COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Decision-layer interpretability for CNN-based glaucoma classification via sparse feature selection and ANFIS

Etik Irijanti[a,b], Igi Ardiyanto[a], Hanung Adi Nugroho*[a]

[a]*Department of Electrical and Information Engineering; Universitas Gadjah Mada, Yogyakarta 55281, Indonesia*
[b]*Department of Information Technology; Universitas Muhammadiyah Yogyakarta, Yogyakarta 55183, Indonesia*

Article history:
Received: 24 April 2026 / Received in revised form: 25 June 2026 / Accepted: 25 June 2026

Abstract

Glaucoma is a leading cause of irreversible vision loss, and its early detection remains challenging due to the presence of subtle structural variations in retinal fundus images. Convolutional neural networks (CNNs) have demonstrated strong performance for automated classification of glaucoma; however, the relationship between extracted features and prediction outcomes frequently proves challenging to interpret. Most existing explainable artificial intelligence (XAI) approaches rely on post hoc visualizations, which provide limited insight into the decisions-making process. To address this limitation, this present study proposes a hybrid CNN–feature selection–ANFIS framework (CNN–FS–ANFIS) that integrates interpretability directly within the decision layer. In this framework, the first stage of the process involves adapting a CNN backbone to the glaucoma classification task through the use of transfer learning. This is then used as a fixed feature extractor to obtain retinal representations for decision-layer modeling. Subsequently, a feature selection stage is applied driven by sparsity to construct a compact and structured subset of informative features. These features are then fed into an Adaptive Neuro-Fuzzy Inference System (ANFIS), enabling predictions to be expressed through explicit fuzzy rule-based reasoning. The impact of feature compactness is examined in a controlled experimental setting, where the feature subset size is varied from three to nine. The findings demonstrate that compact feature subsets can achieve consistent and competitive performance. By means of LASSO-selected features, the ANFIS decision layer achieved an AUC of 0.84±0.01, sensitivity of 0.82±0.13, specificity of 0.74±0.10, and an F1-score of 0.79±0.04. Rule-base analysis further exhibited that two- to three-rule ANFIS configurations-maintained AUC values of approximately 0.84 while preserving a transparent and manageable decision structure. The proposed framework, therefore, enables direct analysis of the relationship between selected CNN features, fuzzy rules, and model outputs. This traceable decision pathway has the potential to support more transparent and auditable glaucoma screening systems.

Keywords: Adaptive neuro-fuzzy inference system (ANFIS); Classification; Convolutional neural network (CNN); Explainable AI; Glaucoma

1. Introduction

Glaucoma is a chronic and progressive ocular neuropathy and a leading cause of irreversible blindness worldwide [1–4]. The disease is characterized by gradual structural damage to the optic nerve head, frequently progressing silently until substantial and permanent visual field loss has occurred [5]. It is therefore critical that early detection is prioritized to prevent irreversible vision loss and reduce the global burden of avoidable blindness.

The retinal fundus image is a widely utilized modality for glaucoma classification due to its non-invasiveness, cost-effectiveness, and wide availability [6, 7]. Fundus images have been shown to contain structural features that are indicative of glaucoma, including cupping of the optic disc and thinning of the neuroretinal rim [8, 9]. Previous studies have also explored optic cup segmentation from retinal fundus images using adaptive thresholding and morphological image processing, further supporting the relevance of optic-disc and optic-cup analysis for glaucoma-oriented retinal image processing [10]. Fundus cameras are more readily deployed in primary and secondary healthcare in comparison to more specialized diagnostic instruments such as optical coherence tomography (OCT) or perimetry. These features make fundus imaging suitable for automated analysis and the implementation of large-scale automated systems [11, 12].

*Corresponding Author.
Email: adinugroho@ugm.ac.id
https://doi.org/10.21924/cst.11.1.2026.1968

Recent advances in deep learning have enabled automated analysis of retinal images by learning hierarchical features. Convolutional neural networks (CNNs) have been successfully employed to learn discriminative features from fundus images for the purpose of glaucoma classification [13, 14]. Different CNN architectures including VGG, ResNet, DenseNet, and EfficientNet have demonstrated the capacity to learn structural patterns associated with glaucoma-related changes [15–18]. However, it is important to note that most of the CNN models are regarded as black-box systems that provide limited explanations on the conversion of learned representations into classification decisions. This paucity of transparency remains a major barrier to the broader application of deep learning models in image analysis tasks.

Several studies have investigated glaucoma classification through deep learning approaches on publicly available retinal datasets. The PAPILA dataset was initially presented by Kovalyk et al. [19] and contains retinal fundus images, clinical data, and accurate annotations of the optic disc and optic cup.

Advanced research has explored a range of modeling strategies, including ensemble learning, regression-based models, and architectural modifications, purposely to address challenges specific to particular dataset, such as class imbalance and small sample sizes [7, 20–22]. While these works highlight the potential of glaucoma classification tasks, their main focus is on enhancing predictive performance, rather than explicitly structuring and explaining decisions.

Deep learning models have limitations in interpretability. To address these challenges, many explainable artificial intelligence (XAI) techniques have been employed, including saliency maps [23], and class activation maps, [24]. In medical image analysis, CAM-based explanation methods have also been evaluated with deep segmentation models with the objective of enhancing the transparency and understandability of model decisions [25]. However, these approaches commonly do not explicitly model the underlying decision-making process and instead offer post hoc visual explanations. Thus, the relationship between the extracted features and the ultimate prediction is characterized by a lack of structure and is challenging to analyze systematically. This limitation highlights the necessity to move beyond post hoc explanation methods and instead focus on the structural formation of decisions within the model. The process of feature selection provides an effective and systematic method to restrict the high-dimensional feature space induced by CNN representations [26]. The process of feature selection has been shown to reduce redundancy and preserve a compact subset of useful features, thereby enabling the construction of a compact and structured decision space, and providing an explicitly interpretable decision layer.

Fuzzy inference systems are among the interpretable machine learning methods that inherently and systematically provide a transparent basis for decision modeling by expressing information in terms of explicit rule-based structures. Adaptive Neuro-Fuzzy Inference Systems (ANFIS) effectively integrate the principles of neural network and fuzzy rule-based reasoning. This enables the implementation of data-driven models in practice while maintaining interpretable decision rules [27–29]. The paradigm is also supported by previous work on deep convolutional fuzzy classifiers, which demonstrates that the synergy of deep feature extraction and fuzzy rule-based reasoning can result in inherently interpretable models [30]. However, in the context of high-dimensional feature representations, the implementation of such approaches may introduce increased rule-based complexity, potentially reducing the clarity and manageability of the resulting decision structure. In this setting, ANFIS is especially well-suited as a decision-layer model in conjunction with compact feature representations. This allows the creation of a structured and interpretable decision system while keeping model complexity controlled.

The proposed framework is distinguished by its deviations from conventional XAI approaches that rely on post hoc explanation methods. In contrast, the framework under discussion embeds interpretability directly within the decision layer through structural design. In this work, the focus is shifted from post hoc interpretability to the enforcement of decision-layer interpretability by design. Rather than explaining a trained black-box model, the decision space is explicitly constrained through sparsity-driven feature selection and employ a neuro-fuzzy inference system to construct a structured and transparent, rule-based decision mechanism.

The present study proposes a structured hybrid CNN–FS–ANFIS framework that explicitly separates deep representation learning from decision-layer modeling. A CNN backbone is initially adapted to the glaucoma classification task through the utilizationof transfer learning and is subsequently employed as a fixed feature extractor to obtain deep representations from retinal fundus images. The application of sparsity-driven feature selection is intended to transform high-dimensional representations into a compact decision space. The ANFIS then employs the extracted features to generate explicit rule-based decisions. This architecture under discussion has been demonstrated to facilitate a structured, transparent decision-making process by placing interpretability directly into the decision layer, rather than relying on post hoc explanations. In this paper, we propose a framework that enables inherent interpretability by explicitly separating deep feature representations from sparsity-controlled, rule-based decision modeling via ANFIS.

The following section provides a summary of the most important contributions of this study:

- A structurally interpretable paradigm is introduced to separate deep feature representation from decision-layer modeling. This separation enables a more systematic analysis of the decision behavior, without being affected by end-to-end optimization.
- Sparsity-driven feature selection has been demonstrated to constrain the decision space to a more compact and structured form. When combined with ANFIS, this compact representation enables the construction of an explicit and manageable fuzzy rule-based decision system, reduces unnecessary model complexity, and allows the decision logic to be examined directly without relying on post hoc explanation techniques.

2. Materials and Methods

This section describes the dataset, the preprocessing techniques, and the model used to construct the proposed interpretable decision-layer framework for glaucoma classification.

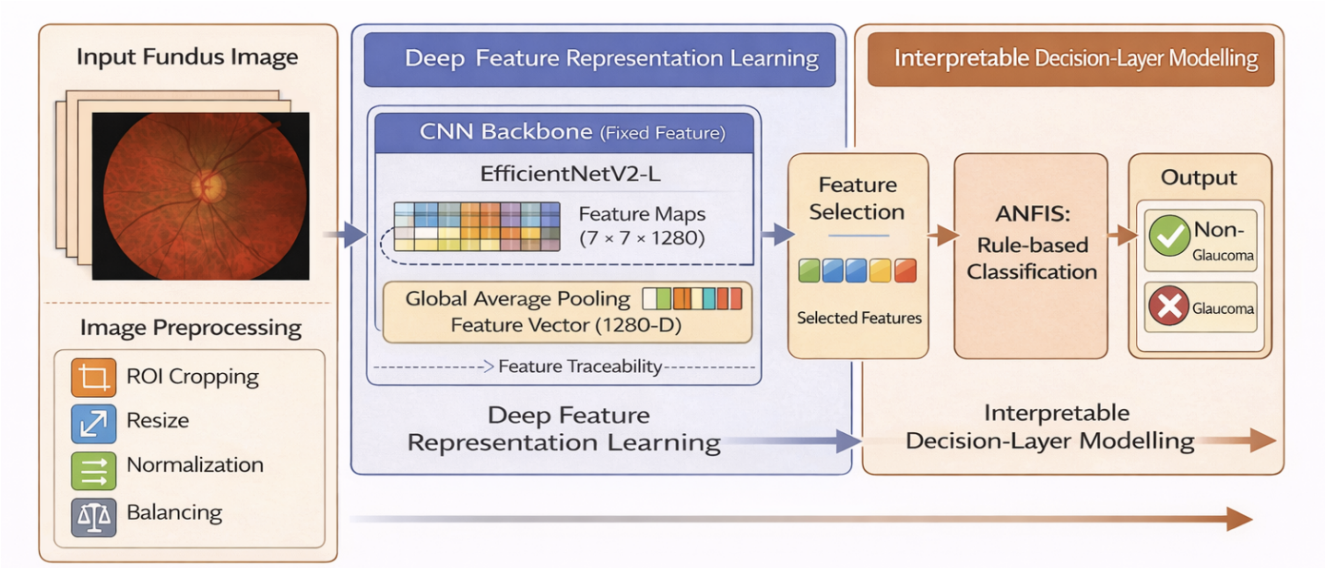


Fig. 1. Proposed framework for interpretable glaucoma detection. The pipeline integrates CNN-based feature extraction, sparsity-driven feature selection, and ANFIS-based decision inference to construct a compact and explicitly interpretable decision layer.

The interpretability of model is an inherent feature of the decision-making process by the design of the model in contrast to post hoc explanation approaches.

The experimental design focuses on controlled comparison rather than end-to-end performance optimization. A CNN backbone is utilized as a feature extractor, and identical deep feature representations are employed across different feature selection and classification methods. The design under consideration has been developed for the purpose of isolating the decision layer's contribution and enabling a systematic evaluation of interpretability-related properties.

2.1. *Dataset*

The present experiments were conducted using the PAPILA dataset, which contains 488 color fundus images from 244 subjects, with each subject contributing images from both eyes. The images have been found to have the same resolution, i.e., 2576 $\times$ 1934 pixels, and were acquired under the same imaging conditions [19]. The dataset contains cases of healthy individuals, individuals with glaucoma, and suspected cases of glaucoma [19]. In the binary classification protocol commonly employed in previous studies, suspected glaucoma samples were merged into the glaucoma class, resulting in two classes: a total of 333 images were deemed to be healthy, while 155 images were classified as positive consisting of cases of glaucoma or suspected glaucoma. The primary motivation for this decision was the objective of the study oriented to screening, where the minimization of false negatives was prioritized over strict separation between confirmed and suspected glaucoma cases.

Most images in the PAPILA dataset are centered on the optic disc region. This region is commonly associated with structural variations relevant to glaucoma-related patterns [11]. In addition, the dataset provides a controlled experimental setting for the analysis of decision-layer behavior under fixed deep feature representations. In contrast to the prevailing emphasis on large-scale generalization in related studies, this study focuses on the structural evaluation of interpretability-oriented decision modeling. A controlled context enables a systematic study of the effect of feature selection on the compactness and interpretability of the downstream decision model.

Image preprocessing was performed to standardize input images, emphasize relevant retinal structures, and enable a controlled evaluation of the proposed decision-layer modeling framework.

2.2. *Image Preprocessing*

Expert-annotated segmentation of the PAPILA dataset was utilized to crop the region of interest (ROI). The cropping region is determined by the boundary between the optic disc and cup, allowing a more accurate localization of relevant retinal structures.

This annotation-guided approach enables consistent inclusion of key anatomical regions associated with glaucoma while reducing background variability. The focus on this region has been shown to enhance the relevance of the extracted visual features for subsequent analysis, as is typically adopted in optic-disc-centered preprocessing strategies [31].

An expert-defined segmentation guides the preprocessing procedure. This approach assists in the reduction of the uncertainty surrounding the concept of ROI localization. This also provides a more controlled experimental setting, facilitating analysis of decision-layer behavior. The ROI extraction protocol follows the optic-disc-centered approach as described in our previous studies [32].

Following ROI extraction, all images were resized to 224 $\times$ 224 $\times$ 3 to match the input size of the CNN backbone. To achieve comparable pixel distributions and minimize the influence of ambient and imaging conditions, intensity normalization was performed. Data balancing was applied to the training set by means of under-sampling to address the class imbalance, resulting in a balanced subset of 310 images, with 155 samples per class. The proposed approach reduces bias against the majority class, thereby enabling fair performance evaluation of the classification problem without the reliance on synthetic data. To

prevent overfitting and improve robustness, data augmentation techniques, including random rotation and horizontal flip, were exclusively applied to the training set.

The validation and test datasets were randomly split to minimize potential bias. The aim of the preprocessing method is to standardize image data, preserve relevant retinal structures, and ensure that the CNN backbone consistently extracts features.

The proposed framework is modular in that its main components can be replaced or extended independently. In future applications, the ROI extraction stage may be replaced by automated optic disc detection, localization, or segmentation methods [33, 34], while the feature selection and decision-layer modules can also be substituted with alternative methods without changing the overall CNN–FS–classifier structure.

2.3. Framework Overview

The proposed framework is illustrated in Fig. 1. In general, it is a sequential processing pipeline, with the objective of mapping raw fundus images to interpretable decisions, which consists of two fundamental steps: deep feature representation learning and interpretable decision-layer modeling.

Firstly, image preprocessing is performed, including region-of-interest (ROI) cropping, scaling, and intensity normalization to highlight task-relevant retinal structures and to ensure input consistency.

Subsequently, the pre-processed images are passed to a CNN backbone that has been adapted through transfer learning. This is then used as a fixed feature extractor during the decision-layer experiments, generating high-level visual representations related to glaucoma patterns.

Afterward, feature selection is applied to reduce the dimensionality of the CNN-derived representation by retaining only the most informative features. This stage has been shown to construct a more compact decision space, thus facilitating the control of the complexity of the subsequent rule-based model. The selected feature set is then used as input to an Adaptive Neuro-Fuzzy Inference System (ANFIS), which performs classification using structured fuzzy rules.

The proposed framework provides a clear transformation from image data to interpretable decisions by separating representation learning from decision-layer modeling, thus embedding interpretability directly into the model structure, instead of relying on post hoc explanation techniques.

2.4. Deep Feature Extraction

To efficiently parameterize and learn high-level visual representations, the EfficientNetV2-L convolutional backbone, which has been pretrained on ImageNet [35], was adopted. The selection of architecture is further supported by our preceding study on the PAPILA dataset, wherein EfficientNet-based models exhibited reliability in glaucoma classification [36]. Building upon these results, we have applied EfficientNetV2-L as the CNN backbone and adapted it to the glaucoma classification task through transfer learning. Subsequent to this adaptation stage, the trained backbone is utilized as a fixed feature extractor to obtain consistent deep representations of fundus images for subsequent feature selection and decision-layer modeling.

Within the EfficientNet family, EfficientNet-B0 and EfficientNet-B7 serve as meaningful comparison points because they represent compact and high-capacity configurations, respectively. As reported in previous studies on the PAPILA dataset, AUC values for EfficientNet-B0, and EfficientNet-B7 were 0.78 and 0.84 respectively [32, 36]. Consequently, these models provide relevant references for assessing the capacity of the proposed CNN–FS–ANFIS framework to achieve competitive performance while improving decision-layer interpretability.

In the proposed framework, CNN adaptation and decision-layer modeling are treated as separate stages. The CNN backbone is first adapted to the glaucoma classification task through the utilization of transfer learning. Subsequent to this adaptation stage, its parameters are maintained constant during the extraction of deep features for feature selection and downstream classifier comparison. This design provides a controlled experimental setting in which all subsequent models operate on identical deep feature representations, thus allowing the study to focus on the impact of feature selection and interpretable decision-layer modeling.

The transferability of learned visual representations across domains supports the use of ImageNet-pretrained models. The early and intermediate layers of deep convolutional networks have been shown to capture generic visual patterns such as edges, textures, and structural configurations. These patterns are not domain-specific and are still applicable to retinal fundus images. Although ImageNet consists of natural images, these transferable features provide a robust basis for capturing structural patterns in fundus images.

EfficientNetV2-L transforms the input image of size $224 \times 224 \times 3$ into $7 \times 7 \times 1280$ feature maps, encoding high-level spatial patterns associated with retinal anatomical structures.

Global average pooling (GAP) is a process that condenses spatial information from convolutional feature maps into a compact representation for further analysis. This operation condenses the $7 \times 7 \times 1280$ tensor into a 1280-dimensional feature vector while preserving the semantic information obtained by the convolutional filters.

At this stage, no classification decision is performed. Instead, the resulting deep feature vector serves as the input representation for the subsequent feature selection stage.

2.5. Feature Selection

The proposed framework is centered on the feature selection stage, the purpose of which is to convert high-dimensional deep representations into a compact, interpretable decision space. This step selects the most informative features from the entire set retrieved instead of providing them all to the classifier immediately. This process eliminates redundancy and directly improves the clarity of the ensuing decision-making process.

In this study, feature selection is systematically applied to a 1280-dimensional feature vector obtained from the CNN backbone after global average pooling (GAP), where each feature corresponds to a specific activation channel of the final convolutional layer. Deep representations in high dimensions frequently prove to be redundant, with discriminative patterns being localized in few features [37–39]. Accordingly, effective decision-making can be achieved using a compact set of dominant features rather than the full feature space.

This dimensionality reduction plays a key role in ANFIS-

based modeling, as it directly impacts the complexity of the fuzzy rule base. The generation of rules is implicitly controlled by the feature selection stage, which restricts the number of input features. It can thus be concluded that the resulting system is structurally manageable for rule-based modeling.

An important property of the proposed framework is that the selected features remain traceable to their corresponding convolutional filters. It can be demonstrated that, since each GAP-derived feature represents the averaged activation of a specific feature map, the index of a selected feature directly identifies the associated channel in the CNN. This process facilitates the mapping of the selected features back to the original spatial feature maps of size 7×7, thereby ensuring structural traceability between the decision layer and the learned deep representations.

This traceability can be employed for interpretability, whereby the CNN filters that contribute to the final conclusion can be examined. The approach serves to bridge the gap between the deep feature representations and the rule-based reasoning.

This stage is of particular importance in the proposed framework, as it directly influences the complexity and interpretability of the subsequent ANFIS-based decision model. In rule-based systems such as ANFIS, the number of rules increases combinatorially with the number of input features. It is therefore important to limit the feature dimensionality to ensure that the decision structure remains comprehensible and tractable.

In this study, multiple feature selection methods are explored to analyze their impact on performance consistency and decision-space compactness. The investigation focused on the impact of feature compactness on the stability and interpretability of the prediction made by the decision layer. was investigated in a controlled experimental environment. Feature subset sizes ranging from 3 to 9 were extensively investigated. This increase in range allows for a more thorough study of the impact of feature compactness on the stability of the decision, and the structural transparency of the decision layer. The analysis includes lower and higher-dimensional configurations in the trade-off between information sufficiency and structural simplicity for interpretable decision modeling. It is evident that the design under discussion facilitates the identification of a representative compact feature subset suitable for interpretability-oriented analysis.

As a supporting perspective, compact feature representations are also consistent with the notion that human decision-making benefits from a limited number of information units [40–42]. However, this merely serves as an interpretability intuition rather than constituting the primary basis for selecting feature subset sizes.

To analyze the impact of feature selection strategies, three methods were considered: Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Mutual Information (MI). These methods provide additional ways to reduce the feature dimensionality.

LASSO achieves sparsity by applying ℓ_1 regularization, which drives coefficients of less informative features to zero [43]. Elastic Net is a combination of ℓ_1 and ℓ_2 regularization that addresses some issues of LASSO when features are correlated [44]. MI measures the statistical dependency between single features and the target variable under the assumption that there is no linear relationship [45].

For LASSO and Elastic Net, features were ranked according to the magnitude of the learned coefficients, with those exhibiting the most significant absolute weights being retained. In the context of MI, the features were categorized based on their dependence on the target variable. Subsequently, the top-k features were selected. All approaches were restricted to selecting the same number of features for the purpose of ensuring a fair and controlled comparison.

The selection process was accomplished in two stages. In the first stage, an ablation study was undertaken utilizing LASSO to evaluate the impact of various sizes of feature subsets (3–9) on sparsity and model stability in repeated experimental iterations. The second step was to conduct a comparison analysis using LASSO, Elastic Net, and MI. For consistency, the number of features in the selected subset was fixed for all methods.

The CNN-derived feature representation was utilized consistently across all experiments to isolate the effect of feature selection from that of representation learning. This controlled design enables systematic evaluation of decision-layer behavior under fixed deep feature representations.

The selected feature subsets are then used as input for downstream classifiers such as standard machine learning models and ANFIS. The objective of this study is to identify and thoroughly investigate the impact of feature dimensionality on the prediction accuracy and interpretability of the decision layer.

In this work, interpretability is defined as the capability of the decision model to explicitly provide human-understandable reasoning through transparent and inspectable structures at the decision layer. The proposed system embeds interpretability in the model design characterized by compact feature representations and clear fuzzy rules as opposed to post hoc explainability systems that provide explanations after model inference. This approach coincides with the purpose of building a transparent decision layer, wherein compact feature representations allow for direct examination of rule-based reasoning without the imposition of additional extra structural complexity.

2.6. Feature Traceability to CNN Representations

An important characteristic of the proposed framework is the traceability of selected features to their corresponding convolutional representations in the CNN backbone. This property establishes a direct link between the interpretable decision layer and the underlying deep feature extraction process.

The final convolutional layer of the CNN architecture generates feature maps of dimensions $7 \times 7 \times 1280$. Each of the 1280 channels can be viewed as an individual activation map that encodes a specific visual pattern in the input fundus image. The feature maps are subsequently passed through a global average pooling (GAP) operation that reduces each 7×7 feature map to a scalar value, thereby yielding a 1280-dimensional feature vector, where each scalar element corresponds directly to one convolutional channel.

In the context of feature selection, a subset of these scalar features is selected according to their relevance to the classification problem. It is important to note that each selected feature retains its original channel index, enabling it to be mapped back to its corresponding convolutional feature map.

This mapping facilitates the reconstruction of the selected

features at the representation level through the retrieval of the associated 7×7 activation maps from the CNN. The decision-making process can thus be related not only to abstract feature values, but also to spatially dispersed activation patterns within the image.

This traceability introduces an additional level of interpretability, linking the deep feature representations to rule-based reasoning within the ANFIS model. It enables further analysis of how specific convolutional filters contribute to the final classification decision, thereby supporting a more transparent and explainable prediction framework.

2.7. Decision-Layer Modeling and Classifier Comparison

The stage of feature selection produces compact feature subsets that have a direct impact on the complexity of the resulting fuzzy rule base. This demonstrates the significant impact of feature dimensionality on interpretable decision modeling.

The selected feature subsets are provided as inputs to multiple decision models to systematically evaluate how different decision-layer strategies behave under identical feature representations. The controlled experimental setup is employed to isolate the effect of the decision models, thereby ensuring that any observed performance differences may be attributed to the decision-layer mechanisms rather than variations in feature extraction. This design is in line with the standard practice of classifier comparison studies, which require consistent input representations to ensure a fair and reliable evaluation [46, 47].

The evaluated models include the C4.5 decision tree [48], RIPPER rule learner [49], support vector machine (SVM) [50], random forest [51], and an Adaptive Neuro-Fuzzy Inference System (ANFIS) [29]. These models encompass a wide range of learning paradigms, including tree-based, rule-based, kernel-based, ensemble, and neuro-fuzzy techniques. This comprehensive approach provides a thorough evaluation of the decision-layer behavior across diverse model families. This diversity is important to ensure that the evaluation captures differences not only in predictive performance but also in model structure, interpretability characteristics, and decision-space representation [52].

Recent clinical decision-support studies have also emphasized the necessity of evaluating rule-based classifiers should not only in terms of their predictive performance but also in terms of the transparency, complexity, and clinical meaningfulness of the rules they generate [53]. This perspective supports the evaluation of decision-layer models in terms of both classification performance and interpretability, particularly in medical classification tasks where transparent decision logic is important.

Among these models, ANFIS provides an explicit and inherently interpretable decision mechanism. In this study, interpretability is defined as the ability of the model to produce transparent and human-readable decision rules. In contrast to popular black-box systems or models with post hoc explanations, ANFIS integrates interpretability into its framework via fuzzy membership functions and if-then rules that allow for visible and traceable decision-making. The ANFIS model has been developed using compact features selected by LASSO, thereby yielding a structured and manageable rule base with essential discriminative information and reduced redundancy.

The premise parameters of the ANFIS model, namely centers and spreads of the Gaussian membership functions, are initialized by means of a clustering-based method inspired by Fuzzy C-Means (FCM). The clustering technique is applied to the specified feature subsets, and the membership functions are constructed in the same compact feature space where the decision model is developed. It is evident that this initialization process facilitates the capture of the underlying distribution of the input features by the membership functions. This, in turn, provides a meaningful and interpretable initialization for rule construction.

Each input feature is represented by Gaussian membership functions. The membership degree of the j-th input feature for the i-th rule is hereby defined as follows:

$$\mu_{ij}(x_j) = \exp\left(-\frac{(x_j - c_{ij})^2}{2\sigma_{ij}^2}\right) \tag{1}$$

where x_j denotes the j-th input feature, while c_{ij} and σ_{ij} represent the Gaussian membership function's center and spread associated with the i-th rule, respectively.

Each fuzzy rule in the first-order Sugeno ANFIS model produces an output defined as:

$$f_i = w_i(p_1x_1 + p_2x_2 + \cdots + p_nx_n + r) \tag{2}$$

where w_i denotes the firing strength of the i-th rule, and $\{p_1, \ldots, p_n, r\}$ are the consequent parameters.

The final model output is formulated as a normalized weighted sum of all rule outputs:

$$v = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{3}$$

The ANFIS output y serves as a raw logit. Because y is continuous and unbounded, it cannot be directly interpreted as a probability. To address this, a sigmoid activation function is applied to map the output into the $[0, 1]$ range:

$$P(\text{Glaucoma}) = \frac{1}{1 + e^{-y}} \tag{4}$$

During training, all model parameters, including both premise parameters (membership function centers and spreads) and consequent parameters, are jointly optimized. The optimization is performed by means of the Adam optimizer, which updates parameters through backpropagation to minimize the classification loss. This unified optimization strategy allows the model to simultaneously adapt the fuzzy partitions and the local linear consequents in a data-driven way.

Adam was selected approach due to its provision of adaptive first- and second-moment updates, which are regarded suitable for jointly optimizing heterogeneous ANFIS parameters, including Gaussian membership centers, spreads, and consequent parameters, with limited manual tuning [54]. Although AdamW, Nadam, and RAdam are valid alternatives [55–57], this present study focuses on the effect of sparsity-driven feature selection and compact ANFIS rule-base modeling rather than optimizer comparison. Therefore, Adam was utilized as a stable baseline optimizer to avoid introducing additional optimizer-specific effects into the decision-layer analysis.

The ANFIS model is characterized by its utilization of a compact feature space to avoid combinatorial rule explosion and

maintains a transparent, stable, and interpretable decision structure. This property is pivotal for enabling direct inspection of the decision process through a limited set of fuzzy rules.

To examine the effect of rule-base complexity on decision-layer behavior, the number of fuzzy rules in the ANFIS model is systematically varied within a compact range (2 to 5 rules). This design facilitates a systematic assessment of the impact of the rule-base size on prediction accuracy, model stability, and interpretability under a given feature representation. It highlights the trade-off between expressive capability and structural transparency within the proposed framework.

The learned ANFIS model provides an explicit set of fuzzy rules that describe the link between the selected feature space and the classification output. These rules provide a structured representation of the decision process, enabling direct inspection of how feature combinations contribute to the final prediction. In the following section, these learned rules are analyzed to examine the interpretability of the proposed decision-layer framework.

2.8. Rule-Based Interpretability Analysis

To study the interpretability of the proposed CNN–FS–ANFIS framework, a rule-based analysis is performed, utilizing the learned fuzzy inference structure and its relation to the underlying convolutional feature representations.

The CNN backbone extracts these deep visual features, which are then translated to channel-wise activation vectors by global average pooling (GAP). Each element of the activation vector is directly proportional to the activation strength of a convolutional filter. The obtained features are then utilized as input to the ANFIS model after feature selection to obtain a compact subset of informative features.

The ANFIS model is capable of learning a set of fuzzy if-then rules with Gaussian membership functions defining the antecedent part and linear Sugeno type functions defining the consequent part. The antecedent segment represents the activation pattern of selected convolutional features, while the consequent models their contribution to the final decision score.

During inference, each rule generates a firing strength that is equivalent to the product of the memberships. The final result is derived via a normalized weighted sum of the rule outputs, and then a sigmoid transformation provides the classification probabilities.

The present study investigates the effect of decision-layer complexity by systematically varying the number of ANFIS rules within a compact range (2 to 5). A controlled environment allows us to study the impact of rule-base size on predictive accuracy, model stability, and interpretability for fixed feature representations.

In the context of interpretability analysis, a compact two-rule configuration is presented as a representative exemplar. This configuration enables the explicit examination of each rule and direct linkage to the selected feature space, thereby highlighting the transparency of the decision mechanism.

To further support interpretability, activation maps of the selected convolutional filters are visualized on retinal fundus images. The visualizations demonstrate the spatial regions of feature activations and provide a contextual understanding of how image patterns are translated into decisions by the fuzzy rules.

This combined analysis establishes a direct relationship between image regions, feature activations, and rule-based reasoning. This enables a structured and transparent interpretation of the decision process in the proposed framework.

The results of the rule-base size analysis are presented in Section 3.5.. This section explores the trade-off between model expressiveness and interpretability.

2.9. Experimental Protocol

The dataset was divided into three distinct segments: training set, validation, and test sets, with 70% for training, 15% for validation, and 15% for testing. This strategy avoids data leakage and provides a good estimate of model generalization, as it is guaranteed that at most one split contains images of the same subject.

To eliminate any possibility of sampling bias and evaluate model stability, all the trials were repeated on five separate times with different random seeds. In this study, a fixed CNN backbone was employed to extract deep features from each image, and all models were assessed with the same feature representations.

The controlled experimental design enables a fair comparison of feature selection methods and decision-layer strategies by isolating the decision layer's contribution. In instances where possible, the hyperparameters were maintained as constant across models, such that the performance differences are mostly related to the feature selection and behavior at the decision layer rather than an optimization bias. In addition, the evaluation was performed uniformly over all feature subset sizes, allowing for a systematic comparison of different levels of sparsity under identical experimental settings.

The training configuration and hyperparameters were defined as follows. Due to the relatively small size of the PAPILA dataset, data augmentation was applied to prevent overfitting. A total of 20 augmented samples were created for each training image using a pipeline of random transformations, including horizontal flips, rotations ($k \times 90^\circ$) and brightness ($\Delta_{\max} = 0.2$). This was further complemented by random zooming (80% area crop) followed by resizing to 224×224 pixels, thereby increasing the diversity of the training set.

The classification framework utilizes an EfficientNetV2-L backbone that has been pretrained on ImageNet. The original top layers were replaced with a GAP layer, followed by a dropout layer (rate = 0.5) and a dense output layer with sigmoid activation.

Training was conducted in two stages. Firstly, the classifier head was trained for 10 epochs, whilst the backbone was maintained in a frozen state ($\mathrm{LR} = 10^{-4}$). This was followed by a fine-tuning phase, during which the entire network was unfrozen and trained for up to 20 epochs utilizing a lower learning rate ($\mathrm{LR} = 10^{-6}$). After this CNN adaptation stage, the classification head was removed, and the GAP output was used as a 1280-dimensional deep feature representation. These extracted features were then maintained constant across all subsequent feature selection and decision-layer experiments. Therefore, LASSO, Elastic Net, Mutual Information, conventional classifiers, and ANFIS were evaluated using the same fixed CNN-derived feature representations.

Table 1. Comparison of decision behavior across feature selection methods using ANFIS (mean ± standard deviation).

Method	AUC	Sensitivity	Specificity	F1-score
All features	0.84±0.01	0.81±0.09	0.69±0.12	0.77±0.02
LASSO + ANFIS	0.84±0.01	0.82±0.13	0.74±0.10	0.79±0.04
Elastic Net + ANFIS	0.84±0.02	0.80±0.13	0.77±0.08	0.78±0.05
MI + ANFIS	0.81±0.05	0.85±0.16	0.76±0.12	0.81±0.06

2.10. *Evaluation Metrics*

Model performance of the model was assessed with threshold-independent and task-appropriate metrics, including sensitivity, specificity, F1-score, and the area under the receiver operating characteristic curve (AUC).

Sensitivity is defined as the proportion of glaucoma patients who were correctly identified as glaucoma cases, while specificity is the proportion of healthy patients correctly identified as being healthy. The F1-score reflects how well precision and recall are balanced. The area under the receiver operating characteristic curve (AUC) evaluates the model's discrimination ability across different decision thresholds.

The statistical significance of the different testing configurations was evaluated by using a non-parametric test, the Friedman test. This test allows the comparison of many models over multiple runs.

2.11. *Computational Environment*

The experiments conducted in this study were executed on the Google Colaboratory cloud platform, utilizing an NVIDIA Tesla T4 GPU with approximately 15 GB of RAM and a Python 3 runtime environment with high-RAM configuration enabled.

The implementation of model development and deep feature extraction were implemented utilized TensorFlow and Keras, while feature selection and fuzzy inference components were implemented using standard Python libraries.

From a computational-complexity perspective, the cost of the proposed framework is mainly dominated by the EfficientNetV2-L feed-forward process. For an input image size of 224×224 pixels, the feed-forward complexity of EfficientNetV2-L is estimated to be approximately 11.5 GFLOPs per image. This estimation is obtained by scaling the reported 53 GFLOPs at 480×480 resolution according to the squared spatial-resolution ratio [35]. After feature extraction, a conventional fully connected classifier utilizing all 1280 GAP features requires 1281 trainable parameters and has an inference complexity of $O(1280)$. In contrast, the proposed ANFIS decision layer employs only six selected GAP features and two fuzzy rules, resulting in approximately 38 trainable parameters with an inference complexity of $O(Rn) = O(12)$, where R is the number of rules and n is the number of selected features. It can thus be concluded that, despite the overall computational cost being dominated by the EfficientNetV2-L backbone, the proposed GAP feature selection followed by ANFIS substantially reduces classifier-level complexity while providing an interpretable rule-based decision structure. The principal foucs of this study was not empirical runtime benchmarking, since the primary contribution is to be found in decision-layer interpretability rather than deployment-time optimization.

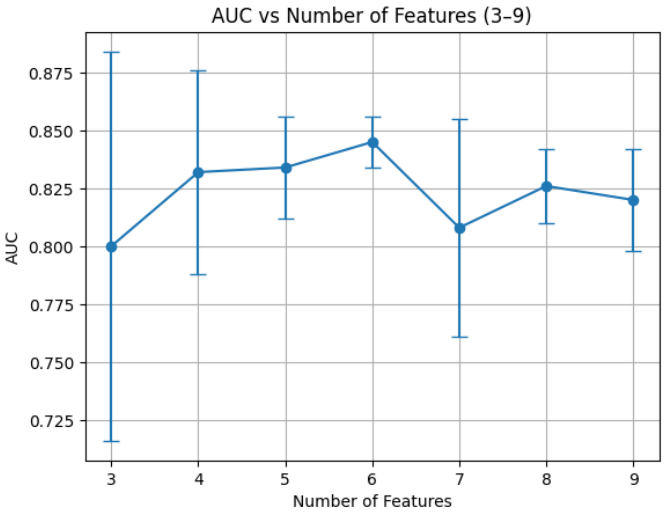


Fig. 2. AUC scores across feature subset sizes ranging from 3 to 9 using LASSO-based feature selection. Error bars represent the standard deviation across five experimental runs.

3. Results and Discussion

This section discusses the proposed framework from the point of view of decision-layer modeling. The discussion is focused on the impact of the feature selection and classifier design on the structure, compactness, and interpretability of the resulting decision process. The evaluation is structured into two main components. Firstly, the impact of different feature selection methods on the resulting feature space is analyzed under a fixed decision-layer configuration. Secondly, multiple classifiers are evaluated using identical feature representations to assess the behavior of different decision-making strategies under reduced dimensionality.

The analysis emphasizes consistency in decision behavior and the structural properties of the resulting models, as opposed to the utilization of absolute performance metrics. This perspective enables a more meaningful assessment of interpretability within decision support systems in medical imaging. Moreover, the impact of feature-space reduction on the control of decision space complexity is discussed. Finally, the interpretability of the ANFIS decision layer is evaluated by means of rule-based analysis and case-based assessment.

3.1. *Ablation Study on Feature Subset Size*

An ablation study was accomplished on feature subset sizes ranging from 3 to 9 under cautiously controlled experimental settings purposely to investigate the effect of feature dimensionality on classification performance and model stability. The design enables a systematic investigation of the effect of compact-

ness of features on the predictive stability and interpretability of the decision layers by isolating the impact of feature dimensionality from other effects.

As demonstrated in Fig. 2, a clear and consistent performance trend is evident. From the experimental results, the AUC increases, and the variance of the performance decreases as the number of selected features from small-size subsets increases, indicating that the stability of the model increases across multiple runs.The performance further stabilizes within a moderate range of feature subset size, wherein the model has the potential to maintain high AUC with a relatively low variance.

This behavior suggests that a compact range of feature subset sizes is sufficient to capture the essential discriminative information while maintaining stable predictive behavior. In the event of an insufficient number of features being selected, the resulting model has limited representational capacity, resulting in higher variance and reduced discriminative power. Conversely, if the number of features is increased beyond this range, the model includes redundant or less informative features, which leads to diminishing returns and potential instability.

This observation is of particular relevance in the case of rule-based models, such as ANFIS, where the complexity of decision layer increases combinatorially with the number of input features. The dimensionality of the features needs to be limited so as to keep a compact and manageable decision structure, enabling a clearer interpretation of the resulting rule base.

From an alternative perspective, the range of compact features identified is consistent with constraints of human working memory capacity commonly reported [40–42]. While this observation provides a natural explanation, it is not utilized as the main basis for feature selection.

Overall, the results obtained indicate that performance stabilizes within a compact and stable feature range, thus supporting the use of a representative subset for subsequent analysis. Accordingly, a representative compact subset is adopted and consistently used in the following experiments to ensure a controlled evaluation of decision-layer behavior across different modeling configurations.

3.2. *Comparison of Feature Selection Methods*

The present study was the first to investigate the influence of feature selection strategies on decision-layer behavior. Four configurations were evaluated in this study: (i) no feature selection (all pooled features), (ii) LASSO-based selection, (iii) Elastic Net-based selection, and (iv) MI-based selection. To isolate the effect of feature selection, the downstream decision model was fixed to ANFIS across all configurations.

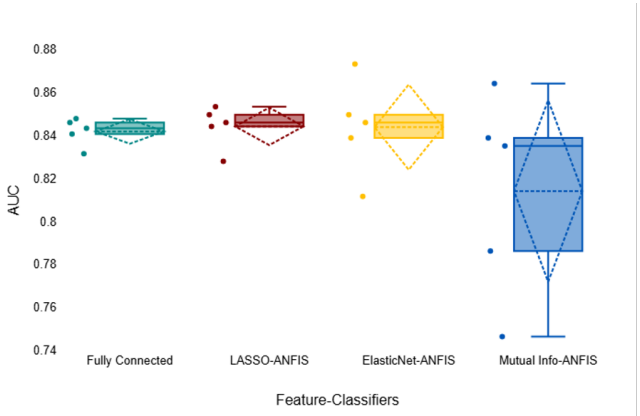
As summarized in Table 1, the mean and standard deviation of AUC, sensitivity, specificity, and F1-score are presented, with these metrics obtained across repeated experimental runs.

Overall, only minor numerical differences are observed across feature selection methods, as summarized in Table 1. Performance metrics demonstrate uniformity across all settings, indicating that the decision behavior is similar, even with variation in feature sparsity. In particular, similar AUC values suggest that the compressed feature representations keep the most significant discriminative structure extracted by the convolutional backbone.

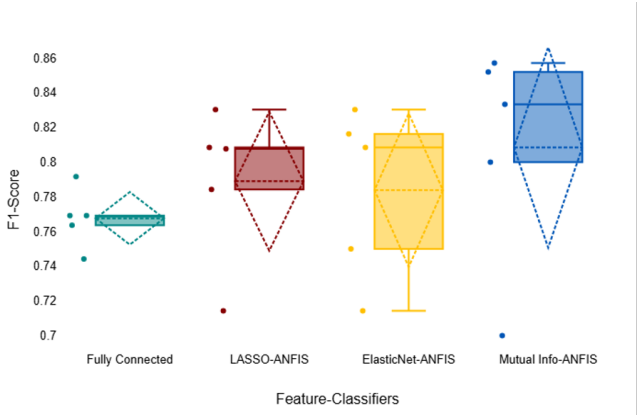
Fig. 3 shows boxplots that further support this observation,

Table 2. Summary of Friedman test results for feature selection comparison.

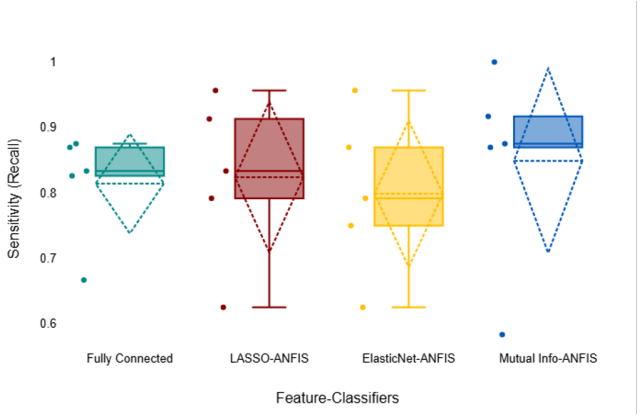
Metric	χ^2	df	p-value
Sensitivity	1.86	3	0.602
Specificity	4.98	3	0.173
F1-score	4.02	3	0.259
AUC	4.20	3	0.241



(a) AUC



(b) F1-score



(c) Sensitivity

Fig. 3. Decision behavior distribution across feature selection strategies using ANFIS. The boxplots illustrate the variability of (a) AUC, (b) F1-score, and (c) sensitivity across repeated experimental runs. Substantial overlap among distributions indicates consistent decision behavior despite differences in feature sparsity.

revealing a significant overlap in the distribution of performance metrics across the different feature selection methods. The variability and central tendencies of AUC, F1-score, and sensitivity remain comparable, suggesting similar variability and consistent decision-making behavior across multiple runs.

To assess whether these numerical differences are statistically meaningful, non-parametric Friedman tests were conducted for each evaluation metric. The results are summarized in Table 2. For all detection-related metrics, we cannot reject the null hypothesis of equal performance across feature selection methods ($p > 0.05$). The results show that performance measures alone are not sufficient to fully describe how feature selection affects model behavior and interpretability.

However, despite the overall similarity in performance, a detailed examination of the behavior of the metrics over the feature selection methods indicates some subtle differences. In particular, MI exhibits relatively high sensitivity and specificity but a lower AUC. This behavior can be explained in terms of the nature of the evaluation metrics, since sensitivity and specificity are threshold dependent and reflect the performance at a given operating point, whereas AUC is threshold independent and measures the ability of the model to maintain consistent ranking across all possible thresholds. As shown in Table 1 and Fig. 3, differences in ranking consistency can be obscured by overlapping distributions. These results suggest that MI might perform well at a specific decision threshold but provides less stable class separation overall, resulting in a lower AUC.

In such a case, other model design considerations beyond raw performance metrics become more important. In this study, interpretability-oriented criteria provide a meaningful basis for selecting the most suitable feature selection method.

MI selects features based on their individual statistical dependence with the target label. This approach can identify useful features but does not explicitly address the redundancy among the selected variables, which can lead to less compact feature sets.

The limitation of this method is that correlated features cannot be retained. Elastic Net overcomes this limitation by combining elasticity ℓ_1 and ℓ_2 regularization. However, doing so will also increase the dimensionality of the decision space and potentially impact interpretability.

In contrast, LASSO explicitly enforces sparsity by ℓ_1 regularization, encouraging the selection of a small subset of informative features. Apart from its competitive performance, LASSO provides a consistent and controlled approach to feature sparsity, allowing for the analysis of feature subset size in the ablation study. Finally, the low variability across LASSO-based configurations indicates stable decision behavior over repeated runs.

This type of sparsity is particularly beneficial in terms of interpretability of the decision layer. Reducing the input space dimensionality naturally limits the complexity of the subsequent rule-based inference process. In rule-based models such as the ANFIS, the number of potential rules as well as the interactions between membership functions increase exponentially with the number of input variables. By limiting the number of features, LASSO provides a way to avoid rule explosion while preserving the inspection capability of the decision architecture. This property is essential for enabling a transparent decision process at the decision layer.

Therefore, although all feature selection strategies exhibit comparable performance on conventional evaluation metrics, LASSO is selected as the primary strategy. The reason for this choice is that it can enforce sparsity, stabilize the performance and limit the complexity of the decision space, which are all important aspects for obtaining a compact, interpretable rule-based model in the ANFIS decision layer.

3.3. Classifier Comparison on LASSO-Selected Features

After identifying LASSO as the most suitable feature selection method, a compact and interpretable feature subset was obtained. These features were then used to evaluate several classification models. Rather than identifying a universally superior classifier, this comparison examines how different decision mechanisms behave under sparsity-constrained feature representations.

Five classifiers were considered, namely, the C4.5 decision tree, RIPPER rule learner, support vector machine (SVM), random forest, and Adaptive Neuro-Fuzzy Inference System (ANFIS). These models represent different decision-making strategies such as rule-based (C4.5, RIPPER), margin-based (SVM), ensemble (random forest), and neuro-fuzzy (ANFIS).

We did not emphasize accuracy and other measures that depend on a threshold since they are less reliable in the presence of class imbalance, which is common in medical datasets. Instead, we focused on task-appropriate, threshold-independent metrics such as AUC, sensitivity, specificity, and F1-score. These metrics provide a more complete view of the classifier's performance.

Friedman tests were conducted across multiple experimental runs to assess the statistical significance of the observed differences among classifiers. The results indicate statistically significant differences across classifiers for AUC ($p = 0.002$), sensitivity ($p = 0.009$), specificity ($p = 0.023$), and F1-score ($p = 0.011$). This finding suggests that, beyond feature selection, the structure of the decision layer continues to influence model behavior under the same LASSO-selected feature representation.

We summarize the results in Table 3. As shown in Table 3, ANFIS achieved the highest AUC, sensitivity, and F1-score, while maintaining specificity comparable to SVM and random forest. RIPPER obtained the highest specificity, but its lower AUC, sensitivity, and F1-score indicate less balanced overall behavior. Therefore, ANFIS provides the most suitable trade-off between discriminative performance and interpretable rule-based decision modeling.

These results lead to several observations. The SVM, random forest and ANFIS behave similarly, suggesting that the features selected by LASSO still carry the essential information required for decision-making. Even with the reduction in dimensionality, these models are stable. This suggests that a small subset of features is sufficient to capture the underlying patterns, while the original CNN representation is largely redundant.

In this setting symbolic rule learners like C4.5 and RIPPER are less successful by comparison. They have simpler structures, which are less capable of modeling nonlinear relations in the reduced feature space, and thus are less effective.

While SVM and random forest can model nonlinear patterns

Table 3. Summary of classifier behavior across evaluation metrics using LASSO-selected features. Values are reported as mean ± standard deviation across five experimental runs. Sens. = sensitivity; Spec. = specificity.

Classifier	AUC	Sens.	Spec.	F1-score
C4.5 Decision Tree	0.67±0.07	0.68±0.09	0.66±0.12	0.67±0.06
RIPPER	0.70±0.08	0.60±0.15	**0.81±0.04**	0.67±0.12
SVM	0.82±0.02	0.79±0.12	0.74±0.09	0.77±0.04
Random Forest	0.81±0.03	0.75±0.10	0.74±0.08	0.74±0.05
ANFIS	**0.84±0.01**	**0.82±0.13**	0.74±0.10	**0.79±0.04**

effectively, their decision processes remain difficult to interpret. In contrast, ANFIS provides a trade-off between expressive modeling power and interpretability by explicitly modeling the decision process via fuzzy rules.

This advantage becomes more pronounced when combined with sparse feature representations. The ANFIS model built on a small set of input variables extracted with LASSO is structurally manageable, with a small number of fuzzy rules that can be directly inspected. Such properties are particularly important in medical image analysis contexts, where transparency and traceability are essential.

More generally, these results highlight the importance of *decision space compactness* as a key ingredient of interpretable deep learning. The reduction of the feature space prior to the decision step preserves a structure that is more interpretable and still sufficiently expressive.

In this framework, LASSO reduces the high-dimensional CNN features into a compact set of decision-relevant variables, while ANFIS translates them into an explicit tractable rule-based system. The pipeline they form allows for fully transparent decision-making with no need for post hoc explanations and offers a more direct and auditable approach for medical image analysis.

Based on these observations, ANFIS is selected as the final decision layer in the proposed framework. This selection is motivated not only by its stable discriminative behavior but also by its structural compatibility with sparsity-driven feature representations and its support for inherently interpretable inference.

3.4. Decision Space Considerations

As shown in Table 1 and supported by the Friedman test results in Table 2, the evaluation outcomes across different feature selection strategies are statistically indistinguishable ($p > 0.05$). This indicates that multiple feature selection methods can preserve the essential decision-relevant information in deep convolutional representations.

When decision behavior remains largely comparable across feature selection strategies, the choice of method cannot be justified solely based on evaluation metrics. The structure of the resulting feature space is therefore a critical factor, as the dimensionality and organization of this space directly affect the complexity and interpretability of the downstream decision model.

In this framework, feature selection acts as a structural constraint to reduce the redundancy of the high-dimensional CNN feature space. Deep convolutional features often exhibit correlated or overlapping patterns. If left unfiltered, these patterns can lead to unnecessarily complex and less clear interpretable decision structures.

This issue becomes particularly important for rule-based models. When operating on high-dimensional inputs, the number of possible rule combinations increases exponentially, often leading to *rule explosion*. Such growth increases decision complexity, making transparent and interpretable decision-making more difficult.

In this study, decision space compactness is formally defined as the ability to represent high-dimensional deep features using a small subset of informative features that preserve the essential decision structure while enabling stable and interpretable rule-based inference.

The ablation study further supports this observation, showing that performance consistently stabilizes within a compact range of feature subset sizes. In particular, the results around six features indicate that a compact representation is sufficient to preserve discriminative information while avoiding unnecessary complexity in the decision layer.

The sparsity-inducing property of LASSO is therefore particularly beneficial. LASSO provides a parsimonious set of decision-relevant features while ensuring consistent decision-making via ℓ_1 regularization. As seen in the previous section, this reduction to the feature space does not have a significant impact on the classification outcomes, which indicates that the selected features contain the necessary information for classification.

The compact representation combined with the ANFIS decision layer leads to a more transparent inference mechanism. The fuzzy rule base remains manageable and can be inspected completely due to the reduction of input variables. This helps to prevent excessive rule growth and simultaneously allows the model to capture nonlinear relationships.

The results underscore the importance of a compact decision space for deep learning-based prediction systems. CNNs provide rich feature representations, but their high dimensionality can be detrimental to the decision stage and interpretability. We argue that constraining the feature space prior to inference allows a trade-off between the expressivity of the representations and their interpretability.

In summary, the combination of LASSO-based feature selection and ANFIS-based rule inference is a powerful means to obtain interpretable decision layers, where interpretability is intrinsic to the model structure rather than a post hoc explanation.

This capability is especially important for medical image analysis, where transparent and interpretable decision processes are necessary for understanding model behavior and allowing for reliable analysis.

Table 4. Performance comparison across different ANFIS rule-base sizes (mean ± standard deviation from 5 experimental trials).

Rules	AUC	Sensitivity	Specificity	F1-score	Precision
2 Rules	0.8420±0.0168	0.8069±0.1223	0.7450±0.0959	0.7795±0.0449	0.7676±0.0573
3 Rules	**0.8445±0.0072**	**0.8232±0.0980**	0.7797±0.1047	**0.8039±0.0475**	0.7963±0.0727
4 Rules	0.8428±0.0358	0.7989±0.1473	**0.7957±0.0897**	0.7928±0.0761	**0.8021±0.0573**
5 Rules	0.8286±0.0205	0.7895±0.1304	0.7612±0.0751	0.7745±0.0584	0.7734±0.0382

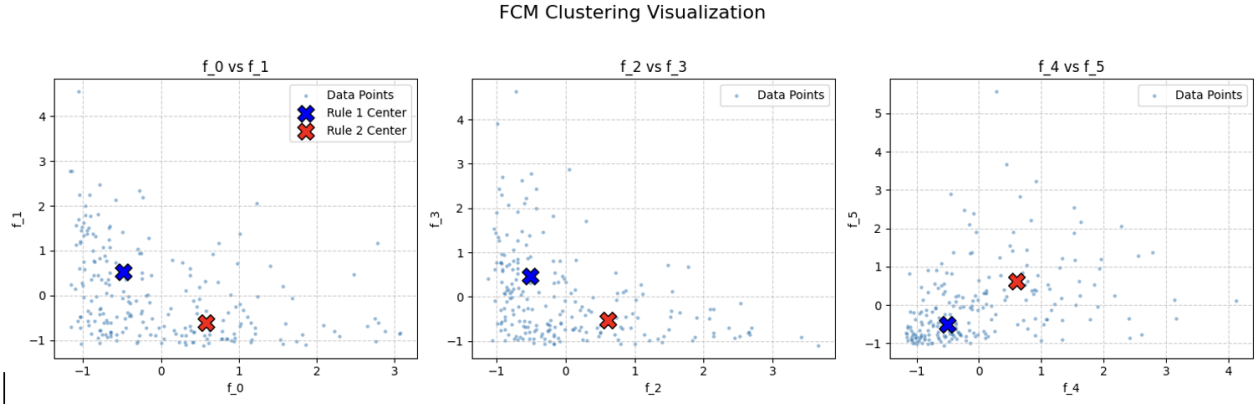


Fig. 4. Visualization of the reduced CNN feature space and the centers of fuzzy rules obtained during clustering initialization. Data points represent samples projected onto pairwise feature dimensions (f_0, f_1), (f_2, f_3), and (f_4, f_5), while the markers indicate the centers of the two fuzzy rules learned by the ANFIS decision layer.

3.5. Analysis of Rule-Base Size

To further investigate decision-layer complexity, a controlled experimental design is employed in which the number of ANFIS rules is systematically varied within a compact range from 2 to 5, while the feature representation is kept fixed. The controlled environment provides a unique opportunity to study the effect of rule-base size on predictive behavior, model stability, and interpretability without the confounding effects of feature variability.

Table 4 summarizes the performance metrics obtained for different rule-base sizes.

The three-rule configuration shows slightly higher values in AUC and F1-score, but with stable behavior on sensitivity and specificity (see Table 4). Nevertheless, the two-rule configuration achieves comparable performance in the explored range and provides a more compact and clear interpretable decision structure.

Increasing the number of rules beyond three does not consistently improve performance. The four-rule configuration gives the best specificity and precision but also a higher degree of variability, as shown by larger standard deviations over experimental runs. The five-rule configuration exhibits a drop in performance for most of the metrics, which can be an indicator of over-parameterization relative to the available data.

On the modeling side, more rules provide more flexibility in modeling nonlinear decision boundaries. However, this comes at the cost of reduced interpretability, as the number of rule interactions and the decision complexity increase.

Overall, the results indicate that compact rule configurations (particularly 2 to 3 rules) are sufficient to capture the essential decision patterns in the reduced feature space. Furthermore, the statistical test indicates that the performance differences between the configurations are not significant ($p > 0.05$), implying that the increase in rule complexity does not result in significant improvements in the predictive performance.

These results support the design principle of the trade-off between the expressive capability and the structural transparency in the decision layer. In particular, simpler rule bases allow for clearer interpretation without compromising predictive performance, which supports the principle of parsimony formalized in Occam's Razor [58].

From a parsimony perspective, these findings suggest that the addition of rule-based complexity adds little value when performance differences are small and inconsistent. Under such conditions, a simpler configuration is preferred, as additional complexity does not yield consistent improvements while reducing analytical transparency and introducing unnecessary structural complexity in the decision layer.

Accordingly, these observations guide the selection of a representative compact rule base for subsequent interpretability analysis. In particular, a two-rule configuration is adopted as a representative setting, as it provides a clear and tractable decision structure while maintaining stable behavior across the explored range.

This configuration provides a compact and stable decision structure, which is subsequently used for interpretability analysis in the following section.

3.6. Rule-Based Interpretability Analysis

This section investigates the interpretability of the proposed framework by explicitly analyzing the learned fuzzy rules in the ANFIS decision layer, building on the rule-based size analy-

sis presented in Section 3.5. The rule variation study indicates that compact rule configurations are sufficient to maintain stable, consistent predictive behavior within the explored setting while enabling a more transparent, tractable decision structure.

Following LASSO-based feature selection, the original high-dimensional CNN feature representation is reduced to a compact subset of six informative variables. This type of dimensionality reduction limits the complexity of the decision layer and prevents the explosion of the rule base, enabling interpretable modeling. In this controlled environment, configurations with a limited number of rules strike a good balance between predictive stability and interpretability.

Therefore, we choose a two-rule configuration as a representative case for interpretability analysis, since it offers a straightforward and manageable decision structure while remaining consistent with the performance trends observed in Section 3.5. This selection is consistent with the principle of parsimony, which favors simpler configurations when performance differences are marginal.

This compact rule base enables direct inspection of the decision logic, i.e., each rule can be analyzed explicitly without the need for post hoc explanation methods. The relation between deep feature representations and decision outcomes can be investigated in a structured, transparent and inherently interpretable way.

Feature Space Structure and Rule Centers

To examine how the ANFIS model organizes the reduced feature space, the distribution of the selected CNN features is visualized together with the centers of the fuzzy rules. This analysis focuses on the representative two-rule configuration selected based on the findings presented in Section 3.5. It provides a compact setting for analyzing decision-space partitioning.

Fig. 4 illustrates the spatial distribution of samples along with the rule centers obtained during fuzzy clustering initialization. The visualization is presented using pairwise feature projections: (f_0, f_1), (f_2, f_3), and (f_4, f_5). In each plot the distribution of samples is shown in the corresponding feature subspace and the centers of the two fuzzy rules are marked.

Several observations can be drawn. First, the features extracted by the selected CNN have structured distributions, which indicates that the convolutional backbone learns meaningful and discriminative features. Second, the rule centers are distributed in different parts of the reduced feature space, which indicates a well-defined partition. Third, the separation of the centers shows that the ANFIS model divides the feature space into regions associated with different decision behaviours, which supports interpretable decision boundaries.

Gaussian Membership Function Adaptation

The Gaussian membership function is used to model each antecedent condition of the proposed ANFIS model:

$$\mu(x) = \exp\left(-\frac{(x-c)^2}{2s^2}\right) \tag{5}$$

where c and s are the center and spread parameters that are learned during the training process. These functions define soft activation regions of the reduced feature space, which helps in modeling the nonlinear relationships among the chosen CNN features.

Fig. 5(a) and Fig. 5(b) show the Gaussian membership functions before and after the training process. The initial membership functions are symmetrically distributed and highly overlapping, thus forming a rough partition of the feature space. The learning process modifies both the centers and spreads to the characteristics of the underlying data distribution, leading to feature-specific activation regions that model class separation more accurately.

This adaptation shows the effect of the learned feature representations on the decision boundaries. The Gaussian parameters adapt to the structure of the selected features instead of learning fixed partitions, thereby bridging deep representations to interpretable fuzzy regions.

Explicit Rule Representation

The extracted representative rules from the trained model are shown in Table 5. Each rule consists of Gaussian antecedent conditions defined over the six selected features and a linear Sugeno-type consequent that determines the final decision output.

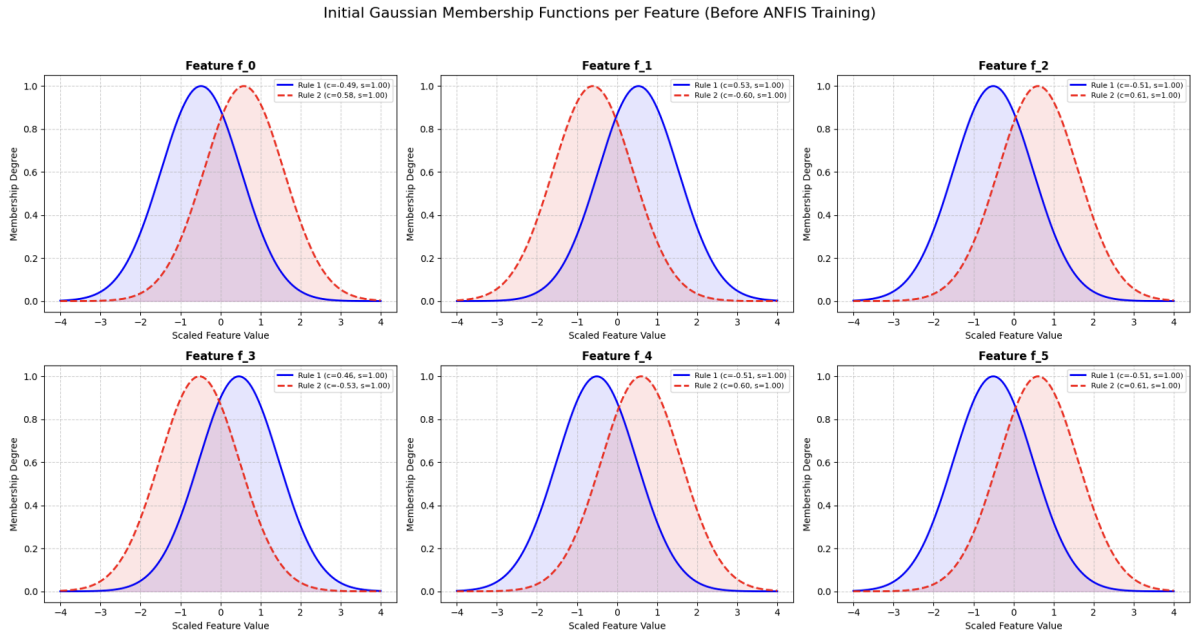
Table 5. Representative fuzzy rules learned by the ANFIS decision layer.

Rule 1
IF f_0 is Gauss($c = -0.733, s = 1.264$) AND f_1 is Gauss($c = 0.941, s = 1.214$) AND f_2 is Gauss($c = -0.946, s = 0.635$) AND f_3 is Gauss($c = 0.903, s = 0.648$) AND f_4 is Gauss($c = -0.803, s = 0.832$) AND f_5 is Gauss($c = -0.938, s = 0.673$)
THEN $y = 0.326f_0 - 0.311f_1 + 0.339f_2 - 0.330f_3 + 0.311f_4 + 0.347f_5 - 0.335$
Rule 2
IF f_0 is Gauss($c = 0.510, s = 0.678$) AND f_1 is Gauss($c = -0.293, s = 0.910$) AND f_2 is Gauss($c = 0.896, s = 0.903$) AND f_3 is Gauss($c = -0.965, s = 0.643$) AND f_4 is Gauss($c = 0.373, s = 1.482$) AND f_5 is Gauss($c = 0.711, s = 1.142$)
THEN $y = 0.309f_0 - 0.341f_1 + 0.292f_2 - 0.340f_3 + 0.336f_4 + 0.315f_5 + 0.337$

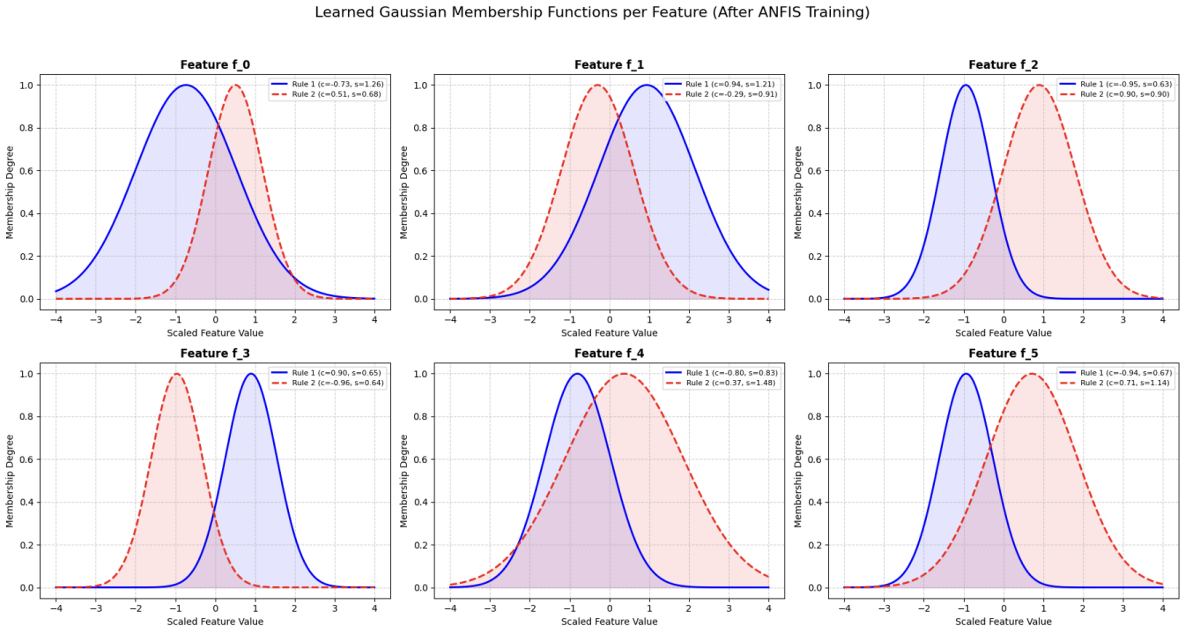
Several structural characteristics can be observed from the learned rule base. Each rule is defined on a limited set of six chosen features, which shows that the original high-dimensional CNN representation can be efficiently compressed without any loss of relevant decision information. Moreover, the Gaussian membership functions define localized activation regions, i.e., each rule models specific patterns in the reduced feature space, thus improving interpretability.

The linear consequent coefficients explicitly quantify the contribution of each feature to the final decision output. If the coefficients are positive, increasing the value of the corresponding feature increases the decision score. If the coefficients are negative, increasing the feature value suppresses the decision score. This fully explicit formulation allows direct inspection of the impact of features in the decision layer, which is usually not possible in conventional deep neural networks.

Another important feature is the reduced number of rules. Only two rules are generated, which leads to a compact and



(a) Initial Gaussian membership functions before ANFIS training.



(b) Learned Gaussian membership functions after ANFIS training.

Fig. 5. Evolution of Gaussian membership functions during ANFIS training: (a) initial symmetric membership functions used for model initialization; (b) learned membership functions after training, where the centers and spreads adapt to the distribution of the selected CNN features.

tractable decision structure, avoiding the rule explosion typically faced by fuzzy systems with higher-dimensional inputs. This compactness allows a global analysis of the decision process without the need for additional interpretation procedures.

These results show the importance of the compactness of the decision space for interpretable modeling. CNN-based approaches often produce high-dimensional feature representations that increase decision complexity. By integrating feature selection with rule-based inference, the decision space is constrained into a structured and manageable form while preserving stable predictive behavior.

To summarize, the results indicate that interpretability in medical image analysis can be obtained by structural design, not by post hoc explanation. In the proposed framework, interpretability is embedded directly into the model architecture, where feature reduction and rule-based inference jointly yield a transparent, explicitly analyzable decision process.

3.7. Case-Based Interpretability Analysis

To further evaluate the interpretability of the proposed CNN–ANFIS framework, three representative prediction cases are analyzed: a confirmed glaucoma case, a borderline case near the decision boundary, and a clearly normal retinal sample. These cases are selected to illustrate model behavior across different classification scenarios and to demonstrate how selected deep features contribute to the activation of fuzzy rules and the final prediction outcome.

Unlike post hoc explanation techniques, which interpret model behavior after training, the proposed framework explicitly embeds interpretability directly within its architecture. Deep feature representations are first extracted by the convolutional backbone, followed by feature selection to significantly reduce the dimensionality of the decision space. The resulting compact feature set is then processed by the ANFIS layer, which provides a transparent and structured rule-based inference mechanism.

Each prediction can be directly traced through a structured decision pathway in which selected features contribute to rule activation and influence the final output. This enables one to identify discriminative information and different visual patterns associated with different classification outcomes.

From an analytical point of view, the model shows similar behavior for different cases, where the different activations of features reflect the differences in the underlying input patterns. In particular, borderline cases show an intermediate activation pattern and thus have a lower prediction confidence than cases clearly separable.

Case 1: Glaucoma Sample

Fig. 6 shows a representative glaucoma sample based on the proposed framework. Fig. 6(a) shows an ANFIS inference visualization of the activation of the selected deep features of the fuzzy membership functions that determine the firing strength of the corresponding fuzzy rules.

Rather than relying on a single dominant rule, the final prediction is obtained by aggregating multiple rule activations. This mechanism provides a transparent, structured inference process, in which the contribution of each feature to the final prediction can be examined through its associated membership activations and rule weights.

The feature channels shown in Fig. 6(b) correspond to the subset of deep features retained after LASSO-based feature selection. LASSO does not directly select the convolutional filters, rather it acts on the pooled CNN feature representation and selects the most informative feature dimensions for classification. The proposed framework reduces the redundancy of the deep representation by limiting the input space of the decision layer to a compact subset of discriminative features, while preserving informative patterns for the classification task.

The spatial activation maps are shown in Fig. 6(c) to further demonstrate the regions in the retinal image where the responses of these selected features are the strongest. A number of filters have focused responses around the optic disc and nearby regions. Although the activation maps should not be considered as absolute anatomical markers, the fact that they correspond to structurally important regions suggests that the model learns consistent and interpretable visual patterns.

Case 2: Borderline Sample

Fig. 7 shows a borderline case, where the model prediction is close to the decision boundary. In this case, the ANFIS inference provides a probability slightly below the classification threshold, hence normal prediction with low confidence. This behavior is consistent with the intrinsic uncertainty in samples with intermediate visual features of healthy and glaucomatous structures.

The dominant CNN features shown in Fig. 7(b) exhibit weaker and less structured responses compared with the glaucoma case, providing limited discriminative evidence for either class. Consequently, multiple fuzzy rules are activated with comparable strengths, leading to competing contributions during inference.

The activation maps in Fig. 7(c) show that filter responses are spatially spread over several retinal areas, not concentrated on a particular anatomical structure. Such distributed activation patterns correspond to the moderate confidence observed during rule aggregation, indicating that the model avoids overconfident predictions when the visual evidence is ambiguous.

Case 3: Normal Sample

Fig. 8 presents a clearly normal retinal image. In this case, the ANFIS inference yields a probability well below the classification threshold, leading to a confident normal prediction. The rule-aggregation process, therefore, strongly favors the non-glaucoma class.

The CNN feature responses retained after feature selection are relatively weak and diffuse, indicating that the convolutional backbone does not detect strong pathological patterns within the retinal image. This observation is further confirmed by the activation maps where the filter responses are not dominantly focused on specific retinal structures.

Cross-Case Analysis

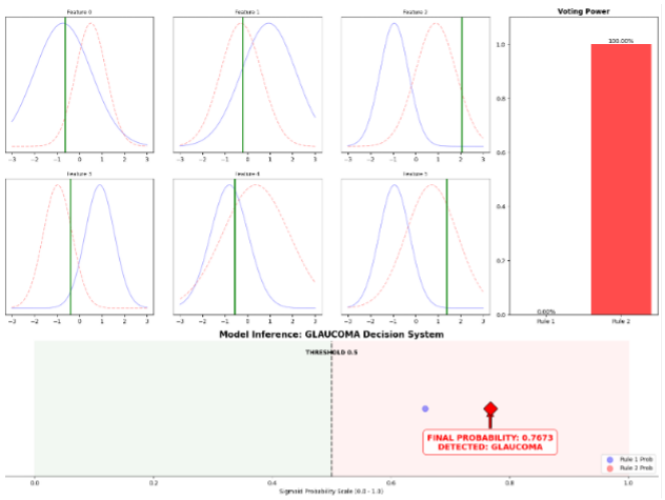
Comparing across the three cases shows how the proposed framework works across different classification scenarios. The glaucoma sample shows strong and localized feature activations, which result in dominant rule firing and a confident positive prediction. In contrast, the normal sample shows weaker and more scattered feature responses resulting in rule activations towards the healthy class. The borderline case is located in between the two extremes, with moderate feature responses resulting in competing rule activations and prediction probabilities close to the decision threshold.

This smooth transition between cases indicates that the model learns a continuous mapping between feature responses and prediction confidence, rather than sharp decision boundaries. It is a model behavior that indicates the model can represent different levels of evidence in a structured decision process.

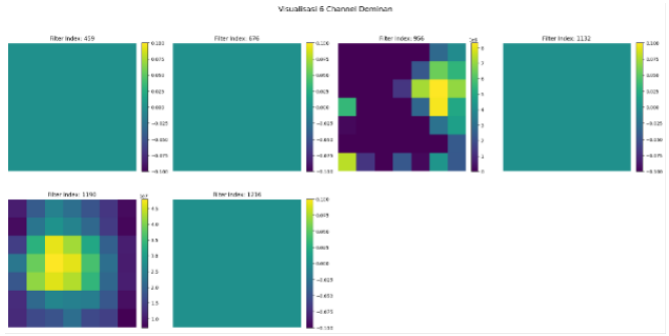
Feature Selection Consistency

Another important aspect of the proposed framework is the consistency of the feature selection mechanism. The LASSO procedure does not directly select convolutional filters but rather works on the pooled CNN feature representation and identifies feature dimensions that consistently contribute to classification across experimental runs.

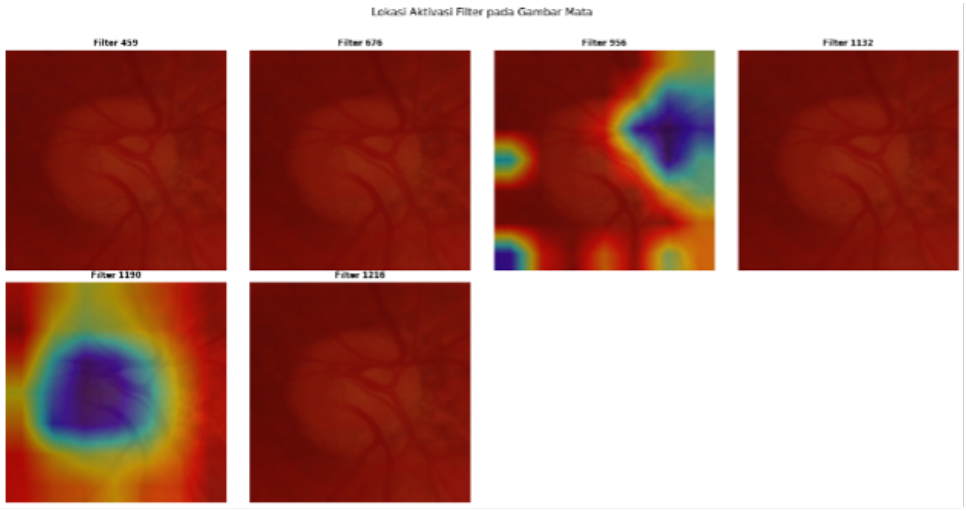
The stability of the selected features confirms that the infor-



(a)



(b)



(c)

Fig. 6. Interpretability analysis for a glaucoma sample: (a) ANFIS inference and rule activation; (b) dominant CNN feature channels retained through LASSO-based feature selection; (c) spatial activation patterns of the selected filters projected onto the retinal fundus image.

mation used to discriminate between glaucomatous and healthy images is not randomly distributed in the feature space, and some CNN-based representations tend to be repeatedly found to be informative for glaucoma classification. The stability of the selected features also confirms that the feature selection stage captures meaningful visual patterns in the deep representation.

Consistent feature selection is also important for interpretability. A common subset of features selected by experiments makes the rule base more stable and easier to analyze. This property further improves the transparency of the proposed framework by making sure that the decision process is not dictated by arbitrary fluctuations of features.

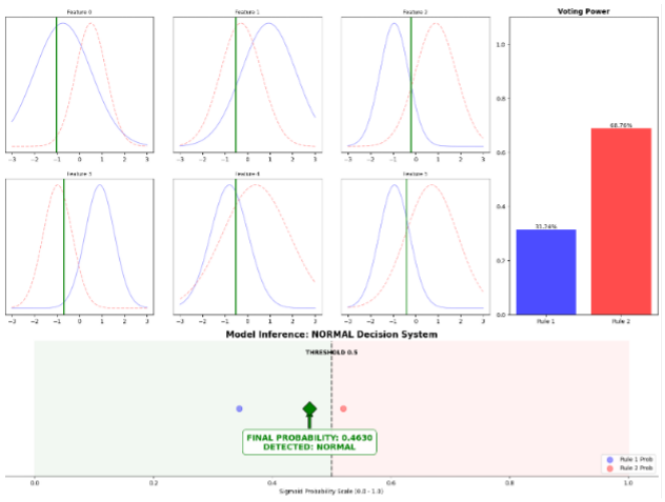
Decision Space Compactness

The proposed design also has the advantage of a compact decision space. Typically, deep convolutional networks generate high-dimensional feature representations with hundreds or even thousands of latent variables. Directly feeding such representations into a rule-based inference system would result in excessive rule complexity and reduced interpretability [59].

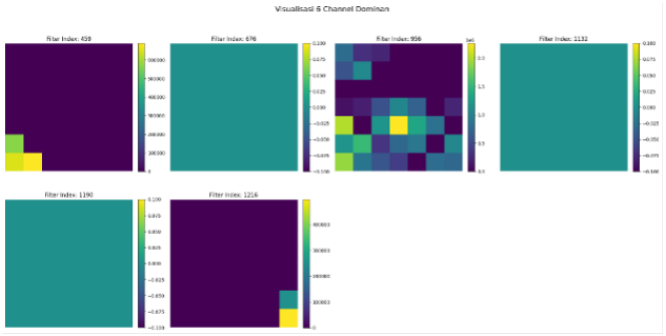
The framework implements sparsity-based feature selection before the decision layer, limiting the dimensionality of the input space on which ANFIS functions, leading to a fuzzy rule-based system working with a small set of informative features. This allows for a compact and manageable inference structure while maintaining competitive performance [60].

From an interpretability perspective, this compact decision space improves traceability of the classification process. Each prediction is broken down into a series of interpretable steps: extraction of deep features, filtering of the features by LASSO selection, and rule-based reasoning with the ANFIS inference mechanism. This structured pipeline provides a transparent path from the evidence derived from the retinal image to the final classification decision.

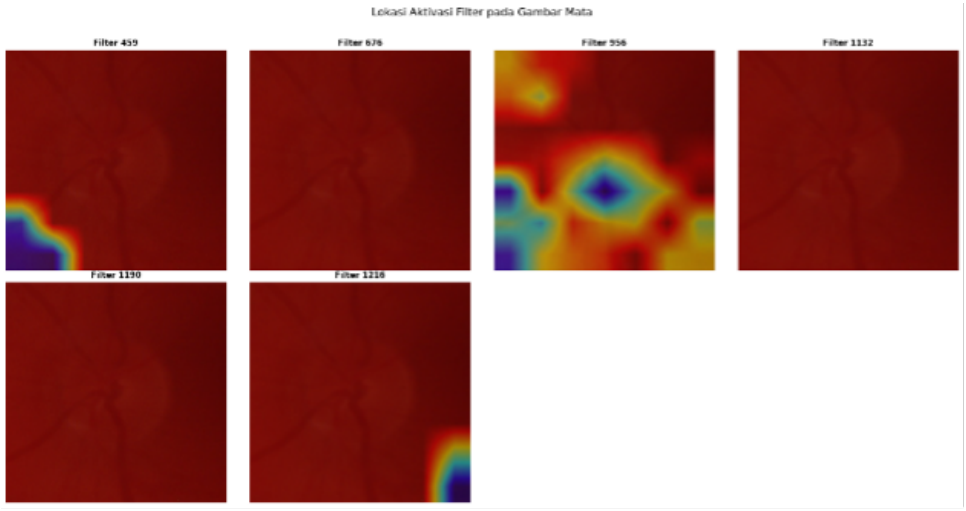
In summary, the case-based analysis demonstrates the potential of the proposed CNN–FS–ANFIS framework to provide interpretable predictions in various diagnostic scenarios. The combination of deep representation learning with a lean rule-based decision layer allows the framework to find a compromise between the predictive performance and the transparency of the model, a compromise that is essential for constructing reliable medical AI systems in the future.



(a)



(b)



(c)

Fig. 7. Interpretability analysis for a borderline retinal sample: (a) ANFIS inference producing a probability close to the decision threshold; (b) dominant CNN features retained after feature selection; (c) spatial activation patterns of the selected filters.

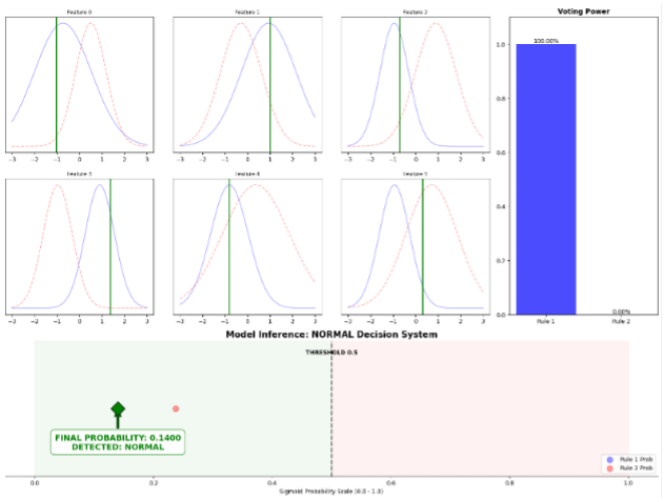
Table 6. Comparison with existing methods evaluated on the PAPILA dataset. Reported AUC values are taken from the respective original studies, as AUC is the most consistently available metric across these works.

Method	Backbone	Decision Model	Interpretable	AUC
Various CNN methods [19]	CNN	Fully connected	No	0.71
Regression CNN [7]	CNN	Regression layer	No	0.77
Ensemble CNN [20]	CNN ensemble	Neural classifier	No	0.78
EfficientNet-B0 [32]	EfficientNet-B0	Fully connected	No	0.78
EfficientNet-B7 [36]	EfficientNet-B7	Fully connected	No	0.84
CNN–FS–ANFIS (Proposed)	EfficientNetV2-L	Fuzzy rule inference	**Yes**	**0.84**

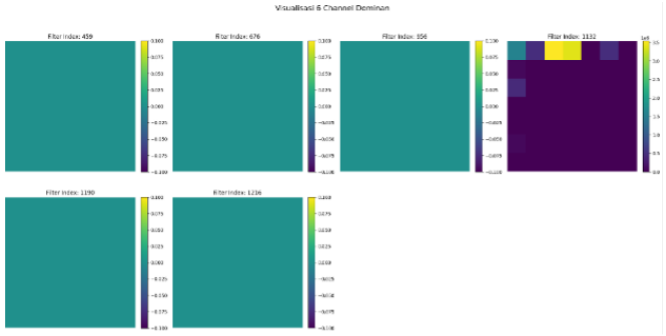
3.8. Comparison with Existing Methods on the PAPILA Dataset

In order to have a reference of the performance of the proposed framework, we compare it against the state-of-the-art methods tested on the PAPILA dataset. The comparison on a common dataset allows to have a more reliable basis to judge the relative effectiveness, because the performance of models may vary from dataset to dataset due to different image characteristics, annotation protocols, and evaluation settings.

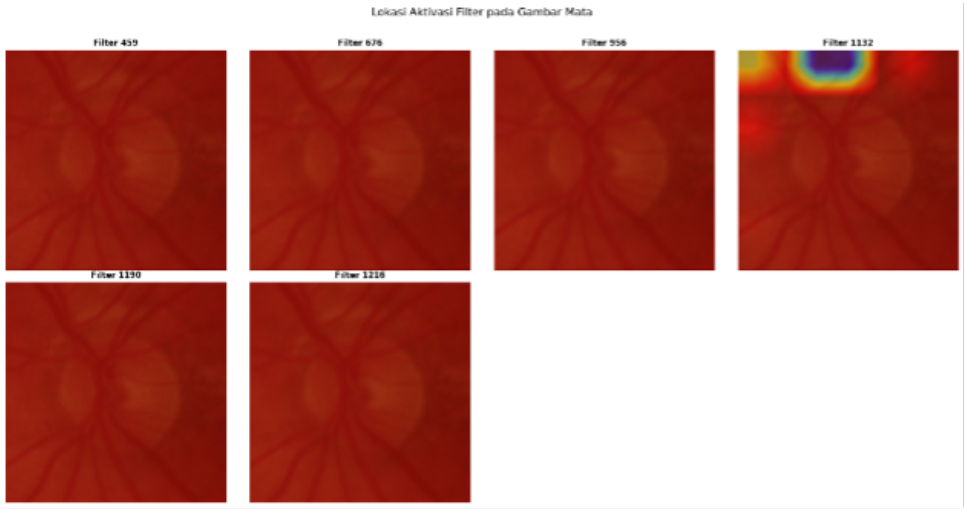
For comparison with previous studies, AUC was used as the main reference metric because it is the most consistently reported metric across existing works on the PAPILA dataset. Sensitivity, specificity, and F1-score were not included in Table 6 because these metrics were not consistently reported in the respective original studies or were evaluated under differ-



(a)



(b)



(c)

Fig. 8. Interpretability analysis for a normal retinal sample: (a) ANFIS inference producing a probability well below the classification threshold; (b) dominant CNN features selected by LASSO; (c) spatial activation patterns projected on the retinal image.

ent experimental protocols, data splits, and decision thresholds. Therefore, directly comparing these metrics could lead to an unfair interpretation. In contrast, the proposed method is evaluated in detail using AUC, sensitivity, specificity, and F1-score in Table 3 and Table 4.

Table 6 compares a number of representative methods on the PAPILA dataset. Previous works such as Kovalyk et al. [19] achieved an AUC of 0.71 using a variety of CNN-based models with fully connected decision layers. We considered alternative modeling strategies to improve predictive performance. Hemelings et al. [7] applied regression-based CNNs and achieved an AUC of 0.77, Sanchez-Morales et al. [20] an AUC of 0.78 with an ensemble CNN.

Additional improvements were achieved with newer methods based on the EfficientNet family. EfficientNet-B0 achieved an AUC of 0.78 [32] and the larger EfficientNet-B7 architecture achieved an AUC of 0.84 [36]. The proposed CNN–FS–ANFIS framework achieves an AUC of 0.84, which is comparable to that of the high-capacity EfficientNet-B7 model while additionally providing an interpretable decision mechanism through fuzzy rule-based inference.

Despite these competitive results, most existing approaches rely on conventional neural network classifiers that produce opaque decision boundaries. In contrast, the proposed framework introduces interpretability at the decision layer through sparsity-driven feature selection and rule-based fuzzy inference. The model narrows the decision space and transforms deep feature representation into explicit fuzzy rules, enabling transparent, structured reasoning and consistent decision-making behavior.

From a methodological point of view, these findings underscore the need for a trade-off between interpretability and decision consistency in medical AI systems. Deep CNNs offer an efficient way to learn hierarchical feature representations, but the decision mechanism remains hard to interpret due to the dense transformations of the high-dimensional latent spaces.

The proposed architecture circumvents this limitation by imposing structural constraints in the decision layer. LASSO-based feature selection is used to eliminate redundancy while keeping the most informative features, which are then mapped by ANFIS to a compact, interpretable rule base. This design preserves predictive ability and allows traceable, auditable

decision-making.

Unlike the post hoc explanation approaches, the proposed framework attains interpretability inherently by model design. Such structural integration leads to a more reliable and transparent pathway to interpretable AI-based glaucoma classification.

3.9. Limitations of the Study

There are several limitations of this study which must be noted. This is particularly true regarding the scope of experimental validation and the desire for decision layer analysis.

First, the experiments were performed on the PAPILA dataset, which is relatively small in comparison to large-scale retinal imaging datasets. However, the study deliberately employs a controlled experimental setting to allow for a systematic investigation of the decision-layer behavior with fixed deep feature representations. The design offers methodological clarity but needs validation on larger and more diverse datasets to establish generalizability.

Second, the proposed framework is validated on a single primary dataset with relatively homogeneous imaging characteristics and acquisition protocols. Although a subject-level split is employed to prevent data leakage, cross-dataset evaluation remains an important direction for future work. Future validation may include ACRIMA, REFUGE, and ORIGA [61–63], since these datasets provide different data distributions, imaging conditions, annotation protocols, and optic-disc-centered evaluation settings for assessing model generalizability beyond PAPILA.

Third, after the CNN backbone was adapted to the glaucoma classification task through transfer learning, the extracted GAP features were kept fixed during the feature selection and decision-layer experiments. This design choice was intended to isolate and analyze the contribution of the interpretable decision layer under identical CNN-derived feature representations. Nevertheless, further investigation of joint optimization between feature extraction, feature selection, and decision-layer modeling could be explored in future studies.

Finally, although ANFIS offers an explicit rule-based interpretability, the current analysis is restricted to structural transparency and representative case-based interpretability. A thorough expert-based interpretability analysis would be useful to further explore the consistency, stability, and domain-specific relevance of the extracted rules.

In sum, these limitations are a conscious trade-off between controlled methodological analysis and broader applicability, and they create opportunities for future research to extend the proposed framework to more diverse data settings.

3.10. Summary of Findings

Some useful insights into the behavior of the proposed CNN–FS–ANFIS framework can be obtained by the experimental results.

The first comparison of feature selection strategies shows that the differences of evaluation results between the compared methods are relatively small. This suggests that the deep CNN representations have considerable redundancy, which allows selecting compact feature subsets that contain the essential information for consistent decision-making.

Second, the classifier comparison shows that the choice of a decision mechanism is still important even with sparsity-constrained feature representations. While some classifiers show similar discriminative behaviors, ANFIS provides the most balanced performance across a range of evaluation metrics while maintaining an explicit rule-based inference structure.

Third, the joint use of LASSO-based feature selection and ANFIS allows for the construction of a compact and interpretable decision layer. By reducing the dimensionality of the feature space before classification, the framework prevents rule explosion and ensures that the resulting inference process remains transparent and inspectable.

Fourth, the case-based analysis shows that the model captures a gradual relationship between feature responses and prediction confidence. When features are strongly activated and are localized, the model makes a confident prediction of glaucoma. When the activation is less or spread out, the model predicts glaucoma with less confidence or is borderline, indicating ambiguity near the decision boundary.

In conclusion, the current results highlight the relevance of *decision space compactness* as a design principle for interpretable deep learning systems. These results indicate that discriminative power and interpretability are not incompatible objectives and can be achieved at the same time through a proper structural design of the decision layer. Moreover, the proposed framework provides transparent and traceable decision-making processes, making it possible to conduct systematic analysis of model behavior in image classification tasks.

4. Conclusion

In this study, we proposed a CNN–FS–ANFIS framework for interpretable glaucoma classification by separating deep feature extraction from decision-layer modeling. A CNN backbone was first adapted to the glaucoma classification task through transfer learning and then used to obtain retinal feature representations, sparsity-driven feature selection was applied to construct a compact decision space, and ANFIS was employed to generate explicit fuzzy rule-based decisions. The results show that the proposed framework achieved competitive classification performance using LASSO-selected features, with ANFIS obtaining an AUC of 0.84 ± 0.01, sensitivity of 0.82 ± 0.13, specificity of 0.74 ± 0.10, and F1-score of 0.79 ± 0.04. The analysis further shows that compact feature representations are sufficient to maintain stable decision behavior while reducing rule-based complexity. In particular, two- to three-rule ANFIS configurations achieved AUC values of approximately 0.84, indicating that increasing rule complexity does not necessarily improve predictive performance. These findings support decision-space compactness as an important design principle for interpretable deep learning. Overall, the proposed CNN–FS–ANFIS framework provides a structured alternative to black-box CNN classifiers by enabling direct inspection of the relationship between selected CNN features, fuzzy rules, and model outputs.

References

1. J. D. Steinmetz, R. R. A. Bourne, P. S. Briant, S. R. Flaxman, H. R. B. Taylor, J. B. Jonas, et al. *Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the Global Burden of Disease Study*. Lancet Glob. Health 9 (2021) e144–e160.
2. X. Li, G.-X. Huang, R. Li, Z.-J. Zhang, S.-Q. Zhang, L. Zhu, et al. *2021 Global Burden of Disease Study: a comprehensive analysis of glaucoma*

in the middle-aged and older adult population at global, regional, and national levels. Front. Public Health 13 (2025) 1526061.
3. S. Shan, J. Wu, J. Cao, Y. Feng, J. Zhou, Z. Luo, et al. *Global incidence and risk factors for glaucoma: a systematic review and meta-analysis of prospective studies*. J. Glob. Health 14 (2024) 04252.
4. Z. Wang, C. C. Xue, Y. Li, et al. *Global glaucoma prevalence: burden and projection to 2060*. Am. J. Ophthalmol. 283 (2026) 324–335.
5. W. M. Liao, B. J. Zou, R. C. Zhao, Y. Q. Chen, Z. Y. He, M. J. Zhou. *Clinical interpretable deep learning model for glaucoma diagnosis*. IEEE J. Biomed. Health Inform. 24 (2020) 1405–1412.
6. X. Yang, J. Wu, X. Wang, Y. Yuan, J. Li, G. Chen, et al. *Multi-scale spatio-temporal transformer-based imbalanced longitudinal learning for glaucoma forecasting from irregular time series images*. IEEE J. Biomed. Health Inform. 29 (2025) 2859–2870.
7. R. Hemelings, B. Elen, A. K. Schuster, M. B. Blaschko, J. Barbosa-Breda, P. Hujanen, et al. *A generalizable deep learning regression model for automated glaucoma screening from fundus images*. NPJ Digit. Med. 6 (2023) 112.
8. S. A. Haja, V. Mahadevappa. *Advancing glaucoma detection with convolutional neural networks: a paradigm shift in ophthalmology*. Rom. J. Ophthalmol. 67 (2023) 222–237.
9. S. Sangchocanonta, P. Pooprasert, N. Lerthirunvibul, K. Patchimnan, P. Phienphanich, A. Munthuli, et al. *Optimizing deep learning models for glaucoma screening with vision transformers for resource efficiency and the pie augmentation method*. PLOS ONE 20 (2025) e0314111.
10. H. A. Nugroho, T. Kirana, V. Pranowo, A. H. T. Hutami. *Optic cup segmentation using adaptive threshold and morphological image processing*. Commun. Sci. Technol. 4 (2019) 63–67.
11. S. Saha, J. Vignarajan, S. Frost. *A fast and fully automated system for glaucoma detection using color fundus photographs*. Scientific Reports 13 (2023) 44473.
12. R. Arnay, J. Hernández-Aceituno, T. Díaz-Alemán, J. M. Martínez-de-la Casa, J. Sigut. *Optic cup segmentation of stereo retinal fundus images using virtual reality*. Med. Biol. Eng. Comput. 61 (2023) 1421–1434.
13. Y. Xu, M. Hu, H. Liu, H. Yang, H. Wang, S. Lu, et al. *A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis*. NPJ Digit. Med. 4 (2021).
14. X. Huang, M. R. Islam, S. Akter, F. Ahmed, E. Kazami, H. A. Serhan, et al. *Artificial intelligence in glaucoma: opportunities, challenges, and future directions*. Biomed. Eng. Online 22 (2023) 126.
15. J.-A. Kim, H. Yoon, D. Lee, M. Kim, J. Choi, E. J. Lee, et al. *Development of a deep learning system to detect glaucoma using macular vertical optical coherence tomography scans of myopic eyes*. Scientific Reports 13 (2023) 13642.
16. Y. Hagiwara, O.-A. Ciora, M. Monnet, G. Lancho, J. M. Lorenz. *AI-driven approaches for glaucoma detection – a comprehensive review*. Diagnostics 14 (2024) 52.
17. D. M. H. Nguyen, H. M. T. Alam, T. Nguyen, D. Srivastav, H.-J. Profitlich, N. Le, et al. *Deep learning for ophthalmology: the state-of-the-art and future trends*, 2025. arXiv preprint.
18. V. K. Velpula, L. D. Sharma. *Automatic glaucoma detection from fundus images using deep convolutional neural networks and exploring networks behaviour using visualization techniques*. SN Computer Science 4 (2023).
19. O. Kovalyk, J. Morales-Sánchez, R. Verdú-Monedero, I. Sellés-Navarro, A. Palazón-Cabanes, J.-L. Sancho-Gómez. *PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment*. Scientific Data 9 (2022) 291.
20. A. Sánchez-Morales, J. Morales-Sánchez, O. Kovalyk, R. Verdú-Monedero, J.-L. Sancho-Gómez. *Improving glaucoma diagnosis assembling deep networks and voting schemes*. Diagnostics 12 (2022) 1382.
21. D. J. Nugraha, N. Yudistira, A. W. Widodo. *Weighted loss for imbalanced glaucoma detection: insights from visual explanations*. Comput. Biol. Med. 196 (2025) 110875.
22. X. Kui, Z. Hai, B. Zou, Y. Li, W. Liang, Z. Ming, et al. *PK-Net: a prior knowledge-driven dual-path network for enhanced glaucoma screening*. Knowl.-Based Syst. 329 (2025) 114374.
23. J. Sigut, F. Fumero, J. Estévez, S. Alayón, T. Díaz-Alemán. *In-depth evaluation of saliency maps for interpreting convolutional neural network decisions in the diagnosis of glaucoma based on fundus imaging*. Sensors 24 (2024) 239.
24. M. D. Abramoff, M. K. Garvin, M. Sonka. *Retinal imaging and image analysis*. IEEE Rev. Biomed. Eng. 3 (2010) 169–208.
25. D. N. K. Hardani, I. Ardiyanto, H. Adi Nugroho. *Decoding brain tumor insights: Evaluating CAM variants with 3D U-Net for segmentation*. Commun. Sci. Technol. 9 (2024) 262–273.
26. G. Chandrashekar, F. Sahin. *A survey on feature selection methods*. Comput. Electr. Eng. 40 (2014) 16–28.
27. B. Singh, M. Doborjeh, Z. Doborjeh, S. Budhraja, S. Tan, A. Sumich, et al. *Constrained Neuro-Fuzzy Inference Methodology for Explainable Personalised Modelling with Applications on Gene Expression Data*. Scientific Reports 13 (2023) 456.
28. Y. A. Sagar, M. S. R. L. Reddy, S. Shilpa, N. Jyothi, A. Velivela, A. N. Rao. *Neuro-fuzzy systems: neural networks and fuzzy logic integration in soft computing*. In *Cybernetics, Human Cognition, and Machine Learning in Communicative Applications*, pages 39–49. Springer Nature Singapore, 2025.
29. J.-S. R. Jang. *ANFIS: adaptive-network-based fuzzy inference system*. IEEE Trans. Syst. Man Cybern. 23 (1993) 665–685.
30. M. Yeganejou, S. Dick, J. Miller. *Interpretable deep convolutional fuzzy classifier*. IEEE Trans. Fuzzy Syst. 28 (2020) 1407–1419.
31. H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, et al. *Disc-aware ensemble network for glaucoma screening from fundus image*. IEEE Trans. Med. Imaging 37 (2018) 2493–2501.
32. E. Irijanti, H. A. Nugroho, I. Ardiyanto. *Performance analysis for glaucoma diagnosis using transfer learning on the PAPILA dataset*. In *2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP)*, pages 213–218, 2023.
33. F. Calimeri, A. Marzullo, C. Stamile, G. Terracina. *Optic Disc Detection Using Fine Tuned Convolutional Neural Networks*. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 69–75, 2016.
34. H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, X. Cao. *Joint Optic Disc and Cup Segmentation Based on Multi-label Deep Network and Polar Transformation*. IEEE Trans. Med. Imaging 37 (2018) 1597–1605.
35. M. Tan, Q. V. Le. *EfficientNetV2: Smaller Models and Faster Training*. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR, 2021.
36. E. Irijanti, H. A. Nugroho, I. Ardiyanto. *EfficientNet model for detecting glaucoma on the PAPILA dataset*. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 62–67, 2024.
37. Y. Bengio, A. Courville, P. Vincent. *Representation learning: a review and new perspectives*. IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.
38. B. O. Ayinde, T. Inanc, J. M. Zurada. *On correlation of features extracted by deep neural networks*. IEEE Trans. Neural Netw. Learn. Syst. 30 (2019) 3138–3150.
39. I. Guyon, A. Elisseeff. *An Introduction to Variable and Feature Selection*. J. Mach. Learn. Res. 3 (2003) 1157–1182.
40. A. D. Baddeley. *Working memory*. Science 255 (1992) 556–559.
41. N. Cowan. *The magical number 4 in short-term memory: a reconsideration of mental storage capacity*. Behav. Brain Sci. 24 (2001) 87–114.
42. G. A. Miller. *The magical number seven, plus or minus two: some limits on our capacity for processing information*. Psychological Review 63 (1956) 81–97.
43. R. Tibshirani. *Regression shrinkage and selection via the lasso*. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1996) 267–288.
44. H. Zou, T. Hastie. *Regularization and variable selection via the elastic net*. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005) 301–320.
45. T. M. Cover, J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006.
46. J. Demšar. *Statistical Comparisons of Classifiers over Multiple Data Sets*. J. Mach. Learn. Res. 7 (2006) 1–30.
47. R. Caruana, A. Niculescu-Mizil. *An empirical comparison of supervised learning algorithms*. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 161–168. ACM, 2006.
48. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
49. W. W. Cohen. *Fast effective rule induction*. In *Proceedings of the 12th International Conference on Machine Learning (ICML)*, pages 115–123. Morgan Kaufmann, 1995.
50. C. Cortes, V. Vapnik. *Support-vector networks*. Machine Learning 20 (1995) 273–297.
51. L. Breiman. *Random forests*. Machine Learning 45 (2001) 5–32.
52. T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.
53. F. Fityah, N. A. Setiawan, D. W. Anggrahini. *Interpretability evaluation of rule-based classifier in myocardial infarction classification based on syntactical features of ECG signal*. Commun. Sci. Technol. 10 (2025) 460–

466.
54. D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization*. In *International Conference on Learning Representations*, 2015.
55. I. Loshchilov, F. Hutter. *Decoupled Weight Decay Regularization*. In *International Conference on Learning Representations*, 2019.
56. T. Dozat. *Incorporating Nesterov Momentum into Adam*. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.
57. L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, et al. *On the Variance of the Adaptive Learning Rate and Beyond*. In *International Conference on Learning Representations*, 2020.
58. A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth. *Occam's razor*. Inf. Process. Lett. 24 (1987) 377–380.
59. Y. Wang, A. Paschke. *Extracting interpretable hierarchical rules from deep neural networks' latent space*. In *Rules and Reasoning*, pages 238–253. Springer Nature Switzerland, 2023.
60. D. Zhang, T. Chen. *Scikit-ANFIS: a scikit-learn compatible Python implementation for adaptive neuro-fuzzy inference system*. Int. J. Fuzzy Syst. 26 (2024) 2039–2057.
61. A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, A. Navea. *CNNs for Automatic Glaucoma Assessment Using Fundus Images: An Extensive Validation*. Biomed. Eng. Online 18 (2019) 29.
62. J. I. Orlando, H. Fu, J. B. Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, et al. *REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs*. Med. Image Anal. 59 (2020) 101570.
63. Z. Zhang, F. S. Yin, J. Liu, D. W. K. Wong, N. M. Tan, B. H. Lee, et al. *ORIGA(-light): An Online Retinal Fundus Image Database for Glaucoma Analysis and Research*. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3065–3068, 2010.