Jurnal Teknologi Informasi, Komputer dan Aplikasinya (JTIKA)
Vol. 8, No. 1, Maret 2026, (Terakreditasi Sinta-4, SK No:164/E/KPT/2021) ISSN:2657-0327

EYE DISEASE CLASSIFICATION USING DEEP LEARNING: A COMPARATIVE STUDY OF MOBILENETV2, XCEPTION, AND EFFICIENTNET-B0

(*Klasifikasi Penyakit Mata Menggunakan Deep Learning: Studi Perbandingan MobileNetV2, Xception, dan EfficientNet-B0*)

Latifa Zahra Agustini*[1], Fitri Bimantoro[1], Ramaditia Dwiyansaputra[1]
[1]Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: latifazahraa18@gmail.com, [bimo, rama]@unram.ac.id

Abstract

This study presents a comparative analysis of three convolutional neural network (CNN) architectures—MobileNetV2, Xception, and EfficientNet-B0—for classifying retinal fundus images into four categories: Cataract, Diabetic Retinopathy, Glaucoma, and Normal. Using a dataset of 4,217 images, the models were trained with transfer learning, image augmentation, and regularization techniques, and evaluated through 5-fold cross-validation. EfficientNet-B0 achieved the highest mean accuracy (0.85) and demonstrated stable performance across all metrics, while MobileNetV2 provided competitive accuracy with lower computational requirements, making it suitable for resource-limited environments. Xception showed the lowest and least stable performance, indicating a higher tendency to overfit. External validation with clinical images revealed a significant drop in accuracy for all models, highlighting challenges related to domain shift and limited generalization. Grad-CAM analysis also showed difficulties in detecting subtle pathological features in Diabetic Retinopathy and Glaucoma. The study is limited by the small dataset size, reliance on a single data source, and the absence of additional clinical information. Future work should incorporate larger and more diverse datasets, apply domain adaptation strategies, and integrate multimodal clinical data to enhance robustness and clinical applicability.

Keywords: Fundus Classification, MobileNetV2, Xception, EfficientNet-B0, Transfer Learning.

**Corresponding Author*

1. INTRODUCTION

Vision impairment remains one of the major global health burdens. According to the World Health Organization (WHO), at least 2.2 billion people worldwide live with visual impairment or blindness, and 1 billion of these cases could have been prevented or treated with proper diagnosis and timely management [1]. The prevalence continues to rise, particularly in low- and middle-income countries where access to ophthalmic healthcare services is limited. In Indonesia, disparities in the distribution of ophthalmologists are particularly striking—approximately 59% of specialists are concentrated on Java Island, while many outer regions face severe shortages of medical personnel [2]. These inequalities hinder early detection and timely treatment of eye diseases, especially in rural and underserved communities.

Among the major causes of preventable blindness are diabetic retinopathy, glaucoma, and cataract [3], [4], [5]. These diseases often progress silently in their early stages and frequently remain undetected without regular screening [3]. Retinal fundus imaging is a widely used, non-invasive diagnostic technique that allows clinicians to identify early pathological changes in the retina, optic disc, and microvasculature. However, manual interpretation of fundus images is time-consuming, highly dependent on clinician expertise, and subject to inter-observer variability [3], [4].

To address these challenges, deep learning—particularly Convolutional Neural Networks (CNN)—has emerged as a powerful tool capable of automatically extracting complex visual patterns from medical images. CNN have demonstrated strong performance in classifying retinal abnormalities and have become a foundation for automated screening systems in ophthalmology [3], [4], [6]. Various

architectures such as MobileNetV2, Xception, and EfficientNet-B0 have been widely adopted for fundus image classification due to their efficiency, accuracy, and suitability for real-world deployment. Prior studies have shown that MobileNetV2 provides robust performance with low computational cost [3], [7], Xception effectively captures rich spatial features through depthwise separable convolutions [8], and EfficientNet-B0 achieves strong predictive accuracy through compound scaling despite its relatively small parameter size [9]. Nevertheless, many existing works focus on binary classification or a limited number of disease classes, resulting in a lack of comprehensive multi-class evaluations [6], [9].

Another important requirement in modern AI-based diagnosis is explainability. Techniques such as Grad-CAM enable clinicians to verify whether model predictions correspond to clinically meaningful retinal regions—such as the optic disc for glaucoma or microaneurysms for diabetic retinopathy—thereby improving the transparency and trustworthiness of AI systems [8], [10]. However, not all studies incorporate interpretability analyses or assess the clinical relevance of the generated attention maps, limiting the practical utility of such models [3], [10].

To address these gaps, this study conducts a comprehensive comparison of MobileNetV2, Xception, and EfficientNet-B0 for multi-class classification of retinal fundus images into Cataract, Glaucoma, Diabetic Retinopathy, and Normal categories. The evaluation includes 5-fold cross-validation, external dataset testing, and Grad-CAM visualization to assess not only predictive performance but also clinical interpretability. This research aims to provide a deeper understanding of model behavior, highlight the strengths and limitations of each architecture, and support the development of reliable, interpretable, and efficient AI tools for early detection of eye diseases.

2. LITERATURE REVIEW

Research on automated fundus image classification using deep learning has expanded significantly in recent years. Various Convolutional Neural Network (CNN) architectures—from classical models to lightweight and mobile-optimized designs—have been deployed to identify ocular diseases such as cataract, glaucoma, and diabetic retinopathy. However, despite noticeable progress, existing studies differ in scope, methodology, and evaluation depth, making it necessary to critically synthesize prior work to reveal the current research gaps.

Early studies demonstrated that CNNs consistently outperform traditional image-analysis techniques in extracting retinal features and enabling automated diagnosis. For example, Putri and Rakasiwi employed the VGG-16 architecture for multi-class classification of cataract, glaucoma, and diabetic retinopathy, achieving an accuracy of 88% [5]. While such findings confirm the potential of CNNs for early disease detection, models like VGG-16 remain computationally heavy and less suitable for real-time or resource-limited applications [11]. This limitation creates a need for more efficient architectures that retain high accuracy while minimizing computational cost.

In response to these challenges, several studies explored lightweight CNN models. MobileNetV2, designed for mobile and embedded systems, has been widely adopted due to its efficiency. Indraswari utilized MobileNetV2 for fundus image classification and reported an accuracy of 72% [3], while Huynh achieved an average accuracy of 93.89% in classifying five stages of diabetic retinopathy using transfer learning [12]. These findings highlight MobileNetV2's potential for deployment in low-resource clinical environments.

Xception, another widely used architecture, leverages depthwise separable convolutions to extract fine-grained visual features. Studies such as [7] reported an accuracy of 91.1% for ear disease classification using Xception, indicating its strong feature extraction capabilities. Likewise, [8] applied a modified Xception model for diabetic retinopathy classification and achieved an accuracy of 79.59%. However, the deeper structure of Xception also increases the risk of overfitting, especially when trained on small medical datasets—an ongoing challenge in medical image research.

EfficientNet-B0 has also gained popularity, owing to its compound scaling mechanism that balances depth, width, and resolution. [9] reported 79.22% accuracy when applying EfficientNet-B0 to classify normal, cataract, and glaucoma images . Furthermore, [13] demonstrated that CNN fusion models combining EfficientNet with ResNet50 and DenseNet could reach 92% accuracy and an AUC of 1.00, suggesting that hybrid architectures may further boost diagnostic performance. Nonetheless, most EfficientNet-based studies focus on a limited number of disease categories and rarely examine multi-class classification involving four or more classes.

Additional research emphasizes the importance of preprocessing and augmentation strategies. [14] showed that background removal and data

augmentation substantially improved diabetic retinopathy detection. Such results underscore the influence of data quality and diversity on CNN performance, particularly given the scarcity of large annotated medical datasets.

Despite these advances, existing studies still show several limitations: most focus on binary or three-class classification, lack a unified evaluation framework across different CNN architectures, rarely include external validation, and provide limited analysis of model interpretability. To address these gaps, this study conducts a controlled comparison of MobileNetV2, Xception, and EfficientNet-B0 using identical preprocessing, augmentation, and transfer-learning settings across four disease categories. In contrast to prior work, this study incorporates Grad-CAM and external dataset evaluation to assess not only accuracy but also clinical relevance. This integrated approach provides a clearer and more comprehensive understanding of each model's strengths and constraints in real-world fundus image classification.

2.1. Basic Teory

In this study, the authors used several basic theories to support the research to be conducted:

2.1.1. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a deep learning architecture specifically designed for image processing tasks. CNN consists of several layers, including convolutional layers to extract spatial features, pooling layers to reduce data dimensionality, and fully connected layers to perform classification. CNN can automatically learn visual patterns such as shape, color, and texture without manual feature engineering, making it well-suited for classifying fundus retinal images in medical diagnoses [15].

2.1.2. MobileNetV2

MobileNetV2 is a lightweight CNN architecture that employs inverted residual blocks and depthwise separable convolutions to reduce computational complexity and parameter size. This model is designed for resource-constrained devices such as mobile platforms but remains effective in classifying medical images, including fundus retinal images [3].

2.1.3. Xception

Xception is an advanced CNN architecture developed from Inception, which fully utilizes depthwise separable convolutions. It features a deeper and more complex structure than MobileNetV2 and is capable of capturing more detailed image features. However, it also has a higher risk of overfitting, especially when trained on limited datasets [8].

2.1.4. EfficientNet-B0

EfficientNet-B0 is a CNN architecture that introduces a compound scaling method to balance depth, width, and resolution, achieving higher accuracy with fewer parameters compared to conventional models [11]. With only about 5.3 million parameters, it is lightweight yet effective for medical image classification.

2.1.5. K-Fold

K-Fold Cross-Validation is a resampling technique commonly used to evaluate machine learning models by dividing the dataset into k equal subsets, training the model on k – 1 folds, and validating on the remaining fold, with the process repeated until each fold serves as validation. The final performance is reported as the average across folds, providing a more stable and unbiased estimate compared to a single train–test split [16]. This method is particularly beneficial in medical imaging, where datasets are often limited and imbalanced, as it maximizes data utilization while reducing the risk of overfitting [17].

2.1.6. Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is one of the most widely used Explainable AI (XAI) techniques to provide visual explanations of deep learning models. By generating a heatmap that highlights the regions of the image most relevant to the prediction, Grad-CAM enables researchers and clinicians to verify whether the model focuses on meaningful clinical features rather than irrelevant artifacts [18]. In the context of retinal fundus image classification, Grad-CAM has been applied to visualize areas such as the optic disc, blood vessels, or microaneurysms, thereby improving interpretability and clinical trust in AI-based diagnosis. In this study, Grad-CAM was also employed to visualize the retinal regions learned by MobileNetV2, Xception, and EfficientNet-B0, providing insights into the decision-making process of the models.

3. RESEARCH METHODOLOGY

The main objective of this study is to compare the performance of three CNN architectures—MobileNetV2, Xception, and EfficientNet-B0—in classifying retinal fundus images. The implementation was carried out locally using Python in Visual Studio Code with TensorFlow and Keras. The primary dataset used for training and K-Fold Cross Validation was obtained from Kaggle (Eye Diseases Classification)

(https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification), consisting of 4,217 images across four categories: Cataract, Diabetic Retinopathy, Glaucoma, and Normal.

To assess the generalization capability of the models, an external dataset was included and evaluated after the K-Fold process and before the Grad-CAM visualization. This external dataset comprises 398 images obtained from Rumah Sakit Mata Nusa Tenggara Barat, consisting of 231 Cataract, 112 Diabetic Retinopathy, 23 Glaucoma, and 32 Normal cases. All hospital images were anonymized and reprocessed through resizing, normalization, and illumination adjustment to ensure consistency with the main dataset without altering their pathological characteristics.

To further complete the class representation in the external evaluation, an additional 208 Glaucoma images from the SMDG dataset available on Kaggle (https://www.kaggle.com/datasets/deathtrooper/multichannel-glaucoma-benchmark-dataset?select=full-fundus) were included, along with 199 Normal images and 119 Diabetic Retinopathy images from Mendeley Data(https://data.mendeley.com/datasets/nxcd8krdhg) . All external datasets were processed using the same preprocessing pipeline as the main dataset, but without data augmentation, as they were used exclusively for model evaluation.

The complete research workflow—from model training, K-Fold validation, and external dataset evaluation, to Grad-CAM interpretation—is illustrated in Figure 1.

The dataset was divided using 5-fold cross-validation, where in each iteration four folds were used for training and one fold for validation. Data augmentation (Table I) was applied only to the training portion to improve model generalization. Each fold underwent model training, validation, and performance evaluation, producing metrics such as accuracy, precision, recall, and F1-score.

After completing all five folds, the results were averaged to obtain the final performance for each model. The models were then tested on an external dataset to assess generalizability. Finally, Grad-CAM visualization was generated to highlight the image regions that contributed most to the model's predictions, leading to the final reported results.

Figure 1. Research Flow

TABLE I. AUGMENTATION TABLE

Parameter Value	Values
Rescale	1./255
Rotation range	30
Width shift range	0.2
Height shift range	0.2
Shear range	0.2
Zoom range	0.2
Horizontal flip	True

The three CNN models—MobileNetV2, Xception, and EfficientNet-B0—were initialized with pre-trained ImageNet weights and employed in a transfer learning configuration in which all base convolutional layers were kept frozen. Only the additional top layers, including the classification head, were trained to adapt the models to the four-class classification task (Cataract, Glaucoma, Diabetic Retinopathy, and Normal). This feature-extraction strategy was chosen to ensure stable optimization on a limited dataset, reduce overfitting risk, and lower computational cost compared with full fine-tuning [19], [20], [21]. Freezing the backbone across all models also maintains methodological consistency, ensuring that performance differences reflect the intrinsic architectural characteristics rather than discrepancies in fine-tuning procedures. Accordingly, no fine-tuning was applied to any of the three models; all backbones

were consistently kept frozen to ensure a fair and controlled comparison.

Training for all three models employed Early Stopping with a maximum of 25 epochs. The callback monitored validation loss and halted training when no further improvement occurred, preventing unnecessary computation and reducing overfitting. This ensured that each model was trained only to its optimal generalization point while maintaining fairness across all folds.

The full structural configuration of each model is presented in Tables II, III, and IV; these include the division between frozen and trainable components as well as the classification layers, which incorporate BatchNormalization and Dropout to stabilize feature distributions and improve generalization.

TABLE II. MOBILENET-V2 MODEL ARCHITECTURE

Layer Type	Description	Output
Base	MobileNetV2_ 1.00_224	(7, 7, 1280)
GlobalAveragePooling 2D	Global average of each feature	(1280)
BatchNormalization	Normalization of base output	(1280)
Dense	1024 units, activation: ReLU	(1024)
BatchNormalization	Normalization of dense output	(1024)
Dropout	Prevent overfitting	(1024)
Dense	512 units, activation: ReLU	(512)
BatchNormalization	Normalization of dense output	(512)
Dropout	Prevent overfitting	(512)
Dense (Output)	4 units, activation: Softmax	(4)

TABLE III. XCEPTION MODEL ARCHITECTURE

Layer Type	Description	Output
Base	Xception	(7, 7, 2048)
GlobalAveragePooling 2D	Global average of each feature	(2048)
BatchNormalization	Normalization of base output	(2048)
Dense	1024 units, activation: ReLU	(1024)
BatchNormalization	Normalization of dense output	(1024)
Dropout	Prevent overfitting	(1024)
Dense	512 units, activation: ReLU	(512)
BatchNormalization	Normalization of dense output	(512)
Dropout	Prevent overfitting	(512)
Dense (Output)	4 units, activation: Softmax	(4)

TABLE IV. EFFICIENTNET-B0 MODEL ARCHITECTURE

Layer Type	Description	Output
Base	EfficientNetB0	(7, 7, 1280)
GlobalAveragePooling 2D	Global average of each feature	(1280)
BatchNormalization	Normalization of base output	(1280)
Dense	1024 units, activation: ReLU	(1024)
BatchNormalization	Normalization of dense output	(1024)
Dropout	Prevent overfitting	(1024)
Dense	512 units, activation: ReLU	(512)
BatchNormalization	Normalization of dense output	(512)
Dropout	Prevent overfitting	(512)
Dense (Output)	4 units, activation: Softmax	(4)

4. RESULT AND DISCUSSION

In this study, retinal fundus image classification was performed using three deep learning architectures: MobileNetV2, Xception, and EfficientNet-B0. The dataset contains 4,217 JPG images across four classes—Cataract, Diabetic Retinopathy, Glaucoma, and Normal (Figure 2). All images were resized to 224 × 224 pixels with three RGB channels. Data augmentation was applied to increase training diversity, as summarized in Table I, with sample results shown in Figure 3.

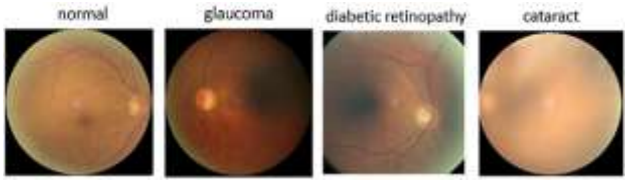


Figure 2. Category of Retinal Fundus Image Dataset

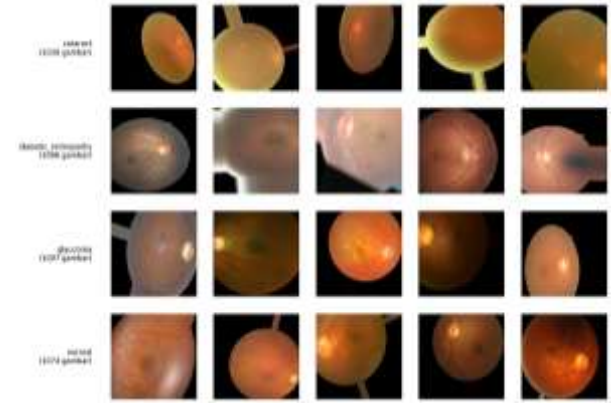


Figure 3. Dataset Augmentation Results

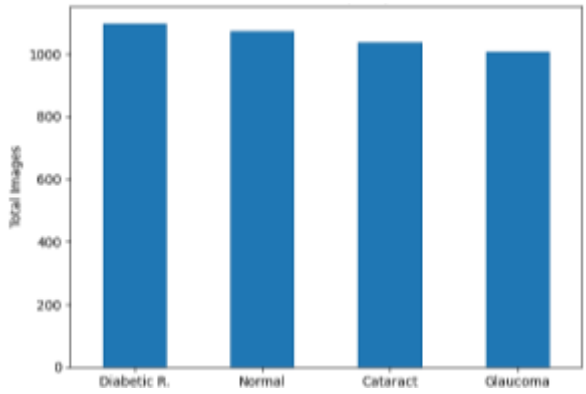


Figure 4. Distribution of Samples per Class

Figure 4 presents the class distribution of the main dataset. With relatively similar sample counts across the four classes—Cataract (1,038), Diabetic Retinopathy (1,098), Glaucoma (1,007), and Normal (1,074)—the dataset can be considered well-balanced. This balance minimizes class bias during the K-Fold Cross-Validation procedure.

Figure 5 illustrates the validation loss curves of MobileNetV2 across five folds. Although the early epochs show noticeable fluctuations, all folds exhibit a clear downward trend as training progresses. By approximately epoch 10 onward, the validation loss stabilizes in the range of 0.80–0.90, indicating that the model gradually converges despite initial variability. Differences between the folds reflect natural variation in validation subsets, yet the overall pattern demonstrates stable learning behavior and no signs of overfitting, as validation loss continues to decrease or remain steady rather than diverging.

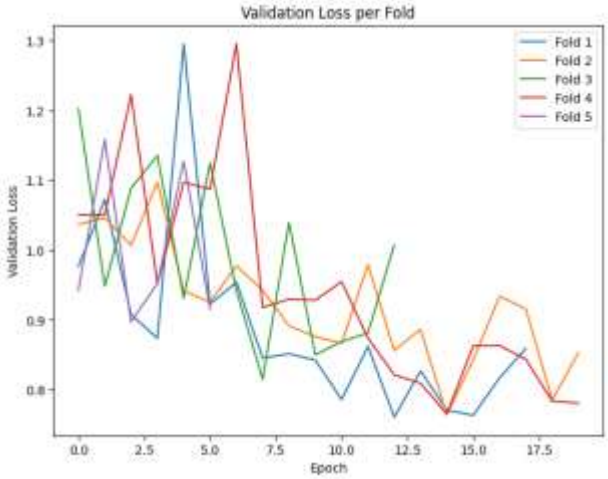


Figure 5. Validation Loss Result of MobileNet-V2 Model

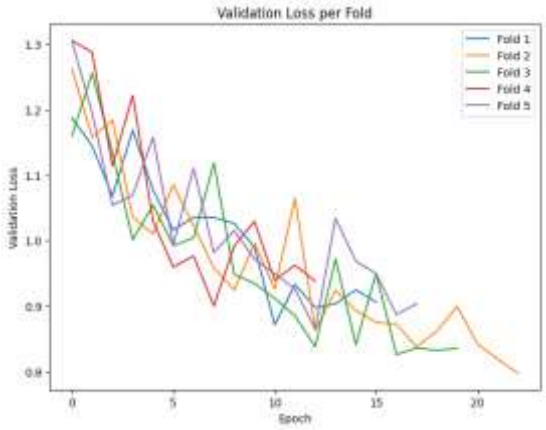


Figure 6. Validation Loss Result of Xception Model

Figure 6 presents the validation loss curves of the Xception model across five folds. The curves show substantial fluctuations in the early epochs, with Fold 3 and Fold 4 exhibiting the highest variability. Despite this instability, all folds display a clear downward trend, and the validation loss gradually stabilizes around 0.85–1.0 after approximately epoch 10. Compared with MobileNetV2, Xception demonstrates higher volatility and slower convergence, reflecting its deeper architecture and greater sensitivity to limited data. Although the loss continues to decrease without signs of divergence or overfitting, the inconsistent fold-to-fold behavior suggests that Xception may require stronger regularization or additional data to achieve more stable generalization.

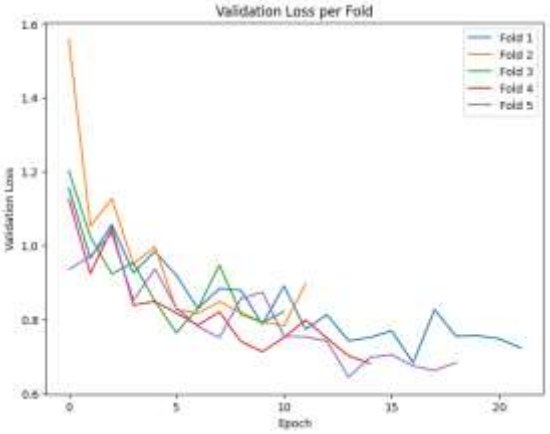


Figure 7. Validation Loss Results of EfficientNet-B0 Model

Figure 7 shows the validation loss curves of EfficientNet-B0 across five folds. Despite an initially high loss in Fold 2, all folds exhibit a clear and consistent downward trend throughout training. After approximately epoch 8, the curves begin to stabilize within the 0.70–0.85 range, with Fold 3 and Fold 4 showing the smoothest convergence. Variations across folds remain relatively small compared with Xception, indicating better stability and more reliable generalization. The absence of upward divergence suggests that EfficientNet-B0 does not experience overfitting and adapts well to the dataset, benefiting from its balanced architecture and efficient parameter scaling.

To determine the performance of the model in classifying eye diseases, an evaluation is carried out using the confusion matrix with the results as shown in Figure 8 for MobileNet-V2 Confusion Matrix Testing Results, Figure 9 for Xception Confusion Matrix Testing Results, and Figure 10 for EfficientNet-B0 Confusion Matrix Testing Results.

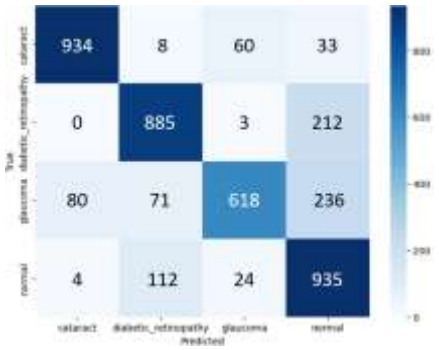


Figure 8. MobileNet-V2 Confusion Matrix Testing Results

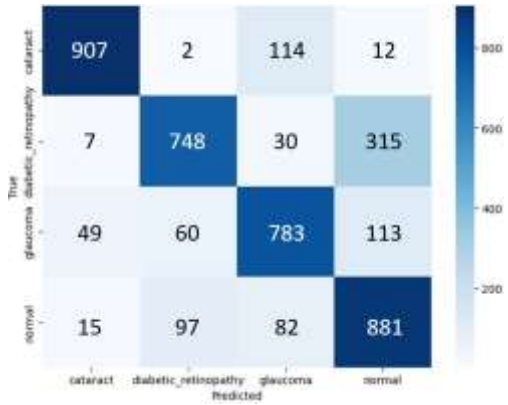


Figure 9. Xception Confusion Matrix Testing Results

Figure 10.EfficientNet-B0 Confusion Matrix Testing Results

In order to provide a comprehensive assessment of model performance, the following evaluation metrics were calculated for each model, with the detailed results presented in Table V (MobileNetV2), Table VI (Xception), and Table VII (EfficientNet-B0).

TABLE V. MOBILENET-V2 ARCHITECTURE TESTING RESULTS

Fold	Accuracy	Precision	Recall	F1 Score
1	0.82	0.85	0.82	0.82
2	0.81	0.83	0.81	0.80
3	0.79	0.81	0.79	0.79
4	0.81	0.82	0.81	0.81
5	0.76	0.79	0.76	0.77
Mean	0.80	0.82	0.80	0.80

TABLE VI. XCEPTION ARCHITECTURE TESTING RESULTS

Fold	Accuracy	Precision	Recall	F1 Score
1	0.79	0.79	0.79	0.79
2	0.79	0.80	0.79	0.79
3	0.80	0.82	0.80	0.80
4	0.79	0.80	0.79	0.79
5	0.78	0.78	0.78	0.78
Mean	0.79	0.80	0.79	0.79

TABLE VII. EFFICIENTNET-B0 ARCHITECTURE TESTING RESULTS

Fold	Accuracy	Precision	Recall	F1 Score
1	0.86	0.87	0.86	0.86
2	0.82	0.85	0.82	0.82
3	0.83	0.84	0.83	0.83
4	0.86	0.87	0.86	0.86
5	0.88	0.89	0.88	0.88
Mean	0.85	0.86	0.85	0.85

The evaluation of MobileNetV2, Xception, and EfficientNet-B0 on retinal fundus images for four eye disease classes showed clear performance differences. EfficientNet-B0 achieved the highest accuracy of 0.85 with balanced precision, recall, and F1-score (Table VII). MobileNetV2 reached 0.80 accuracy with stable performance and low validation loss, suitable for lightweight deployment (Table V). Xception, with 0.79 accuracy, exhibited greater fluctuations and a higher risk of overfitting (Table VI).

To ensure that the performance differences among the models were not caused by random variation, a paired t-test was conducted using the accuracy values obtained from the 5-fold cross-validation. The results of the three pairwise model comparisons are presented in Table VIII.

TABLE VIII. RESULTS OF THE PAIRED T- TEST

Model Comparison	t-Statistic	p-Value	Significance (α = 0.05)
MobileNetV2 vs Xception	0.83	0.46	Not significant
MobileNetV2 vs EfficientNet-B0	-2.85	0.05	Significant
EfficientNet-B0 vs Xception	4.47	0.01	Significant

The paired t-test results indicate a clear significant difference among the three models. The comparison between MobileNetV2 and Xception yielded a p-value of 0.45 (> 0.05), indicating no significant difference between these two models. In contrast, the comparison between MobileNetV2 and EfficientNet-B0 resulted in a p-value of 0.05, demonstrating a significant difference at α = 0.05, with EfficientNet-B0 exhibiting superior performance. The comparison between EfficientNet-B0 and Xception produced a p-value of 0.01, which is also significant and further confirms the superiority of EfficientNet-B0. Overall, these results suggest that EfficientNet-B0 performs statistically better than the other two models, while the difference between MobileNetV2 and Xception is not significant.

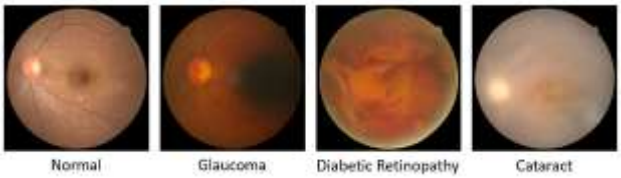


Figure 11.Sample Images From The External Validation Dataset

The next step involves testing the models on this external dataset to evaluate their generalization performance. Figure 11 shows sample images from the dataset, which includes 231 images for each of the four classes (Cataract, Diabetic Retinopathy, Glaucoma, and Normal) and were not used during training. Since the dataset is balanced across all classes, only accuracy was calculated to evaluate overall performance.

The performance of the models in classifying eye diseases on the external dataset is presented through confusion matrices: Figure 12 for MobileNet-V2, Figure 13 for Xception, and Figure 14 for EfficientNet-B0. Additionally, the accuracy results on the external dataset are summarized in Table IX, providing a comprehensive comparison of the overall classification performance of each model.

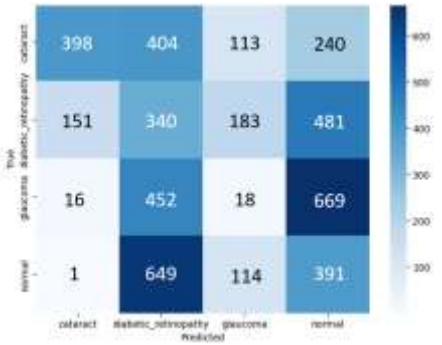


Figure 12.MobileNet-V2 Confusion Matrix on External Dataset Testing Results

Figure 13.Xception Confusion Matrix on External Dataset Testing Results

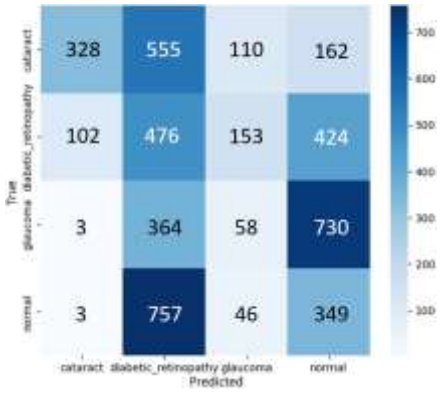


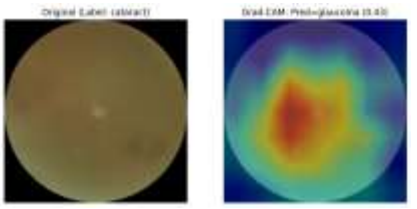
Figure 14.EfficientNet-B0 Confusion Matrix on External Dataset Testing Results

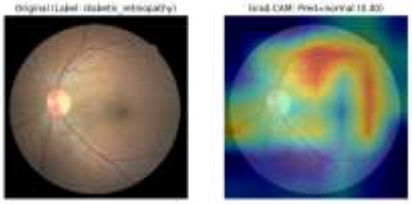
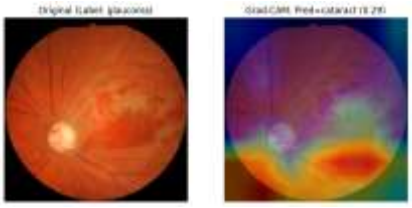
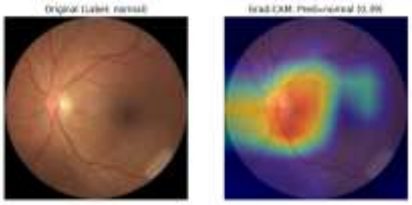
TABLE IX. EXTERNAL DATASET ACCURACY RESULTS

MobileNet-V2	Xception	EfficientNet-B0
0.248	0.263	0.262

On the external dataset, the three models demonstrated relatively similar performance, with accuracies of 0.248 for MobileNet-V2, 0.263 for Xception, and 0.262 for EfficientNet-B0. Compared to the internal testing results, where EfficientNet-B0 achieved the highest accuracy (0.85), the decrease in performance indicates the challenge of handling new data, likely due to differences in data distribution, lighting conditions, image quality, or representation of disease classes. In terms of model analysis, Xception showed a slight advantage with the highest accuracy on the external dataset, indicating strong generalization ability. MobileNet-V2 had relatively lower performance, suggesting greater sensitivity to variations in new data. EfficientNet-B0 remained consistently strong, though slightly below Xception on the external dataset, reflecting its optimal performance on data similar to the training set but slightly less flexible to new variations. Overall, all three models were capable of accurately classifying eye diseases, with Xception excelling in adaptability to external data, MobileNet-V2 being more sensitive to variations, and EfficientNet-B0 maintaining high stability on internal data.

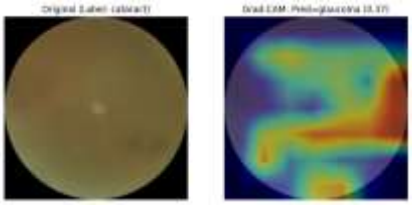
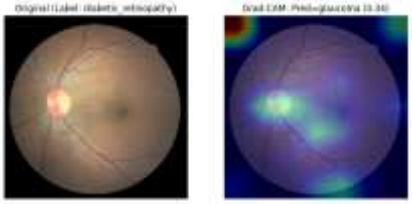
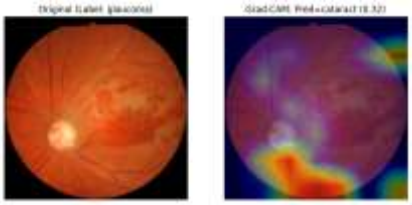
TABLE X. GRAD CAM RESULTS OF MOBILENET-V2

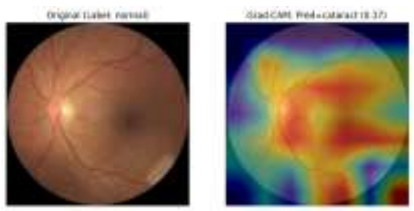
Class	Grad-CAM
Cataract	

Class	Grad-CAM
Diabetic Retinopathy	
Glaucoma	
Normal	

The Grad-CAM visualization for MobileNet-V2 (Table X) shows a strong focus on the opacity regions in Cataract cases, whereas the model's attention is dispersed or less precise in early-stage Diabetic Retinopathy and Glaucoma cases, indicating limitations in detecting subtle pathological features. For the Normal class, the model's focus is relatively accurate, although not as pronounced as in Cataract cases.

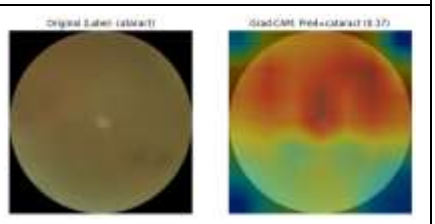
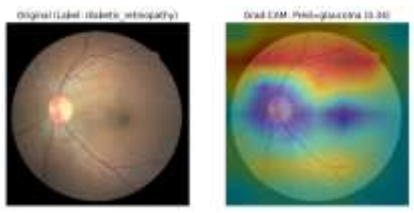
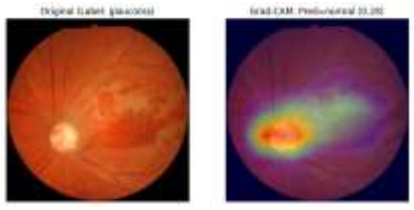
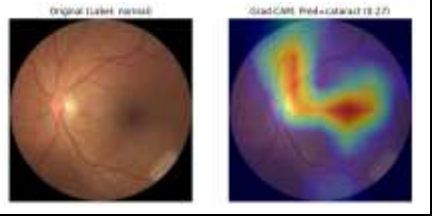
TABLE XI. GRAD CAM RESULTS OF XCEPTION

Class	Grad-CAM
Cataract	
Diabetic Retinopathy	
Glaucoma	

Class	Grad-CAM
Normal	

Grad-CAM heatmaps for the Xception model (Table XI) show that while the model produces clear and focused activations in straightforward cases such as cataract, it frequently highlights peripheral and clinically irrelevant regions in Diabetic Retinopathy and Normal samples. This indicates inconsistent internal feature representation and suboptimal localization of disease-specific regions of interest (ROIs).

TABLE XII. GRAD CAM RESULTS OF EFFICIENTNET-B0

Class	Grad-CAM
Cataract	
Diabetic Retinopathy	
Glaucoma	
Normal	

Grad-CAM results from EfficientNet-B0 (Table XII) reveal that aside from the Cataract sample, the model frequently focuses on non-diagnostic or peripheral regions in Diabetic Retinopathy, Glaucoma, and Normal cases. These inconsistent attention patterns indicate suboptimal localization of clinically relevant features and correspond to the model's observed misclassifications.

To evaluate model complexity, memory requirements, and both training and inference speed, a comparative analysis of the computational efficiency of the three architectures is provided. A summary of these components is presented in Table XIII.

TABLE XIII. COMPUTATIONAL EFFICIENCY

Metric	MobileNetV2	Xception	EfficientNetB0
Total FLOPs	561,229,824	9,106,909,120	790,365,135
Parameter Count	4,107,844	23,500,844	5,899,431
Trainable	1,844,228	2,632,196	1,844,228
Non-trainable	2,263,616	20,868,648	4,055,203
Model Size	30.07 MB	110.03 MB	22.78 MB
Inference Speed (ms/image)	561.11 ms	998.87 ms	399.89 ms
CPU Memory Used	155.93 MB	163.70 MB	190.94 MB
Training Speed (avg per epoch)	240.79 s	631.29 s	317 s

All measurements were performed in a CPU-only environment without GPU acceleration using TensorFlow 2.15, meaning the reported performance fully reflects CPU execution characteristics. The results show that MobileNetV2 is the most computationally efficient model, with the lowest FLOPs and parameter count, making it suitable for low-resource devices, although its inference speed is still slower than EfficientNetB0. Xception exhibits the highest complexity and memory usage, resulting in the slowest training and inference times and making it less practical for CPU-based deployment. EfficientNetB0 provides the best balance: slightly heavier than MobileNetV2 but delivering the fastest inference while maintaining a compact model size. Overall, under CPU

execution, EfficientNetB0 achieves the most favorable performance–efficiency trade-off.

The comparative evaluation of MobileNetV2, Xception, and EfficientNet-B0 shows that the three architectures exhibit distinct characteristics in accuracy, stability, interpretability, and computational efficiency. EfficientNet-B0 achieves the strongest overall performance, with an average accuracy of 0.85 and balanced precision, recall, and F1-scores (0.85–0.86). Its stable learning curves and lower misclassification rates indicate better adaptability to inter-class variability. Although Grad-CAM visualizations show occasional attention to non-critical regions—particularly for Diabetic Retinopathy and Glaucoma—EfficientNet-B0 still outperforms the other models across all evaluation aspects.

MobileNetV2 attains an accuracy of 0.80 and provides the best computational efficiency due to its lightweight architecture, making it suitable for resource-limited deployment. Grad-CAM results reveal that the model reliably identifies clear pathological cues (e.g., cataract opacity), but struggles with subtle lesions characteristic of early-stage Diabetic Retinopathy and Glaucoma, aligning with its misclassification trends.

Xception yields an accuracy of 0.79 but displays more variability in its loss curves and inconsistent Grad-CAM activation patterns, often focusing on peripheral or clinically irrelevant regions. Its deeper architecture leads to the highest memory usage and computational cost, making it the least efficient model for CPU execution.

The confusion matrix analysis reveals that most misclassifications occur in the Diabetic Retinopathy and Glaucoma classes. In Glaucoma, many samples are predicted as Normal because the cup-to-disc ratio—an essential diagnostic indicator—is often subtle and not clearly visible in fundus images [22]. This pattern is clearly visible in the confusion matrix for Xception (Figure 9), where 113 Glaucoma images are misclassified as Normal, representing the dominant error in this class. When the optic disc contour is insufficiently pronounced, the models fail to distinguish pathological structures from normal retinal patterns. The absence of explicit optic disc segmentation further prevents the extraction of fine-grained structural cues, reducing Glaucoma sensitivity [23].

In Diabetic Retinopathy, misclassifications are primarily caused by small lesions such as microaneurysms or hemorrhages, which exhibit low contrast and are easily obscured by illumination noise [24]. At intermediate stages, abnormal vascular patterns may resemble features of other conditions, such as vessel enlargement in glaucoma, or even appear near-normal when lesions are sparse. This behavior is also reflected in the confusion matrix (Figure 9), where 315 Diabetic Retinopathy images are misclassified as Normal, and 30 as Glaucoma, indicating substantial overlap in feature representations between these categories. Grad-CAM findings reinforce this trend, showing that none of the three models consistently attend to the small lesional regions responsible for early or mid-stage Diabetic Retinopathy, explaining the higher misclassification rates in these two classes [25].

All models show a substantial performance decline on the external dataset, with accuracies ranging from 0.248 to 0.263. Xception's external confusion matrix (Figure 13) again shows misalignment between predicted and true labels, including 370 Glaucoma samples misclassified as Normal and 411 Diabetic Retinopathy samples misclassified as Normal, underscoring the strong impact of domain shift—differences in imaging devices, illumination, resolution, and disease prevalence—on generalization [22]. Paired t-tests indicate no significant difference between MobileNetV2 and Xception, while EfficientNet-B0 differs significantly from both.

All training and inference processes were executed on CPU-only settings to reflect realistic deployment scenarios. MobileNetV2 is the fastest and most lightweight architecture, Xception is the most computationally intensive, and EfficientNet-B0 offers a balanced trade-off between accuracy and efficiency.

Overall, EfficientNet-B0 emerges as the most effective architecture for retinal fundus image classification, MobileNetV2 is the most deployment-efficient model, and Xception requires further optimization—such as selective fine-tuning or enhanced regularization—to improve stability and generalization.

5. CONCLUSION

EfficientNet-B0 demonstrated the best performance in fundus image classification, followed by MobileNetV2, which offers high computational efficiency for lightweight devices, while Xception showed the lowest and least stable performance. However, all models experienced a substantial decrease in accuracy when evaluated on external data, indicating limited generalization due to variations in image quality, distribution, and underlying characteristics across different data sources.

The limitations of this study include the relatively small dataset size, reliance on a single primary data source, non-uniform image quality, and the absence of additional clinical information that could support the classification process. For future development, larger and more diverse datasets from multiple institutions are needed, along with the implementation of domain adaptation techniques to mitigate distributional differences between datasets. Incorporating multimodal data such as OCT images or clinical attributes, as well as exploring full fine-tuning and more robust interpretability methods, may further enhance model performance and readiness for clinical deployment.

ACKNOWLEDGMENT

The author sincerely extends gratitude to all individuals who have contributed to the successful completion of this research. Special appreciation is given to the supervisor for dedicating time and effort to provide guidance and engage in discussions on various aspects of this study, thereby facilitating its successful completion.

REFERENCES

[1] World Health Organization, "World report on vision."

[2] D. S. Saputri, "Jokowi: Kekurangan Dokter Spesialis di Daerah Jadi PR Besar," *Republika*, May 06, 2024.

[3] R. Indraswari, W. Herulambang, and R. Rokhana, "Deteksi Penyakit Mata Pada Citra Fundus Menggunakan Convolutional Neural Network (CNN)," *Techno.COM*, vol. 21, no. 2, pp. 378–389, May 2022.

[4] M. N. I. Muhlashin and A. Stefanie, "KLASIFIKASI PENYAKIT MATA BERDASARKAN CITRA FUNDUS MENGGUNAKAN YOLO V8," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 2, pp. 1363–1368, Apr. 2023.

[5] C. A. Putri and S. Rakasiwi, "Diagnosis Dini Penyakit Mata: Klasifikasi Citra Fundus Retina dengan Convolutional Neural Network VGG-16," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 208–216, Apr. 2025, doi: 10.29408/edumatic.v9i1.29571.

[6] William and C. Lubis, "KLASIFIKASI PENYAKIT MATA MENGGUNAKAN CNN," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 10, no. 1, pp. 1–4, 2022.

[7] L. R. Setiawan, G. Pasek, S. Wijaya, and F. Bimantoro, "Ear Disease Clasification Using Deep Learning with Xception and MobileNet-V2 Architecture," *Jurnal Teknologi Informasi, Komputer dan Aplikasinya*, vol. 6, no. 2, pp. 544–555, Sep. 2024.

[8] S. H. Kassani, P. H. Kassani, and R. Khazaeinezhad, "Diabetic Retinopathy Classification Using a Modified Xception Architecture," IEEE, 2019.

[9] Z. Arif, R. Y. Nur Fu'adah, S. Rizal, and D. Ilhamdi, "Classification of eye diseases in fundus images using Convolutional Neural Network (CNN) method with EfficientNet architecture," *Jurnal Riset Tindakan Indonesia*, vol. 8, no. 1, pp. 125–131, Jul. 2023, doi: 10.29210/30032835000.

[10] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," *Information*, vol. 16, no. 195, pp. 1–28, Mar. 2025.

[11] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif Intell Rev*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020.

[12] H. Nhut Huynh *et al.*, "Classification of Stages Diabetic Retinopathy Using MobileNetV2 Model Kalpa Publications in Engineering," Kalpa Publications in Engineering, 2022.

[13] A. Biswas and R. Banik, "CNN Fusion: A Promising Technique for Ophthalmic Disorder Diagnosis," in *Procedia Computer Science*, Elsevier B.V., 2024.

[14] C. Suedumrong, S. Phongmoo, T. Akarajaka, and K. Leksakul, "Diabetic Retinopathy Detection Using Convolutional Neural Networks with Background Removal, and Data Augmentation," *Applied Sciences (Switzerland)*, vol. 14, no. 19, Oct. 2024.

[15] S. Takahashi *et al.*, "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," *J Med Syst*, vol. 48, no. 1, Dec. 2024.

[16] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," Jul. 01, 2023, *Radiological Society of North America Inc.*

[17] J. M. Gorriz, R. M. Clemente, F. Segovia, J. Ramirez, A. Ortiz, and J. Suckling, "Is K-fold cross validation the best model selection method for Machine Learning?," *A PREPRINT*, Nov. 2024.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Computer Vision Foundation*, pp. 618–626, 2017.

[19] I. Matas *et al.*, "Mitigating Overfitting in Medical Imaging: Self-Supervised Pretraining vs. ImageNet Transfer Learning for Dermatological Diagnosis," May 2025.

[20] R. Shadman, M. G. S. Murshed, E. Verenich, A. Velasquez, and F. Hussain, "The Utility of Feature Reuse: Transfer Learning in Data-Starved Regimes," Dec. 2023.

[21] T. Hwang, H. Seo, J. Jung, and S. Jung, "Exploring Selective Layer Freezing Strategies in Transformer Fine-Tuning: NLI Classifiers with Sub-3B Parameter Models," *Applied Sciences*, vol. 15, no. 19, p. 10434, Sep. 2025, doi: 10.3390/app151910434.

[22] S. Thanapaisal *et al.*, "Machine learning technology in the classification of glaucoma severity using fundus photographs," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-11697-1.

[23] H. N. Veena, A. Muruganandham, and T. Senthil Kumaran, "A novel optic disc and optic cup segmentation technique to diagnose glaucoma using deep learning convolutional neural network over retinal fundus images," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6187–6198, Sep. 2022.

[24] D. Virmani, K. B. Umare, A. Devendran, G. Ravikanth, D. Jamthe, and M. Kejariwal, "AI-Powered Early Detection of Diabetic Retinopathy: A Deep Learning Approach for Improved Clinical Decision-Making".

[25] A. Karthik and S. Mynampati, "Explainable AI for Diabetic Retinopathy Detection Using Deep Learning with Attention Mechanisms and Fuzzy Logic-Based Interpretability."