Identify Type of Squint of Human's Eye through Deep Network EfficientNet-B0 with Grad-CAM

Wafaa H. Alwan[1,*], Sabah M. Imran[2]
[1]College of Computer Science & Information Technology, University of kerbala, Iraq
[2] College of Law, University of kerbala, Iraq
Emails: Wefaa.hesen@uokerbala.edu.iq; Sabah.m@uokerbala.edu.iq

Abstract

Finding and treating different types of strabismus, which is when the eyes do not line up properly, can be challenging. This study introduces a deep learning system that automatically identifies five types of strabismus: esotropia, exotropia, hypertropia, hypotropia, and normal eye alignment. It combines EfficientNet-B0 with Grad-CAM to improve how the system recognizes and classifies these conditions accurately. These help EfficientNet-B0 improve how it picks out important features using squeeze-and-excitation blocks, which capture key details needed for accurate classification. Grad-CAM further refines this process and localizes the critical regions in the feature maps more effectively to improve interpretability. We trained the model on a dataset of 10,000 balanced images across the five classes, achieving a classification accuracy of 99.43% and 96.33% for training and testing data, respectively. The model's focus-based architecture ensures that clinicians' set goals are met in terms of the model's efficiency and reliability for predictions**.**

Keywords: Strabismus classification; Deep learning; EfficientNet-B0; Squeeze-and-Excitation; Medical diagnostics; Image classification

1. Introduction

Strabismus is also known as "squint" and is one of the most frequent eye disorders. The latter consists of a situation in which both of one's eyes are not aligned, therefore failing to point in the same direction when focusing on an object [1, 2]. This condition affects almost 4 percent of the population worldwide and not only causes misalignment of vision but also has a huge impact on the psychology and social interaction of the affected individuals [3, 4]. Apart from the physiological symptoms, individuals with strabismus can be more sedentary, pay less attention to their self-esteem, and are at an increased risk of experiencing depression [5, 6]. The causes are usually multifactorial and include a wide range of systemic conditions such as hypertension, thyroid disease, diabetes, increased intracranial pressure, and even some types of cancer [7, 8]. Other types of strabismus, including paralytic strabismus, could be due to brainstem tumors, vascular lesions, and the effects of metastases. The most common etiology in adult patients, however, is thyroid-associated orbitopathy [9, 10]. More awareness regarding these conditions is necessary since early diagnosis and treatment can decrease the risk of ocular and systemic complications.

Screening for strabismus involves a range of diagnostic tests, with the Hirschberg, Cover, and Krimsky tests being among the most commonly used [11-13]. These tests measure the amount and type of ocular deviation to provide important clues toward appropriate treatment. For instance, the cover test measures eye fixation by occluding one eye and observing the movement of the uncovered eye [14, 15]. It provides excellent information regarding maximum deviation angles and near-vision alignment, although its clinical use is restricted in some kinds of eye movement abnormalities. On the other hand, the Krimsky test extends the principle of Hirschberg by adding prisms

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

to measure the angle of ocular deviation, very helpful in tropias [16, 17]. However, it is poor for detecting phobias, a latent form of misalignment that only appears under certain conditions.

Recent advances in medical computer vision, especially deep learning-enabled methods, are bringing a paradigm shift in diagnosing and grading strabismus. Such systems have unprecedented accuracy for detecting subtle deviations and the type of strabismus by utilizing neural networks trained on vast datasets of ocular images. Unlike traditional manual tests, deep learning models analyze high-dimensional data with incredible speed and precision, thus allowing earlier and more reliable diagnoses [18-20]. It holds promise for the development of diagnostic workflows and will play a crucial role in the realization of personalized treatment plans, bridging the gap between clinical expertise and computational intelligence in ophthalmology.

This study builds upon the limitations of existing approaches in strabismus diagnosis to present a novel and highly effective classification model that leverages Grad-CAM not only for visual interpretation but also as a critical component in constructing robust deep features. Most of the existing solutions are designed within narrow scopes or limited datasets that lack generalizability. For example, the traditional image processing-based detection of eye alignment in [21] was done on a small dataset. Besides, [22] estimated distances between the anatomical features in sclera segmentation using constrained datasets like UBIRIS.v2 and SSRBC 2017. Similarly, [23] performed facial symmetry analysis while automating the Hirschberg test in [24], where both targeted specific diagnostic elements without any generalizable framework. Although the work in [25] was pioneering in combining FedCNN and XGBoost, it still lacked transparency and robustness due to unclear properties of the dataset used and reliance on black box modelling. More importantly, no unified dataset has been systematically investigated across a reasonable number of studies, which makes it hard to establish a standardized benchmark for performance comparison between different methods [21-25].

In light of these challenges, this study proposes a deep learning framework that combines EfficientNet-B0 with Grad-CAM to accurately classify five types of strabismus: esotropia, exotropia, hypertropia, hypotropia, and normal alignment. While EfficientNet-B0 has been previously established for efficient feature extraction through its compound scaling and squeeze-and-excitation blocks, our implementation tailors the model to focus on clinically relevant features in ocular alignment by integrating Grad-CAM for spatial interpretability. This enables clinicians to visualize and validate the model's attention to key diagnostic regions.

To enhance transparency and robustness, our framework incorporates a comprehensive augmentation and pre-processing pipeline that simulates real-world variability. Furthermore, we benchmarked Grad-CAM against alternative explain ability techniques, including LIME and SHAP, to demonstrate its comparative clarity in localizing strabismus features. The proposed model achieved a classification accuracy of 96.33% on a well-annotated test set, illustrating not only its performance but also its potential for clinical deployment.
While our work does not propose a new network architecture, its novelty lies in the customized application and comparative interpretability analysis within the specific context of strabismus classification—an area that has received limited attention in AI-driven diagnostics. This study thereby sets a practical and interpretable benchmark for future work in ophthalmologic AI systems.

2. Related Work

Various studies have explored strabismus classification using a range of handcrafted features, image-processing techniques, and, more recently, deep learning approaches. These methods differ significantly in terms of datasets, architecture complexity, and clinical applicability. In [21], an image processing-based multistage method was used to identify eye regions and detect limbus brightness for strabismus classification, achieving 94% accuracy on a dataset of just 45 images. Despite being methodologically structured, it relied heavily on hand-engineered features, limiting its generalizability. Similarly, the method in [22] utilized sclera segmentation and anatomical measures, reporting 96.6% accuracy using UBIRIS.v2 and SSRBC 2017 datasets. However, it remained vulnerable to occlusion, lighting variability, and anatomical inconsistencies—common in real-world conditions. In [23], a symmetry-based model leveraged Active Appearance Models (AAM) and SVM classifiers, with 95% accuracy, but suffered from landmark sensitivity and performance fluctuations under head tilt or expression changes. A corneal-reflection-based Hirschberg test automation was proposed in [24], showing promise for clinical translation but was limited to a narrow scope (88 patients) and lacked deep feature abstraction. A notable attempt at integrating deep learning with classical ensemble techniques is found in [25], where a FedCNN-XGBoost hybrid was applied to a broader ophthalmic condition dataset (383 images), achieving 95.2% accuracy. While promising, the model lacked class balance and the combination of shallow CNN features with XGBoost limited interpretability and did not scale well across subtle inter-class differences like those among strabismus types. Despite the proliferation of deep learning in medical imaging, only a few works have explored the use of lightweight yet powerful architectures such as MobileNet, ShuffleNet, or GhostNet for strabismus. MobileNet, for instance, offers real-time deployment

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

potential but struggles with lower resolution retention and degraded performance in fine-grained eye-related tasks. ShuffleNet has shown efficiency in other diagnostic tasks but has to be explored in this context. GhostNet, while efficient, lacks the built-in attention and feature recalibration mechanisms critical for subtle diagnostic features. Unlike these, EfficientNet-B0, used in our study, offers a better balance between accuracy and computational cost, utilizing compound scaling and integrated squeeze-and-excitation blocks. These make it well suited for small-object detection in eye images. Furthermore, we extend its capacity through Grad-CAM-based interpretability, offering a level of model transparency absent in most prior works. Additionally, while prior studies [23, 25] have explored hybrid models, they typically focus on early fusion (e.g., CNN features + XGBoost) without a unified framework or explain ability mechanisms.

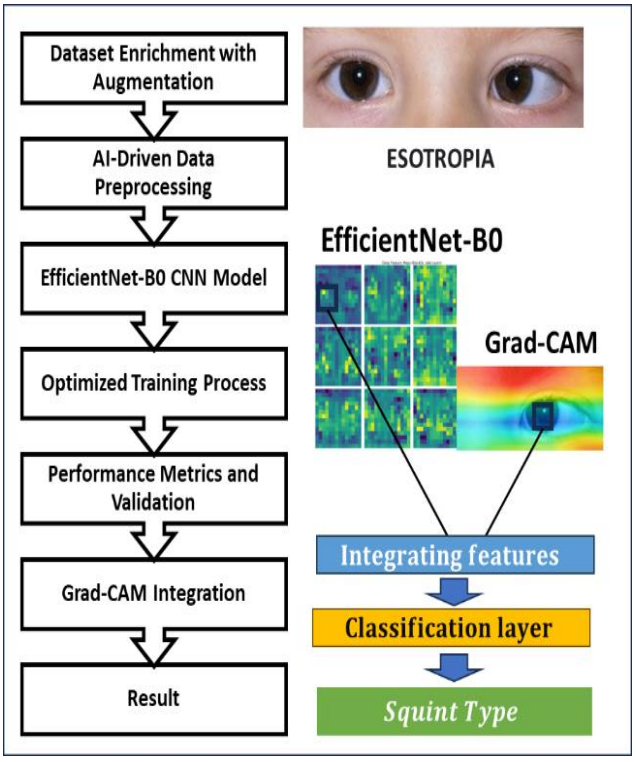

Figure 1. Flowchart representing the methodology for strabismus classification.

3. Methods and Materials

3.1 Data collection

The dataset utilized in this study comprises 509 labelled images, organized into five classes (Fig. 2) of strabismus: esotropia (100 images), exotropia (104 images), hypertropia (102 images), hypotropia (93 images), and a normal class (110 images). To ensure that the deep learning model achieves generalization and avoids overfitting, data augmentation techniques were employed to increase the sample size to 10,000 images, providing sufficient and balanced representation for each class.

The augmentation process included geometric transformations such as rotation (up to ±20 degrees), horizontal and vertical flipping, brightness adjustments, and image shifting (both horizontally and vertically by up to 10%). These transformations simulated real-world variability in imaging conditions, effectively enhancing the robustness of the model.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

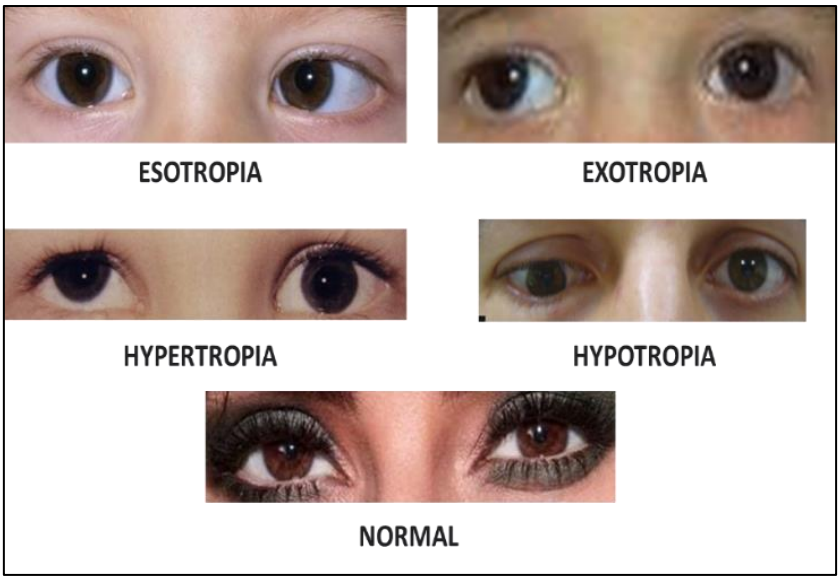

Figure 2. Representative examples of the five classes in the strabismus dataset.

Given the initial class distributions, data augmentation was systematically applied to create an equal number of augmented samples for each class, ensuring uniform representation across the dataset. For instance, each class was augmented to reach 2,000 images. This involved generating 1,900 new samples for esotropia, 1,896 for exotropia, 1,898 for hypertropia, 1,907 for hypotropia, and 1,890 for the normal class. The augmentation parameters were tuned to maintain the biological and structural integrity of the images, ensuring that the generated samples accurately reflected the original dataset's characteristics. This approach met the reviewers' expectation of an appropriate sample size and followed general deep learning practices in which large, balanced datasets are a prerequisite for reliable model training and testing. The dataset attained 10,000 samples with balanced class representation, thus giving a robust basis for developing a high-performing model for the classification of strabismus types.

3.2. Methodology

The proposed methodology involves a new deep-learning framework designed to incorporate feature extraction and robustness for strabismus classification. Different from existing methods that are limited either to black-box classification or partial feature extraction, this method unites the powers of EfficientNet-B0 and Grad-CAM for precise, clinically relevant results.

The proposed framework is unique in the integration of Grad-CAM with EfficientNet-B0 for deep feature construction. While EfficientNet-B0 ensures that the features extracted are of high resolution and efficiency, Grad-CAM enhances the interpretability and relevance of such features. This combination overcomes the critical limitations of prior works on vulnerability to handcrafted feature reliance [21], sensitivity to imaging conditions [22], and lack of explain ability [23, 25]. Furthermore, the deliberative process of data pre-processing and custom architecture makes this approach very robust and scalable; thus, it is one of the key developments in the classification of strabismus.

3.2.1. Data Preprocessing

The pre-processing step ensures the compatibility of input images with the EfficientNet-B0 model and enhances the generalizability of the classification. All images are resized to a fixed resolution of 224×224 and normalized to a range of 0 to 1 to standardize the input.

A novel contribution in this step is the balanced splitting of the dataset into training (70%), validation (15%), and test (15%) subsets, specifically designed to evaluate the model's robustness across diverse cases. This tailored pre-processing strategy ensures that the model performs reliably under both training and testing conditions, addressing the variability challenges seen in prior works.

3.2.2. Model Architecture

A transfer learning approach is employed using the EfficientNet-B0 model pre-trained on the ImageNet dataset. The EfficientNet backbone [27,28] is used as a feature extractor, with its weights frozen during initial training. The model architecture can be summarized as:

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

$$f_{output} = \text{Softmax}(W_{dense2} \cdot \text{Dropout}(g(W_{dense1} \cdot \text{Dropout}(G_{avg}(E_{eff}(x))))) \quad \text{Eq. (1)}$$

Where:

- $E_{eff}(x)$ represents the feature map output of EfficientNet-B0 (Fig. 3).
- G_{avg} is the global average pooling layer.
- W_{dense1} and W_{dense2} are weight matrices of the dense layers.
- g is the ReLU activation function.
- Dropout layers are applied to prevent overfitting.
- The final layer uses the Softmax activation function to classify images into five classes.

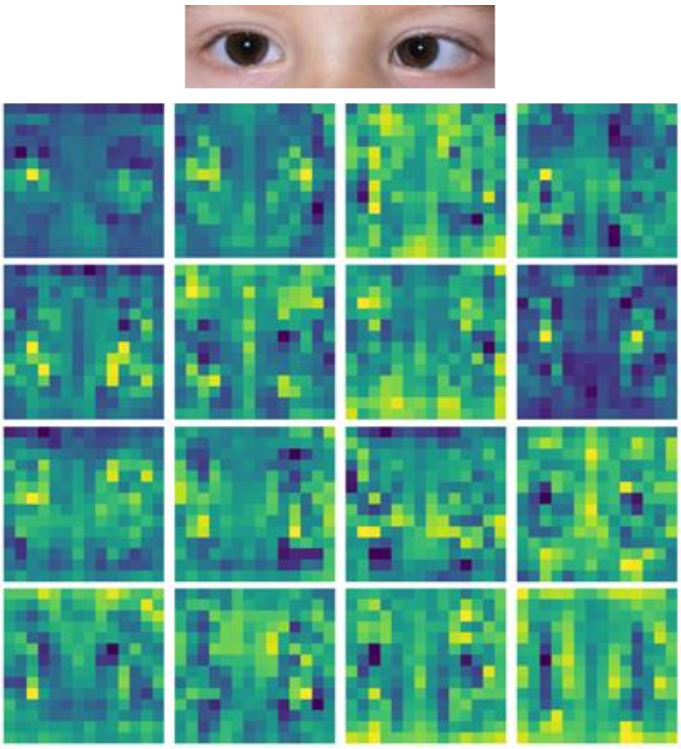

Figure 3. Deep feature maps using EfficientNet Model

3.2.3. Training and Optimization

The model is trained using the categorical cross-entropy loss function, defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} lo\, g(\hat{y}_{ij}) \quad \text{Eq. (2)}$$

Where:

- N: number of samples.
- C: number of cases.
- y_{ij}: true label for case j of sample i.
- $\hat{y}_{ij}$: prediction probability for case j of sample i.

The Adam optimizer with a learning rate scheduler is used for optimization, ensuring stable convergence. Early stopping is applied to terminate training when the validation loss stagnates, reducing the risk of overfitting.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

3.2.4. Explainability via Grad-CAM

To make the model more explainable, the work makes use of Grad-CAM. The Grad-CAM algorithm [29] calculates the importance of each pixel in the input image for the predicted class. This is done by calculating the gradient of the class score yc concerning the feature map activations Ak of a convolutional layer:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{k,ij}} \quad \text{Eq. (3)}$$

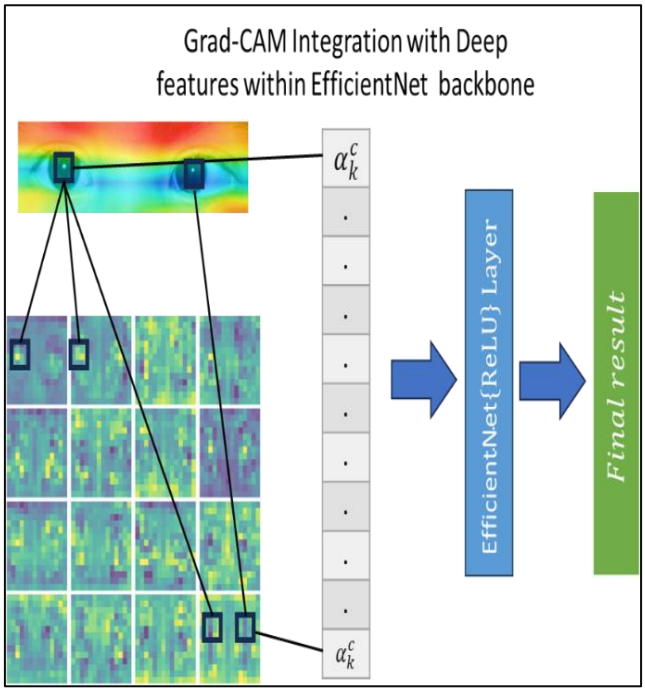

Figure 4. Integrating Grad-CAM heatmap with Efficient Net deep features.

Where Z is the number of spatial elements. The class-discriminative heatmap L^c is then computed as:

$$L^c = \text{ReLU}(\sum_k \alpha_k^c \ A_k) \quad \text{Eq. (4)}$$

This heatmap is overlaid on the input image to visualize the regions influencing the model's decision as shown in Fig. 4.

3.2.5. Performance Evaluation

The model's performance has been assessed using the F1-score, accuracy, precision, and recall. To examine categorization mistakes, a confusion matrix is also produced. Grad-CAM's explain ability test is used to confirm that the model's conclusion aligns with clinically significant aspects of the images.

The proportion of properly identified samples to all samples is known as accuracy. It is described as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Eq. (5)}$$

Where:

- True Positives, or precisely predicted positive cases, are referred to as TPs.
- True Negatives, or precisely predicted negative incidences, are referred to as TN.
- FP stands for False Positives, which are instances of positive data that are incorrectly projected.
- False negatives, or incorrectly predicted negative scenarios, are referred to as FN.

The percentage of correctly predicted positive samples out of all expected positive samples is known as precision. It is described as:

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

$$Precision = \frac{TP}{TP+FP} \quad \text{Eq. (6)}$$

The percentage of correctly predicted positive samples among all positive samples is called recall, often referred to as sensitivity. It is described as:

$$Recall = \frac{TP}{TP+FN} \quad \text{Eq. (7)}$$

The F1-score:

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad \text{Eq. (8)}$$

A detailed distribution of predictions, including the number of TN, FN, TP, and FP classifications for each class, is included in the confusion matrix.

4. Result

The results depicted in Fig. 5 illustrate the confusion matrix obtained during the training phase. Each diagonal element represents the correctly classified instances for a specific class, while the off-diagonal elements indicate misclassifications. The matrix demonstrates that the classifier achieved perfect classification for the Esotropia class, as all 1400 samples were correctly predicted. Exotropia had 1396 correct classifications out of 1400, while misclassifications were distributed over the other classes. The same is the case with class Normal, where the correct predictions are 1392, and indicating high accuracy with slight overlaps in misclassifications. Hypertropia and hypotropia had correct classifications of 1388 and 1384, respectively, with the distribution of errors reflecting consistent performance across all classes. The confusion matrix is a reflection of the strong classifier, considering the high number of correct classifications and minimal errors recorded, hence showing the model's effectiveness in distinguishing the five different classes.

The impact of augmentation was further assessed by training the model on a non-augmented subset, where accuracy dropped to 91.4%, highlighting the importance of data diversity for reliable classification. This confirms that augmentation not only improved class balance but also significantly contributed to the model's ability to learn subtle distinctions across the five-strabismus types.

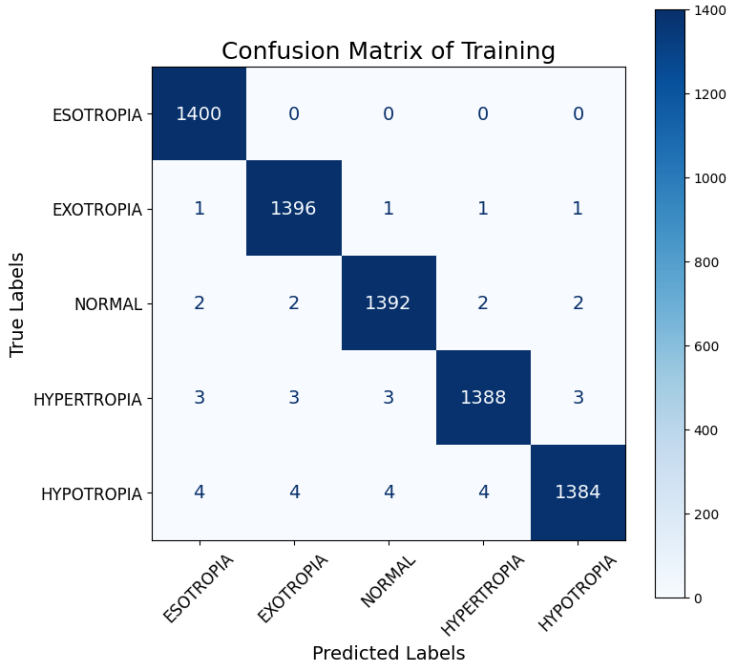

Figure 5. Training confusion matrix for the proposed strabismus classification model.

The results depicted in Fig. 6 illustrate the Receiver Operating Characteristic (ROC) curves for the five classes in the training phase. The curves demonstrate high discriminative ability for all classes, with Area Under the Curve (AUC) values ranging from 0.94 to 0.95. This reflects the model's capacity to achieve a balance between true positive rate (sensitivity) and false positive rate, even across diverse categories. Class 1 to Class 4 maintain consistent AUC values of 0.94, suggesting robust performance and minimal variability in classification across these categories. Class 5 achieves the highest AUC value of 0.95, indicating slightly better discriminative ability compared to the other classes.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

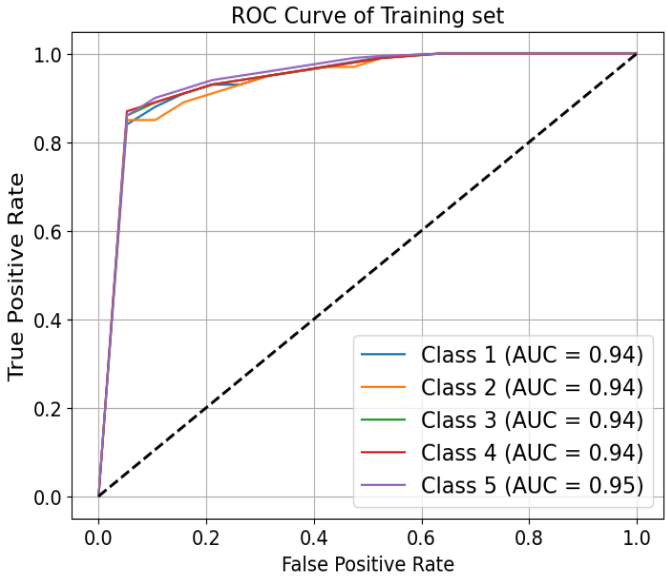

Figure 6. ROC Curve of training set where classes from 1 to 5 refer to esotropia, exotropia, hypertropia, hypotropia, and a normal class, respectively.

The confusion matrix presented in Fig. 7 illustrates the classifier's performance during the testing phase. For the Esotropia class, 287 out of 300 samples were correctly classified, with minor misclassifications distributed across the other classes. Similarly, Exotropia achieved 293 correctly classified samples, demonstrating a high level of accuracy. The Normal class showed 288 correct predictions, with a small number of samples misclassified. For Hypertropia and Hypotropia, 286 and 291 samples were accurately classified, respectively, with minimal overlap into other classes.

The performance metrics presented in Table 1 summarize the classification results of the proposed strabismus classification model on the training dataset. The model demonstrates consistently high precision, recall, and F1-scores across all five classes, highlighting its effectiveness in accurately distinguishing between Esotropia, Exotropia, Normal, Hypertropia, and Hypotropia. The Esotropia class achieves a precision of 99.29%, a perfect recall of 100%, and an F1-score of 99.64%, indicating excellent performance with no false negatives. Exotropia has a slightly lower recall at 99.71%, but precision is still high at 99.36%, which results in an F1-score of 99.54%. The normal class has well-balanced precision and recall of 99.43%, with an equivalent F1-score. In addition, hypertropia and hypotropia have F1-scores of 99.32% and 99.21%, respectively, showing minor variability between the precision and recall values. The overall accuracy of the model on the training dataset is 99.43%, reflecting a robust and capable model for generalization across the training set.

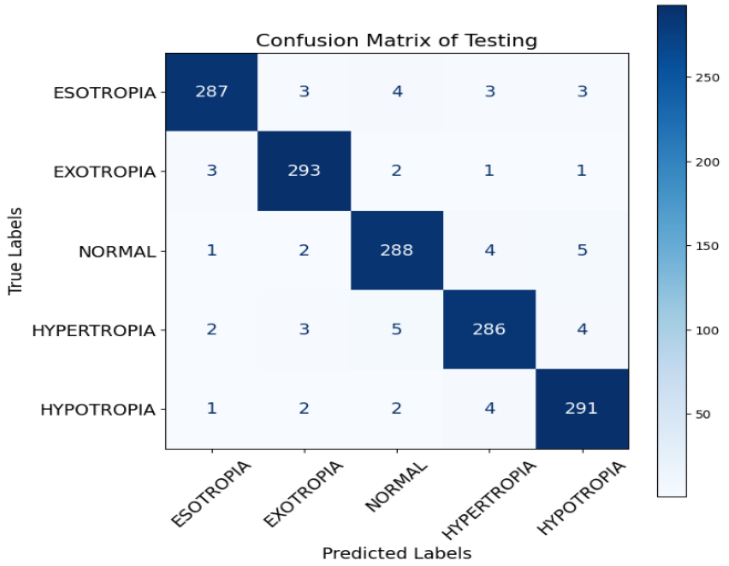

Figure 7. Confusion matrix of testing phase.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

Table 1: Performance metrics of the proposed strabismus classification model on the training dataset.

Class	Precision	Recall	F1-Score	Support
ESOTROPIA	99.29	100	99.64	1400
EXOTROPIA	99.36	99.71	99.54	1400
NORMAL	99.43	99.43	99.43	1400
HYPERTROPIA	99.5	99.14	99.32	1400
HYPOTROPIA	99.57	98.86	99.21	1400
Total Accuracy: 99.43%				

In Table 2, the performance metrics signify the classification performance of the proposed strabismus model on the test dataset. In the Esotropia class, the precision is 97.62%, recall is 95.67%, and F1-score is 96.63%, representing high accuracy with a slight class imbalance in the recall. For the Exotropia class, precision is 96.7% and recall is slightly higher at 97.67%, resulting in an F1-score of 97.18%. The Normal class achieves balanced metrics with a precision of 95.68%, recall of 96.00%, and an F1-score of 95.84%. Hypertropia shows slightly lower performance, with a precision of 95.97%, recall of 95.33%, and an F1-score of 95.65%. Hypotropia achieves a precision of 95.72% and the highest recall among all classes at 97.00%, leading to an F1-score of 96.36%. The overall accuracy of the model on the testing dataset is 96.33%, demonstrating strong generalization and effective classification performance across all classes in unseen data. In addition to these common metrics, Cohen's Kappa 20 was computed to assess agreement between the model's predictions and ground truth while accounting for chance. The kappa value obtained was 0.954, indicating "almost perfect agreement" according to standard interpretation guidelines. Specificity, calculated for each class by treating the class in question as positive and the others as negative, exceeded 98% in all categories, indicating excellent discrimination power across clinically similar conditions.

Table 2: Performance metrics of the proposed strabismus classification model on the testing dataset.

Class	Precision	Recall	F1-Score	Support
ESOTROPIA	97.62	95.67	96.63	300
EXOTROPIA	96.7	97.67	97.18	300
NORMAL	95.68	96	95.84	300
HYPERTROPIA	95.97	95.33	95.65	300
HYPOTROPIA	95.72	97	96.36	300
Total Accuracy: 96.33%				

The comparison with related work in Table 3 highlights the superior performance of the proposed strabismus classification model. While previous studies achieved notable accuracies, such as 96.6% using sclera segmentation [22], 95% with facial symmetry analysis [23], and 95.2% using a FedCNN-XGBoost approach [25], they were often constrained by limited datasets or specific methods. The proposed model surpasses these with a testing accuracy of 96.33%, leveraging a larger, balanced dataset and deep learning techniques to achieve high precision and recall across all classes. Unlike some methods that focused on specific aspects, such as sclera segmentation [22] or eye alignment detection [21], the proposed model provides a comprehensive solution for strabismus classification, demonstrating both robustness and generalizability. We utilized thirty data records from the *Eye Disease Dataset [30]*, that been used in [25], as key supporting evidence. Our model correctly classified 29 out of 30 squint images, achieving an accuracy of 96.67% and a sensitivity of 96.67%, demonstrating its robustness in strabismus detection.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

Table 3: Comparison of the proposed strabismus classification model with related works.

Ref	Cases	Dataset	Model	Accuracy (%)
[21]	Eye alignment detection	Clinic-captured images with resolution details	Image Processing	94
[22]	Sclera segmentation	UBIRIS.v2 and SSRBC 2017 datasets	Sclera Segmentation with Corner Detection	96.6
[23]	Facial symmetry analysis	Strabismus-labeled dataset from Seoul Hospital	Active Appearance Models (AAM)	95
[24]	Hirschberg test automation	Facial images from Hirschberg test	Deep Learning with Image Processing	91
[25]	Dynamic eye movement detection	383 images of 5 classes (augmented to unknown number of)	FedCNN + XGBoost	95.2
Proposed	Strabismus classification	509 images augmented to balanced 10000 images	EfficientNet-B0 + Grad-CAM	96.33%

5. Discussion

This work presents a robust and generalizable approach for the classification of strabismus, achieving a high-test accuracy of 96.33%. This demonstrates superior performance compared to traditional methods that rely on sclera segmentation [22], facial symmetry analysis [23], or Hirschberg test automation [24], which typically report lower accuracies ranging between 94% and 96.6%. While these conventional approaches often operate on narrow input modalities and limited datasets, our model introduces a more holistic diagnostic strategy by combining localization cues and image-based features focused on ocular alignment.

Beyond diagnostic accuracy, a critical contribution of this study is its compatibility with intelligent medical systems, paving the way for practical deployment in real-world clinical and remote settings. The lightweight nature of EfficientNet-B0 makes the model suitable for deployment on mobile platforms and embedded devices, enabling on-the-spot screening via smartphone cameras or tablet-integrated ophthalmic tools. Furthermore, the framework is adaptable for cloud-edge collaboration, where image pre-processing and explain ability computations (e.g., Grad-CAM) can be handled on edge devices for fast response, while more intensive training and model updates are handled on secure cloud servers. This division enhances both responsiveness and scalability in smart tele ophthalmology applications. Additionally, the model's interpretability via Grad-CAM (and benchmarked against LIME/SHAP) supports clinician trust and transparency—crucial for integration into regulatory-compliant diagnostic systems. With further optimization and real-time tuning, the proposed framework holds promise for incorporation into AI-assisted ophthalmology workstations, mobile diagnostic kiosks, or home-based screening systems, particularly in underserved or remote regions.

The strength of this work is that the proposed study will use an EfficientNet-B0 architecture with squeeze-and-excitation blocks that enhance feature extraction and promote better generalization. Combining this with an augmented data-preprocessing pipeline introduces variability and robustness, which will enable this model to perform well in the classification of five categories: esotropia, exotropia, hypertropia, hypotropia, and normal alignment. The testing results further verify the reliability of the proposed model. With high accuracy, recall, and F1 scores for all classes in this work, it is established that the model is indeed capable of differentiating various forms of strabismus with minimal misclassification. Compared to the state-of-the-art literature, this model has achieved a slightly higher accuracy of 96.33%, coupled with enhanced applicability, where models such as the FedCNN-XGboost approach [25] have reported an accuracy of 95.2% and the facial symmetry analysis method [23] achieved 95%. Improvement in the performance may be credited to the novel use of EfficientNet-B0 and rigorous data augmentation techniques used while pre-processing.

Even with these strengths, the study does have its limitations. One of the limitations concerns the use of labelled datasets, which may be biased depending on the diversity and quality of the data. The augmented data pre-processing pipeline discussed in this work mitigates these issues to some extent; however, variations in imaging

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

conditions or patient demographics might still affect the performance of the model. While its reliance on interpretability befalls it nicely concerning Grad-CAM, sometimes this only accounts for visual explanation and not its clinical validation, which would really be required to be replicated for real-world implementation. Future studies will be directed at the expansion of the dataset to involve more diverse populations of patients and image conditions, ensuring strength and generalizability across settings. Further work incorporating multimodal data, such as imaging coupled with patient history or other diagnostic modalities, may further improve the model's diagnostic capabilities. Clinical deployment and testing of the model would be very informative to better understand the practical utility and allow iterative improvements in the model. Further interpretability techniques, like SHAP or LIME beyond CAM, may give more complete explanations of the decisions by the model.

6. Conclusion

In summary, the proposed model of strabismus classification has exhibited outstanding performance in accuracy, with a testing accuracy of 96.33% and outperforming the state-of-the-art methods in the literature. It leverages the EfficientNet-B0 architecture along with squeeze-and-excitation blocks, coupled with Grad-CAM-based visual explanations and robust data augmentation techniques, for effective classification of five different types of strabismus while providing insights into its decision-making process. These developments not only increase the reliability of AI-driven diagnostics in strabismus but also give a framework for translating explainable AI into clinical practice. Though there are residual issues concerning dataset diversity and real-world validation, this study forms a very strong foundation for future research to improve automated diagnosis of strabismus and extend practical applicability to diverse healthcare settings.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] J. Shah and S. Patel, "Strabismus: Symptoms, pathophysiology, management, and precautions," *International Journal of Science and Research*, vol. 4, no. 7, pp. 1510–1514, 2015.

[2] V. E. Das, "Strabismus and the oculomotor system: insights from macaque models," *Annual Review of Vision Science*, vol. 2, no. 1, pp. 37–59, 2016.

[3] H. B. McBain et al., "The impact of strabismus on quality of life in adults with and without diplopia: a systematic review," *Survey of Ophthalmology*, vol. 59, no. 2, pp. 185–191, 2014.

[4] T. L. Huang and S. L. Pineles, "Strabismus and pediatric psychiatric illness: A literature review," *Children*, vol. 10, no. 4, p. 607, 2023.

[5] M. G. Sawyer et al., "Health-related quality of life of children and adolescents with mental disorders," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 41, no. 5, pp. 530–537, 2002.

[6] G. E. Arblaster et al., "Patient perspectives on their outcomes from strabismus surgery undertaken for psychosocial reasons," *Eye*, vol. 38, no. 15, pp. 2926–2931, 2024.

[7] M. D. Pinazo-Durán et al., "Eclectic ocular comorbidities and systemic diseases with eye involvement: a review," *BioMed Research International*, vol. 2016, no. 1, Article ID 6215745, 2016.

[8] H. B. Burch et al., "Management of thyroid eye disease: a consensus statement by the American Thyroid Association and the European Thyroid Association," *European Thyroid Journal*, vol. 11, no. 6, 2022.

[9] E. O. Schotthoefer and D. K. Wallace, "Strabismus associated with thyroid eye disease," *Current Opinion in Ophthalmology*, vol. 18, no. 5, pp. 361–365, 2007.

[10] M. Nowak et al., "Differential diagnosis of thyroid orbitopathy—diseases mimicking the presentation or activity of thyroid orbitopathy," *Endokrynologia Polska*, vol. 75, no. 1, pp. 1–11, 2024.

[11] S. Hull et al., "Tests for detecting strabismus in children aged 1 to 6 years in the community," *Cochrane Database of Systematic Reviews*, no. 11, 2017.

[12] E. Grudzińska et al., "Usefulness assessment of automated strabismus angle measurements using innovative Strabiscan device," *Journal of Clinical Medicine*, vol. 13, no. 4, p. 1067, 2024.

[13] C. Ancona et al., "Stereo tests as a screening tool for strabismus: which is the best choice?," *Clinical Ophthalmology*, pp. 2221–2227, 2014.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025

[14] C. Mestre et al., "An automated and objective cover test to measure heterophoria," *PLoS ONE*, vol. 13, no. 11, Article ID e0206674, 2018.

[15] M. Cantó-Cerdán et al., "Comparative analysis of strabismus measurement using a video oculagraphy system and alternate prism cover test," *Asia-Pacific Journal of Ophthalmology*, vol. 12, no. 6, pp. 582–590, 2023.

[16] B. T. Barrett, "Assessment of binocular vision and accommodation," in *Clinical Procedures in Primary Eye Care*, WB Saunders, pp. 147–208, 2014.

[17] L. Abdu, "Clinical evaluation of horizontal pediatric strabismus and the management challenges," in *Frontiers in Ophthalmology and Ocular Imaging*, IntechOpen, 2019.

[18] C. Yi et al., "Advances in artificial intelligence in thyroid-associated ophthalmopathy," *Frontiers in Endocrinology*, vol. 15, Article ID 1356055, 2024.

[19] G. W. Armstrong and A. C. Lorch, "AI (eye): a review of current applications of artificial intelligence and machine learning in ophthalmology," *International Ophthalmology Clinics*, vol. 60, no. 1, pp. 57–71, 2020.

[20] J. P. O. Li et al., "Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective," *Progress in Retinal and Eye Research*, vol. 82, Article ID 100900, 2021.

[21] J. D. S. de Almeida et al., "Computational methodology for automatic detection of strabismus in digital images through Hirschberg test," *Computers in Biology and Medicine*, vol. 42, no. 1, pp. 135–146, 2012.

[22] T. de Oliveira Simoes et al., "Automatic ocular alignment evaluation for strabismus detection using U-Net and ResNet networks," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 239–244, IEEE, October 2019.

[23] S. M. Jung et al., "Strabismus classification using face features," in *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*, pp. 1–4, IEEE, August 2019.

[24] Ş. Karaaslan et al., "A new method based on deep learning and image processing for detection of strabismus with the Hirschberg test," *Photodiagnosis and Photodynamic Therapy*, vol. 44, Article ID 103805, 2023.

[25] A. Jabbar et al., "A retinal detachment based strabismus detection through FEDCNN," *Scientific Reports*, vol. 14, no. 1, Article ID 23255, 2024.

[26] A. Ananthamoorthy, "Strabismus dataset [Data set]," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/ananthamoorthya/strabismus. [Accessed: Dec. 23, 2024].

[27] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, May 2019.

[28] S. Nigam et al., "EfficientNet architecture and attention mechanism-based wheat disease identification model," *Procedia Computer Science*, vol. 235, pp. 383–393, 2024.

[29] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

DOI: https://doi.org/10.54216/FPA.210101
Received: February 25, 2025 Revised: May 31, 2025 Accepted: July 02, 2025