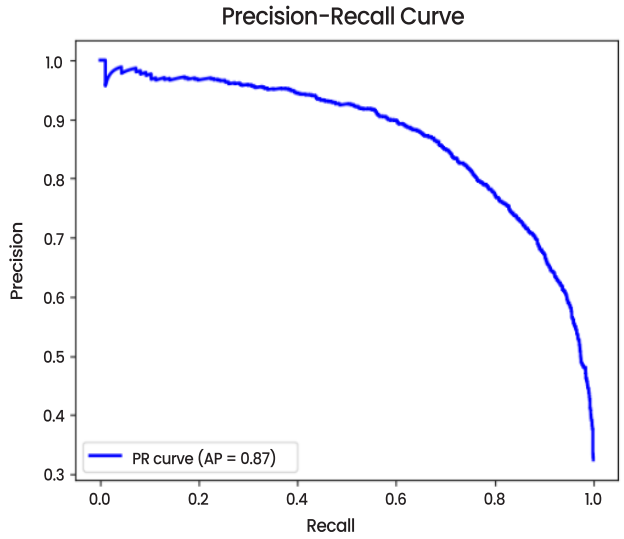

Fig. (18). Precision-recall curve.

Training and validation performance in training and validation performance, as demonstrated by the Loss Over Epochs (Fig. **20**) plot, the model training loss has a decreasing trend, whereas validation loss has an irregular distribution indicating the possibility of overfitting in the later epochs. In the same manner, the Accuracy Over Epochs (Fig. **21**) indicates that the training and validation accuracy are on the rise, and the validation accuracy levels off at 0.86.

Model Metrics:

- **Precision and Recall**: Precision for "Not Looking Ahead" was 0.88, and recall was 0.93. For "Looking Ahead," precision was 0.83, and recall was 0.72.
- **Test Loss and Accuracy**: The test loss was 0.3381, with a corresponding test accuracy of 0.8635.
- **Recall at FPR ≤ 1%**: The model achieved perfect recall of 1.0000 at a false positive rate (FPR) ≤1%.
- **Brier Score**: 0.0995, indicating very good probability calibration.
- **Latency:** The MobileNetV2 (Dual Eye) model achieved a latency of 5615.15 ms per sample.
- **Throughput:** The model demonstrated a throughput of 0.18 FPS.

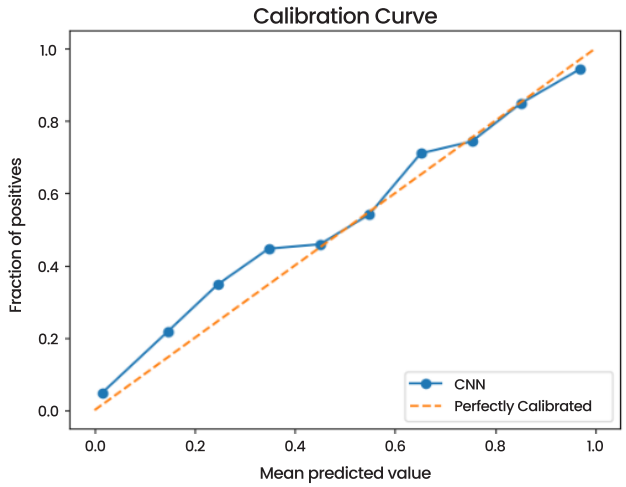

Fig. (19). Calibration curve.

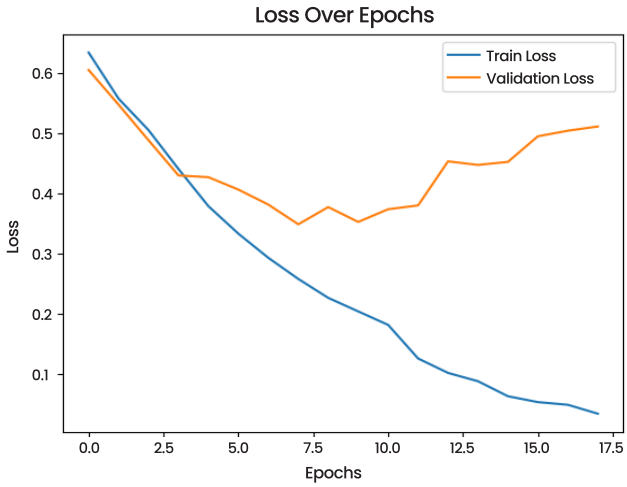

Fig. (20). Loss over epochs.

Finally, the model of MobileNetV2 (Dual Eye) is of a high level, and accuracy, as well as high levels of precision, recall

and AUC, are reached 86%. It is especially applicable in cases where correct and consistent predictions are needed, even though overfitting can occur in small amounts when training. More optimization may be beneficial in augmenting its generalization.

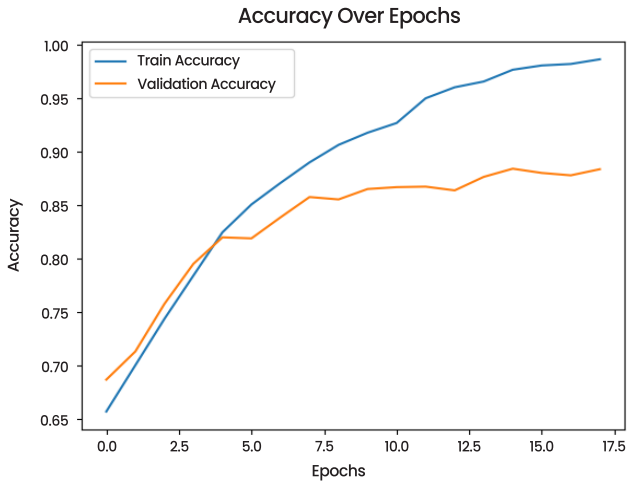

Fig. (21). Accuracy over epochs.

4.4. Model Comparison

This part of the paper compares the performance of three models Convolutional Neural Network (CNN), MobileNetV2 (Single Eye), and MobileNetV2 (Dual Eye) based on different evaluation criteria Table **1**. The classification capability of each model of Looking Ahead *vs.* Not Looking Ahead is evaluated, and their advantages and disadvantages are discussed. CNN model gives a good classification performance with test accuracy of 73.23, AUC is 0.81 and high recall when FPR is less than 1% (0.9990). It however displays certain overfitting since the validation loss variation reflects, which means that it should be further optimized. The model has a balanced accuracy and recall (0.73 in each of the two classes). MobileNetV2 (Single Eye) model has a test accuracy of 75.02, and slightly lower AUC of 0.80. This model has a high recall but a low precision especially when it comes to the "Looking Ahead" class, which has a high recall as compared to precision. This problem is noted by the precision-recall curve yet the model still recalls very well at FPR 1 = 0.9996. MobileNetV2 (Dual Eye) model performs better than the CNN model as well as MobileNetV2 (Single Eye) model with highest test accuracy of 86.35%. It has also got a high AUC of 0.93 that indicates excellent discriminative ability. The model has a good balance in regard to precision-recall especially having an average precision of 0.87 and a significant recall of 0.72 when it comes to the "Looking Ahead" class.

Finally, MobileNetV2 (Dual Eye) model is most advantageous as it offers the most high performance in accuracy, AUC, and precision-recall balance, although the other models, CNN and MobileNetV2 (Single Eye) are also quite promising, especially with respect to recall and recall rates under the low false positive instances.

4.5. Interpretability

To achieve the analysis of interpretability that is more than the actual performance measures, gradient-weighted Class Activation Mapping (Grad-CAM) was applied. The approach will allow visualising the precise regions of the eye images, which the models will focus on in order to make predictions and provide essential information on accuracy and credibility of the system. In the case of single-eye MobileNet model, their Grad-CAM (Fig. **22**) activations revealed that they were constantly localised at the iris and the sclera. These are physiologically salient to gaze direction detection, that is, the network has acquired relevant information and is not working with irrelevant background information. The model also paid attention to the eyelid contours and cornea reflections on some of its samples, which led to the realization of how minor alterations of its texture influence a variation in prediction accuracy.

MobileNet dual eye model offered even more interpretability, as two eye models, right and left are analysed at the same time. The scleral and iris areas activated by the grad-CAM maps (Fig. **23**) were strong and symmetric which revealed that the model merged complementary information between the two eyes with the help of one map. This was particularly evident in problematic cases in which there were glasses, low-light conditions, or partial occlusions; the dual-input structure compensated this case by increasing the consistent signals concerning the gaze that were given by the eye not involved. The visualisations also made sure that the model generalised well under different conditions, which is why there are few chances of biased (or spurious) association.

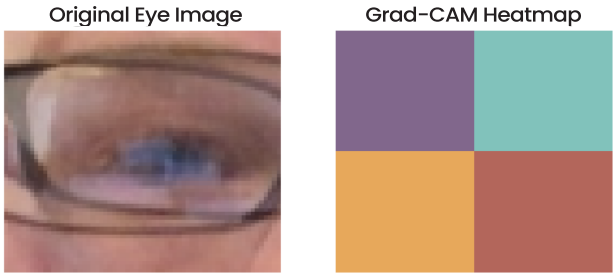

Fig. (22). MobileNetV2 grad-CAM plot.

In addition to visual inspection, interpretability may also be evaluated in terms of quantitative measures, which are used to determine the degree of correspondence between model attention and anatomically relevant areas. As an example, spatial alignment can be objectively measured by the Intersection over Union (IoU) of Grad-CAM heatmaps and manually annotated iris or pupil areas. Meanwhile, focus consistency of attention map across samples can be measured using activation entropy. This kind of quantitative assessment offers some statistical confirmation of interpretability arguments as opposed to the use of qualitative images alone. Moreover, the user-centred evaluations of the plausibility and trustworthiness of visual explanations can be conducted in the future, contributing to the connection of technical interpretability and human perception and usability in the context of real-life usage.

Table 1. Model comparison.

Model	Test Accuracy	AUC	Precision ("Not Looking Ahead")	Precision ("Looking Ahead")	Recall ("Not Looking Ahead")	Recall ("Looking Ahead")	Brier Score
CNN	73.23%	0.81	0.73	0.73	0.73	0.73	0.1782
MobileNetV2 (Single Eye)	75.02%	0.80	0.77	0.66	0.90	0.44	0.1690
MobileNetV2 (Dual Eye)	86.35%	0.93	0.88	0.83	0.93	0.72	0.0995

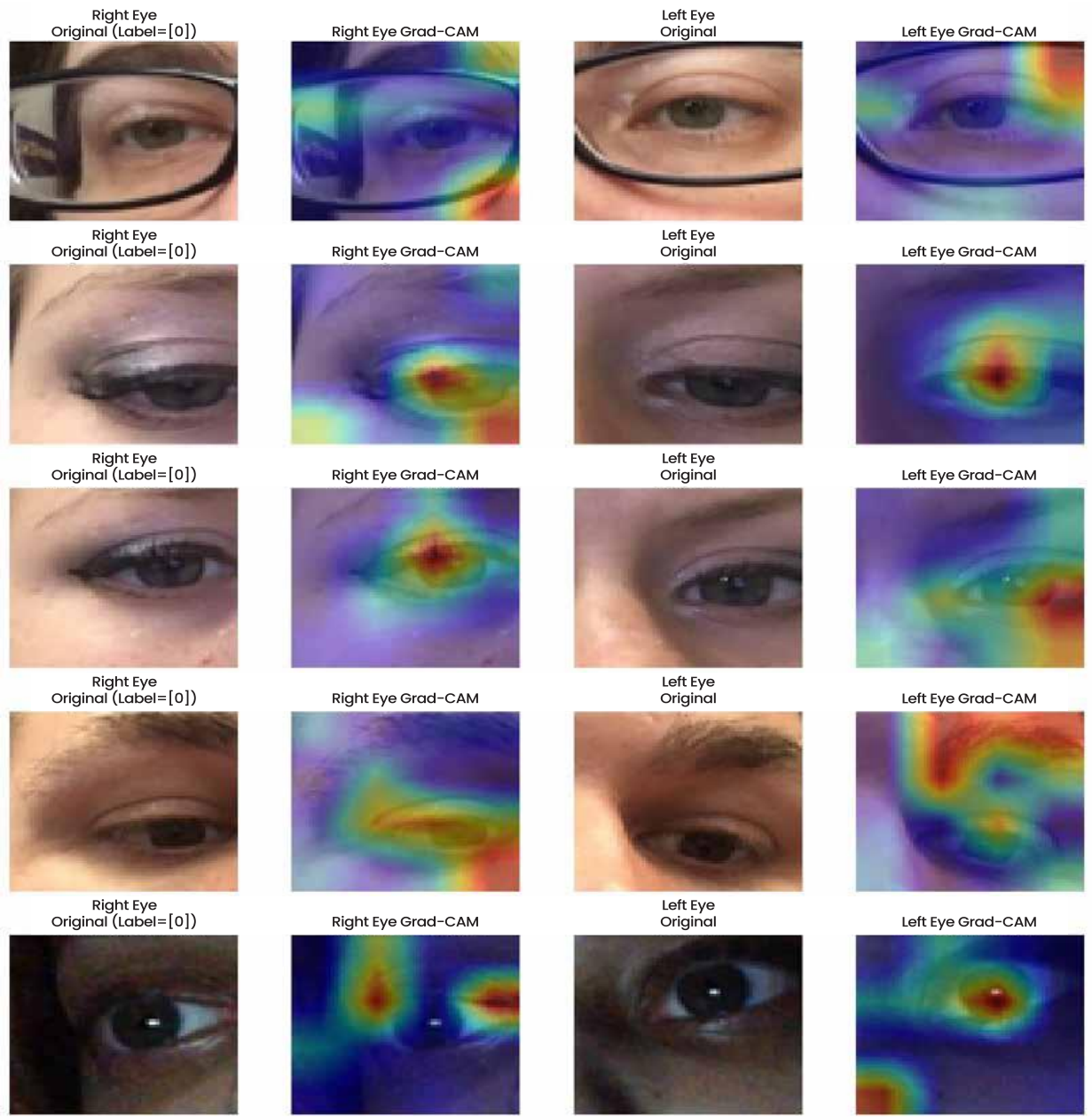

Fig. (23). Dual-input mobileNetV2 grad-CAM plot.

5. DISCUSSION

5.1. Implications

This research has revealed the possibility of using the affordable and real-time gaze tracking solution to create a novel bridge between research prototypes and implementations in practice. Lightweight hardware can be effectively utilised in lightweight architectures such as MobileNet to implement scalable and portable applications of hardware trackers without incurring the high cost of expensive hardware trackers. Gaze-based interfaces can augment virtual reality (AR/VR) to maximise system performance, foveated rendering, and adaptive content delivery. Besides minimising the computational load, these designs have enhanced user interaction and task performance [23]. Gaze tracking has proven to be a viable approach to safety and usability in AR maintenance procedures in industrial settings, supporting the importance of this technology in real-time human-machine interaction [3]. In the same way, gaze tracking would aid predictive systems that identify distraction or fatigue in driver surveillance to enable safer usage on the road [24]. Gaze-enabled interfaces increase access to and diagnostic opportunities in the medical field and assistive technologies, especially with patients with motor impairments who use gaze as their primary communication medium [25].

5.2. Limitations

Although the above implications are promising, the current study has several limitations. To begin with, the dataset was relatively small, which could have caused the problem of class imbalance and lowered generalizability. A larger, more diverse sample can be found in many available datasets, like MPIIGaze or GazeCapture. Nonetheless, their data is noisy and problematic regarding the fidelity of annotations. Poor model performance may be caused by the lack of diversity of datasets as they are applied to unrestricted real-world applications, where head pose, illumination, and occlusion are all highly dynamic. Second, the conditions of the evaluation were virtually simulated. In this way, the implementation scenario would expose additional complications such as latency change on the embedded system, user variability, and hardware-specific constraints. The disadvantages of edge deployment in AR/VR have been identified as the challenge of low latency with thermal and energy limitations, and the discrepancies between controlled experiments and real-world device situations show that this field is blank [26].

Additionally, achieving full cross-domain robustness remains challenging. Subsequent research will incorporate domain-adaptation frameworks and multi-condition datasets to improve resilience to head pose, illumination, and occlusion changes encountered in naturalistic use.Several optimisation strategies can be employed to address the computational overhead observed in the dual-eye MobileNetV2. Quantisation of network weights (*e.g.,* 8-bit integer), structured pruning to remove redundant filters, and model-compression frameworks such as TensorRT or ONNX Runtime can substantially reduce memory usage and inference time without notable accuracy loss. Preliminary tests on comparable lightweight networks indicate that such methods can improve throughput by 10× to 30×, enabling true real-time performance even on resource-constrained AR/VR or automotive edge devices.

The feasibility of deploying the proposed gaze-tracking models on embedded and edge platforms is highly promising given recent advances in low-power AI hardware. Modern chipsets such as Qualcomm Snapdragon XR2, NVIDIA Jetson Nano, and Google Coral TPU support on-device inference with sub-20 ms latency for compressed neural networks. When optimisation techniques such as 8-bit quantisation, layer pruning, and TensorRT acceleration are applied, the single-eye MobileNetV2 can operate at real-time frame rates (50–60 FPS) while consuming less than 5 W of power. Energy efficiency and thermal management remain key design constraints, particularly for AR/VR headsets and in-vehicle driver-monitoring systems; however, adaptive batching and dynamic-frequency scaling can maintain performance within sustainable limits. These results indicate that lightweight, explainable gaze-tracking pipelines can be effectively integrated into edge AI devices, bridging laboratory prototypes with practical AR/VR and automotive applications.

5.3. Future Work

From a future perspective, the scope of research should be pursued in a couple of avenues to make the real-time gaze tracking systems more robust and helpful. The ability to generalise and have a benchmark across devices will be better with increasing assessment to more heterogeneous data, such as MPIIGaze and GazeCapture [25]. The other vital direction is the implementation of Edge AI, since using an efficient gaze model in the AR/VR headsets and in-vehicle systems will ensure a low latency and reduced reliance on cloud services. Of special interest are safety-critical applications, where any delay during response is enough to destroy user confidence and performance of the system [26]. In addition, the continuous tracking of gazes without discrete classification can be further investigated, bringing new medical diagnostic opportunities and a measurement of cognitive load. Constant gaze dynamics modelling may help provide more sophisticated predictions of attention, fatigue, or neurological health, based on the novel VR-based driver monitoring systems [24].

The other research area that has not yet been explored is interpretability, one of the most essential requirements. As more and more gaze-tracking systems are implemented in different sectors, such as health care and transportation, stakeholders need to have understandable and explicable outputs. It is possible to significantly increase the user's trust using methods like Grad-CAM that allow one to visualise the prediction proposal. However, end-users and domain experts have done little systematic research to establish the perspective of such interpretability devices. Future studies should therefore investigate technical accuracy, latency and user-based rating of model transparency.

The study is helpful as it reveals the trade-offs between CNN, MobileNet, and dual-eye MobileNet, using the lightweight and real-time gaze tracking method. Although the application has potential to scale down to AR/VR, driver monitoring and healthcare, there are challenges associated with the diversity of the datasets, real-world implementation and interpretability. These gaps will be addressed using more data, edge deployment techniques and continuous gaze modelling, bringing the field to a stable, affordable, and reliable vision-based implementation.

CONCLUSION

The paper explores the design and testing of the lightweight machine learning based on the real-time gaze tracking, and it is aimed at the accuracy, efficiency, and interpretability. The main objective was to develop a real-time pipeline that can be used to compare various deep learning structures and incorporate interpretability mechanisms to enhance the confidence in the real-world applications. This paper discusses the drawbacks of the old hardware-based trackers, which are typically expensive and obtrusive, and the limitations of the existing software products, which do not provide the accuracy and latency tradeoffs needed in the dynamic, real-world scenario. Three deep learning models were compared: a default convolutional neural network (CNN), a single-eye MobileNetV2 tracker and a dual-eye MobileNetV2 tracker which includes binocular data. The CNN model was successful but had lower accuracy than the MobileNetV2 models. Dual-eye MobileNetV2 model had the highest accuracy (86.35) and the AUC (0.93), and the model was able to perform better at the distinction between the "Looking Ahead" and the "Not Looking Ahead" category. But since it has a high latency (5615 ms/sample), it is not suitable in its current form in real-time application.

The single-eye MobileNetV2 model was the most optimal model to be used in real-time mode, as it provides a good balance between the predictive performance and the latency (1.30 ms/sample), and the throughput is more than 700 FPS. The model is also suitable in the case of resource-constrained settings, like, AR/VR systems and driver monitoring, where real-time response is highly important. This research paper will offer an overall outline of how gaze tracking systems can be critically assessed in the real world in terms of accuracy, latency and interpretability. Grad-CAM visualizations were added to increase the transparency of the model upon which trust and accountability are crucial in the application of the model, including healthcare and automotive systems. All in all, the results show that interpretable, lightweight machine learning pipelines can be used in the future to create vision-based systems that are real-time and have a wide variety of applications in AR/VR, driver monitoring, assistive technologies, and healthcare.

LIST OF ABBREVIATIONS

AP	=	Average Precision
AR	=	Augmented Reality
CNNs	=	Convolutional Neural Networks
EEG	=	Electroencephalography
FPS	=	Frame-Per-Second
Grad-CAM	=	Gradient-Weighted Class Activation Mapping
IoU	=	Intersection Over Union
LSTM	=	Long Short-Term Memory
RNNs	=	Recurrent Neural Networks
SVMs	=	Support Vector Machines
VR	=	Virtual Reality

AUTHOR'S CONTRIBUTION

H.T. has contributed to the study concept, data collection, analysis, manuscript writing, data collection, writing, and proofreading.

ETHICAL APPROVAL & INFORMED CONSENT

All procedures were carried out in accordance with institutional research ethics committee guidelines and Declaration of Helsinki. Informed consent was obtained from all participants. To ensure participant protection, all data were fully anonymized at the point of collection, and no personal or identifiable data was recorded.

AVAILABILITY OF DATA AND MATERIALS

The data will be made available on reasonable request by contacting the corresponding author [H.T.]

FUNDING

None.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this article.

ACKNOWLEDGEMENTS

Declared none.

DECLARATION OF AI

During the preparation of this work the authors used ChatGPT for editing purposes. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

[1] Ozhan FO, Aygun U, Sahin A, Urey H. Dynamic accommodation measurement using Purkinje reflections and machine learning. Scientific Reports. 2023 Dec 7;13(1):21625. https://doi.org/10.1038/s41598-023-47572-0

[2] Jin X, Chai S, Tang J, Zhou X, Wang K. Eye-tracking in ar/vr: A technological review and future directions. IEEE Open Journal on Immersive Displays. 2024 Aug 27;1:146-54. https://doi.org/10.1109/OJID.2024.3450657

[3] Burova A, Mäkelä J, Hakulinen J, Keskinen T, Heinonen H, Siltanen S, Turunen M. Utilizing VR and gaze tracking to develop AR solutions for industrial maintenance. InProceedings of the 2020 CHI conference on human factors in computing systems 2020 Apr 21 (pp. 1-13). https://doi.org/10.1145/3313831.3376405

[4] Van Damme S, Vega MT, De Turck F. Human-centric quality management of immersive multimedia applications. In2020 6th IEEE Conference on Network Softwarization (NetSoft) 2020 Jun 29 (pp. 57-64). IEEE. https://doi.org/10.1109/NetSoft48620.2020.9165335

[5] Bozkir E, Geisler D, Kasneci E. Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In2019 IEEE conference on virtual reality and 3D user interfaces (VR) 2019 Mar 23 (pp. 1834-1837). IEEE. https://doi.org/10.1109/VR.2019.8797758

[6] Shabbir MN, Linh DT. AI and machine learning applications in wearable health devices. Wearable Technology. 2024;5(1):3123. https://doi.org/10.54517/wt3123

[7] Bektaş K, Strecker J, Mayer S, Garcia K. Gaze-enabled activity recognition for augmented reality feedback. Computers & Graphics. 2024 Apr 1;119:103909. https://doi.org/10.1016/j.cag.2024.103909

[8] Lu S, Tan Z, Kong S, Zhang D. Cost-effective gaze tracking system based<? pag\break?> on polymer fiber specklegrams. Optics Letters. 2024 Sep 3;49(18):5027-30. https://doi.org/10.1364/OL.531946

[9] Hallgarten P, Sendhilnathan N, Zhang T, Sood E, Jonker TR. Gears: Generalizable multi-purpose embeddings for gaze and hand data in vr interactions. InProceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization 2024 Jun 22 (pp. 279-289). https://doi.org/10.1145/3627043.3659551

[10] Brousseau B, Rose J, Eizenman M. Hybrid eye-tracking on a smartphone with CNN feature extraction and an infrared 3D model. Sensors. 2020 Jan 19;20(2):543. https://doi.org/10.3390/s20020543

[11] Ahmed I, Ateeb M. Enhancing Education Accessibility through Eye Gaze Technology (EGT). In2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM) 2025 Jan 3 (pp. 1-8). IEEE. https://doi.org/10.1109/IMCOM64595.2025.10857579

[12] Kumari N, Ruf V, Mukhametov S, Schmidt A, Kuhn J, Küchemann S. Mobile eye-tracking data analysis using object detection via YOLO v4. Sensors. 2021 Nov 18;21(22):7668. https://doi.org/10.3390/s21227668

[13] Darapaneni N, Prakash MD, Sau B, Madineni M, Jangwan R, Paduri AR, KP J, Belsare M, Madhavankutty P. Eye Tracking Analysis Using Convolutional Neural Network. In2022 Interdisciplinary Research in Technology and Management (IRTM) 2022 Feb 24 (pp. 1-8). IEEE. https://doi.org/10.1109/IRTM54583.2022.9791826

[14] Chinsatit W, Saitoh T. CNN-Based Pupil Center Detection for Wearable Gaze Estimation System. Applied Computational Intelligence and Soft Computing. 2017;2017(1):8718956. https://doi.org/10.1155/2017/8718956

[15] Gunawardena N, Ginige JA, Javadi B, Lui G. Performance analysis of cnn models for mobile device eye tracking with edge computing. Procedia Computer Science. 2022 Jan 1;207:2291-300. https://doi.org/10.1016/j.procs.2022.09.288

[16] Alzubaidi LH, Hussein AH, Alkhafaji MA, Shilpa N, NP T. Efficient Real-Time Eye Gaze Tracking Detection for Human-Computer Integration Using Advanced Techniques. In2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC) 2023 Dec 4 (pp. 1-6). IEEE. https://doi.org/10.1109/ICMNWC60182.2023.10436021

[17] Chen HH, Hwang BJ, Wu JS, Liu PT. The effect of different deep network architectures upon CNN-based gaze tracking. Algorithms. 2020 May 19;13(5):127. https://doi.org/10.3390/a13050127

[18] Khan W, Ishrat M, Faisal SM, Ahmad F, Nabilal KV, Sharma YK. A robust framework for topology-based anomaly detection in attributed networks using graph attention networks, substructure analysis, and data augmentation. International Journal of Data Science and Analytics. 2025 Nov;20(7):6715-28. https://doi.org/10.1007/s41060-025-00848-2

[19] Khan W, Ebrahim N. Anogat-sparse-tl: A hybrid framework combining sparsification and graph attention for anomaly detection in attributed networks using the optimized loss function incorporating the twersky loss for improved robustness. Knowledge-Based Systems. 2025 Feb 28;311:113144. https://doi.org/10.1016/j.knosys.2025.113144

[20] Donuk K, Ari A, Hanbay D. A CNN based real-time eye tracker for web mining applications. Multimedia Tools and Applications. 2022 Nov;81(27):39103-20. https://doi.org/10.1007/s11042-022-13085-7

[21] Park J, Park S, Cha H. Gazel: Runtime gaze tracking for smartphones. In2021 IEEE International Conference on Pervasive Computing and Communications (PerCom) 2021 Mar 22 (pp. 1-10). IEEE. https://doi.org/10.1109/PERCOM50583.2021.9439113

[22] Martinez, J.P. Gaze Direction Detection: Dataset with separated eyes of different people, looking or not at the camera. 2025; Available from: https://www.kaggle.com/datasets/estopadilla/gaze-direction-detection.

[23] Liu W, Duinkharjav B, Sun Q, Zhang SQ. Fovealnet: Advancing ai-driven gaze tracking solutions for efficient foveated rendering in virtual reality. IEEE Transactions on Visualization and Computer Graphics. 2025 Mar 11. https://doi.org/10.1109/TVCG.2025.3549577

[24] Maruyama R, Takahashi S, Hagiwara T. A Virtual Reality Driving Simulator with Gaze Tracking for Analyzing Driver's Behavior. In2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech) 2022 Mar 7 (pp. 144-145). IEEE. https://doi.org/10.1109/LifeTech53646.2022.9754947

[25] Biradarpatil RS, Chinmayee BL, Hegde S, Mallibhat K, Mudenagudi U. Eye Gaze Tracking Towards User Attention Analysis. In2024 5th International Conference for Emerging Technology (INCET) 2024 May 24 (pp. 1-8). IEEE. https://doi.org/10.1109/INCET61516.2024.10593440

[26] Jawalkar SK. The Metaverse: A New Paradigm for Social and Commercial Interaction. IJLRP-International Journal of Leading Research Publication.;5(2). https://doi.org/10.5281/zenodo.14982529.