IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 November 2024, accepted 23 December 2024, date of publication 26 December 2024,
date of current version 2 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3522862

RESEARCH ARTICLE

Explainable AI-Based Approach for Age-Related Macular Degeneration (AMD) Detection via Fundus Imaging

AINHOA OSA-SANCHEZ[1,2], HOSSAM MAGDY BALAHA[1], ALI MAHMOUD[1], ASHRAF SEWELAM[3], MOHAMMED GHAZAL[4], (Senior Member, IEEE), BEGONYA GARCIA-ZAPIRAIN[2], AND AYMAN EL-BAZ[1], (Fellow, IEEE)

[1]Department of Bioengineering, J. B. Speed School of Engineering, University of Louisville, Louisville, KY 40292, USA
[2]eVIDA Research Group, University of Deusto, 48007 Bilbao, Spain
[3]Ophthalmology Department, Faculty of Medicine, Mansoura University, Mansoura 35516, Egypt
[4]Electrical, Computer, and Biomedical Engineering Department, Abu Dhabi University, Abu Dhabi, United Arab Emirates

Corresponding author: Ayman El-Baz (aselba01@louisville.edu)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by University of Louisville.

ABSTRACT Age-related macular degeneration (AMD) is a leading cause of vision loss in older people and is characterized by subtle retinal changes that make early identification difficult. Previous studies have demonstrated the efficacy of Vision Transformers (ViTs) in classifying medical images by successfully detecting retinal disorders such as AMD. This paper addresses multiple shortcomings in conventional AMD diagnostic techniques by exploring the detection and explanation of various AMD subtypes from numerical features extracted with a ViT model from fundus images through cascaded artificial intelligence (AI) models using transformers, convolutional neural networks (CNNs), and multilayer perceptrons (MLPs). The data were preprocessed to recognize intricate disease-related patterns. The best test results using the cascade method for each model type show that the MLP model achieved an accuracy of 91.86% (with a sensitivity of 92.22% and a specificity of 95.74%). The Transformer model achieved its highest accuracy of 83.72% (with a sensitivity of 83.86% and a specificity of 89.74%). The CNN model demonstrated the best performance, with an accuracy of 94.19% (with a sensitivity of 93.84% and a specificity of 96.00%). This work helps clinicians interpret AMD cases and supports decision-making revealing hidden features of AMD that are not visible to the human eye. Future research will focus on improving these systems by expanding the databases in aggregate and incorporating multimodal data.

INDEX TERMS Age-related macular degeneration (AMD), computer aided diagnosis (CAD), deep learning (DL), eXplainable artificial intelligence (XAI).

I. INTRODUCTION

Age-related macular degeneration (AMD) is a degenerative eye disease affecting the macula of adults aged 50 years or older. In 2020, there were more than 196 million people affected by AMD worldwide (i.e. 8.7% of the total population) which is a leading cause of visual impairment and limited vision worldwide [1]. The number is expected to rise to more than 288 million by 2040 [2], of which about 113 million of these cases are expected in Asia [3].

In the United States, an estimated 19.8 million Americans aged 40 and older were living with AMD in 2019, including 1.49 million with vision-threatening AMD. The prevalence of AMD increases significantly with age, ranging from 2% among people aged 40-44 to 46.6% among those aged 85 and older [4]. Additionally, sex- and age-standardized rates of

The associate editor coordinating the review of this manuscript and approving it for publication was Genoveffa Tortora.

AMD were lower for non-Hispanic Black individuals (7.0%) compared to other racial and ethnic groups, such as non-Hispanic White individuals (13.3%) and Hispanic individuals (13.3%). AMD prevalence also showed significant state-level variation, from 6.2% in the District of Columbia to 18.3% in Florida.

In Europe, the prevalence and incidence of AMD are expected to increase by 15% and 75%, respectively, by 2050, mainly due to population aging. Among people aged from 45 to 85 years, the pooled prevalence of AMD is approximately 8.69% (95% confidence interval [CI], 4.26–17.40). This includes a prevalence of 12.33% (95% CI, 6.46-22.75) in people of European descent, 10.43% (95% CI, 5.27-20.01) in people of Hispanic descent, 7.53% (95% CI, 3.80-14.89) in people of African descent, and 7.38% (95% CI, 3.80-14.89) in people of Asian descent [5], as shown in Figure 1.

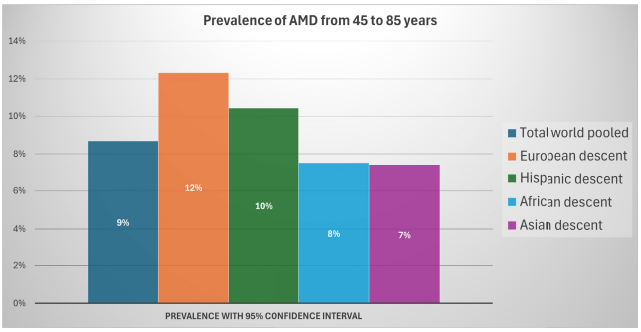


FIGURE 1. Bar chart illustrating pooled prevalence statistics of AMD based on data from of [5].

AMD occurs when aging damages the macula and it contains the highest concentration of photoreceptors in the retina which is crucial for high-resolution central visual acuity [6]. As we get older, debris, known as drusen, accumulates between the retinal pigment epithelium and Bruch's membrane. The presence of drusen is one of the first signs of AMD. There are two forms of AMD, dry and wet, each classified into different stages [7].

Dry AMD occurs in three stages: early, intermediate and late, the latter known as geographic atrophy or advanced non-neovascular AMD [8], see Figure 2. The figure illustrates the three stages of the dry AMD. Wet AMD, also called exudative or neovascular, systematically manifests in the late stage and is distinguished by two grades: inactive and active [9].

Furthermore, wet AMD can be classified as classic, occult or mixed. Both neovascular (wet) AMD and late-stage dry AMD are classified as advanced forms of the disease. Dry AMD is the most common form, wet AMD, although less common, accounts for 90% of AMD-related limited vision [10].

More and more is known about AMD due to new research, but it continues to be difficult to be diagnosed in the early stages, which is why the different challenges have been identified. According to the literature [11], [12], the following challenges exist:

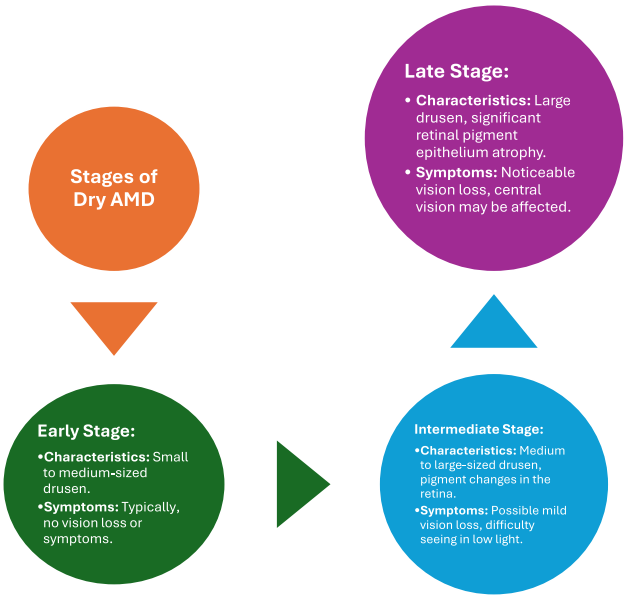


FIGURE 2. Graphical illustration of the three stages of Dry AMD.

- **Early detection**: Because the disease develops quietly and early, detecting it can be difficult. In traditional diagnosis, the lack of ability of current diagnostic techniques to detect subtle alterations in the retina such as visual acuity tests, fundus photography, and even conventional Optical Coherence Tomography (OCT), are often limited in their ability to detect early-stage retinal changes. Specifically, these methods lack the sensitivity required to capture subtle alterations in the retina, such as the slight thinning of photoreceptor layers in the ellipsoid zone. Saleh et al. [11] declared that in the early stages of AMD, there is a slight thinning of the photoreceptor layers, specifically in the ellipsoid zone, but this is difficult to detect without sophisticated imaging techniques.
- **Personalized care**: Given the different natures of AMD, personalized diagnostic methods and care are necessary. For instance, the technique of OCT has been modified to provide comprehensive depictions of retinal changes unique to various forms of AMD [12]. Hamid et al. [13] noted that deep learning algorithms trained on retinal images can help differentiate AMD types by identifying these patterns.
- **Case interpretation**: Determining the diagnosis and course of treatment can be difficult and subjective when interpreting images of the retina. Traditional diagnostic methods involve detailed assessments by clinicians using imaging techniques such as fundus photography and OCT to identify key biomarkers, including drusen deposits and changes in retinal thickness. However, these assessments can be influenced by the clinician's experience and interpretation skills, leading to variability in diagnosis. For instance, a case

can be diagnosed late or intermediate according to the experience of the doctor [11]. Karn and Abdulla [14] mentioned that clinician experience may make the task of assessing biomarkers on OCT images difficult. Classifying variations in retinal thickness or the presence of drusen can be difficult. Aartificial intelligence (AI) and machine learning (ML) technologies assist in this process to reduce subjectivity. By analyzing large datasets and learning to recognize patterns associated with AMD, AI models provide more standardized evaluations, reducing inter-clinician variability and enhancing diagnostic accuracy. These technologies offer objective analyses that support clinicians in making more precise and reliable diagnoses, ultimately improving patient outcomes and streamlining treatment decisions.

Because of these difficulties and challenges discussed in the literature, many research studies such as [15], [16], and [17] focused on creating accurate, effective, and personalized techniques to treat and diagnose AMD. Machine learning and AI have emerged as potential solutions to solve these obstacles. For example, [15], [18], [19] utilized vision transformers (ViTs), [16], [20] used convolutional neural networks (CNNs), and [17], [21] suggested hybrid approaches.

Following the literature, the current study aims to tackle the challenges of distinguishing between various grades of AMD (e.g., normal, have intermediate AMD, geographic atrophy (GA), or wet AMD). It mainly covers two research questions (RQ):

- **RQ1**: How can our AI-based system improve early detection and staging of AMD compared to traditional diagnostic approaches?
- **RQ2**: How does AI-assisted retinal image interpretation impact the efficacy and accuracy of AMD diagnosis and treatment options?

Based on these RQs, our major objective is to improve the classification and interoperability capabilities of our suggested approach using AI models. Compared to the literature, our approach provides the following significant phases and contributions:

- **Development of a non-invasive CAD system**: The study proposes a revolutionary CAD system that analyses retinal images for early identification of AMD using sophisticated machine learning techniques.
- **Implementation of AI Algorithms**: Different stages of AMD were classified using a variety of AI models in three-phase cascade method, such as CNN, Multi-Layer Perceptron (MLP), and Transformers, showing the effectiveness of these algorithms in medical imaging.
- **Hyperparameter Optimization**: The study utilized the Optuna library for hyperparameter tuning, enhancing the performance of the models and ensuring optimal training conditions.
- **Comprehensive Performance Evaluation**: The efficacy of the suggested models was evaluated using a variety of performance measures, including accuracy, sensitivity, specificity, and F1-score, which allowed for a comprehensive examination of their capabilities.
- **Contribution to Personalized Medicine**: Through the use of AI, the study increases the precision of AMD diagnosis, which advances customized treatment regimens and may improve patient outcomes.

This work holds significant value for the clinical sector as it assists physicians in interpreting diverse cases, thereby supporting their decision-making processes. Additionally, it offers various explainability perspectives for AMD and highlights hidden features that may not be visible to the human eye.

Paper Organization: The rest of this paper is organized as follows: Section II gives an overview of related studies in the field, Section III describes the research materials, Section IV describes the methodology used, Section V describes the experimental setup and results, Section VI presents the limitations of the current study, and Section VII concludes the paper and offers insights into future work.

II. RELATED STUDIES

One of the leading causes of vision loss in the elderly is AMD, as mentioned earlier. Because AMD requires individualized care and is characterized by subtle changes in the retina, diagnosis can be challenging as conventional techniques often miss early warning indicators and image interpretation is not always objective [1], [6], [11], [22], [23].

As noted, AI-based systems emerged as solutions for many diseases diagnoses including AMD [24], [25]. These AI-based systems includes Vision transformers (ViTs) that made recent advances in computer vision (CV), particularly in the area of medical image classification [26], [27], [28] such as Diabetic Retinopathy (DR) [27], macular edema [21], and AMD [15], [29].

ViT models have been demonstrated to be useful for utilizing a variety of retinal image analytics to detect a broad spectrum of retinal disorders. Examples of these conditions include AMD, glaucoma, choroidal neovascularization (CNV), and diabetic macular edema (DME) [18], [30].

ViT models are able to extract complex characteristics and patterns from retinal images. This allows to focus on the most important parts of the image and improves their ability to detect subtle anomalies. For instance, Gholami et al. [15] worked on detecting AMD using federated learning (FL) and deep learning (DL) models. They aimed to improve AMD diagnosis by training models on decentralized datasets from different institutions.

They built DL models with data augmentation techniques to handle variations in multiple datasets; implemented FL algorithms to train models locally on a variety of datasets; contrasted the results of multiple test sets between centralized and local models using different encoders. Centralized models were shown to be the best performing models for E = 100. In particular, the ResNet18 encoder showed remarkable accuracy rates, peaking at 94.58% $\pm$

0.62 on DS1. Furthermore, the ViT encoder demonstrated remarkable performance, achieving accuracy rates of 98.18% ± 0.55 and 99.11% ± 0.39 on the DS2 and DS3 test sets, respectively.

Moreover, Xu et al. in [17] examined DeepDrAMD, AI vision model designed with deep learning via hierarchical vision transformers to detect AMD using color fundus images. Two external databases and WMUEH internal dataset were used to train this model. The main objectives of the study were to build DeepDrAMD, and to evaluate its performance in recognizing AMD cases and dividing them into different subtypes such as wet AMD, dry AMD, and NVM type I and type II subtypes of wet AMD. The evaluation metrics yielded a AUC for DeepDrAMD was 96.47% on the WMUEH test set and 98.76% on the external database Ichallenge-AMD. Furthermore, the model showed discrimination capabilities evidenced by an AUC of 99.36% on the WMUEH test set to distinguish between MNV type I and type II subtypes of wet AMD.

Additionally, Yao et al. [21] used the FunSwin approach in to analyze fundus images to quantify the risk of macular edema and grade diabetic retinopathy. This technique extends the Swin Transformer framework, which has been improved and modified to provide features for medical data. They used the MESSIDOR dataset, which has 1,200 color fundus images from the posterior pole of the eye. The goals of the project were to use the FunSwin technique to quantify diabetic retinopathy and calculate the risk of macular edema. They obtained an accuracy of 0.9866 and an F1 score of 0.9896.

Furthermore, Yuka et al. [16] worked with a ViT segmentation model to identify the double layer sign (DLS) on OCT B-scans, because it is a conspicuous hallmark of nonexudative macular neovascularization (neMNV). The dataset they used included 251 eyes of 210 patients, including those with both DLS and drusen. The study wanted to create a deep learning algorithm that could detect neMNV from structural OCT images. The performance of the model got a sensitivity of 82%, specificity of 90%, positive predictive value (PPV) of 79%, negative predictive value (NPV) of 91%, and intersection over union (IoU) score of 61.11% for drusen segmentation and 58.80% for DLS. Furthermore, the model achieved an AUC value of 0.91.

Besides, the work of Jiang et al. [19] also made use of ViT. They used an OCT fundus image dataset from Duke University, which contained images from healthy individuals as well as AMD and DME patients. The objective was to create a CAD technique that correctly classified OCT fundus images for the detection of AMD and DME retinal disorders using the Vision Transformer. They achieved a classification accuracy of 99.69% in differentiating between AMD, DME, and normal images.

In addition, Akcca et al. [18] studied the use of ViTs for automated identification of retinal disorders. A publicly accessible dataset of 84,495 images divided into the four classes CNV, Drusen, DME, and Normal was used. By ensuring a balanced training and testing set, they selected 34,464 images, with an equal proportion for each class. The objectives of the study were to design a diagnostic system to minimize the delay of the increasing number of OCT scans, and demonstrate the effectiveness of transformer models in categorizing retinal diseases. They obtained an accuracy of 95.14% when the model was examined and the values indicated good performance on several categories of retinal disorders.

Instead of using patches and images, Hemalakshmi et al. [20] presented a classification framework, called CNN-RBF classifier, which used a hybrid model combining CNN with radial basis function (RBF) classifiers. This method uses features extracted from the multi-scale discriminant robust local binary pattern (MS-RLBP) method, which used gradient operations, local binary pattern (LBP) descriptors, and RGB component analysis to transform retinal images into numerical representations. The retinal images were further classified into four groups using the suggested framework: diabetic retinopathy (DR), AMD, CNV, and normal (NR). The effectiveness of the method was demonstrated with an accuracy of 97.22%, sensitivity of 96.49%, and specificity of 100%. These results show a progress in these methods and show that numerical features can be used to diagnose retinal disorders.

While previous research has explored the utility of AI-based systems such as ViTs to directly classify retinal images, the current research takes a different approach by focusing on the extraction of numerical features from masked retinal fundus images, instead of processing them directly.

III. MATERIALS

The main objective is to classify AMD grades using numerical features coloured fundus images from patients into different groups of AMD: normal, intermediate, GA and wet. The dataset consists of 216 images for each grade, so in total; the dataset comprises 864 images from the Comparison of AMD Treatment Trials (CATT) and another 864 masks for each of the images. 864 patients were recruited for a study through the Comparison of AMD Treatment Trials (CATT), sponsored by the University of Pennsylvania [31].

This study enrolled individuals over 50 years of age over a two-year period at 43 clinical centers in differents places of the United States. Participants received intravitreal injections of ranibizumab or bevacizumab according to one of three dosing regimens. Samples from the dataset are presented in Figure 3

All clinical and imaging data for this study were identified by the CATT Study Group before being submitted to the University of Louisville. All subjects provided informed consent, and data collection procedures complied with applicable laws and regulations in each setting.

Study treatments were administered through the CATT program during each participant's baseline appointment. Treatment was administered to participants in the monthly

TABLE 1. Summarization of the discussed related studies.

Reference	Year	Binary or Multiclass	Modality	Dataset	Model Type	Cons.	Outcome	Performance
Xu et al. [17]	2023	Binary Classification	CFP and OCT	3,764 CFP images	DeepDrAMD	Data augmentation overfitting	Classification	Accuracy of 96.47%
Yuka et al. [16]	2022	Binary Classification	OCT	648 OCT images	DCNN	the model should be applied only to SS-OCT scans with AMD.	classification, and segmentation	82%, 90%, 79%, and 91% sensitivity, specificity, PPV, and NPV,
Yao et al. [21]	2022	Multiclass classification	color fundus images	1,200 color fundus images of the posterior pole of the eye	FunSwin	Clinical evaluation, data Annotation and generalizability.	Classification	Average accuracy of 98.66%, with an F1-score of 98.96% and a sensitivity of 98.68%
Gholami et al. [15]	2023	binary classification	OCT	105,369 images from 3 datasets	Resnet18 and ViT	Performance Variability, Data Heterogeneity, Training Time and Bottlenecks, Inconsistent Performance, Need for Hyperparameter Tuning, Dependence on Data Quality.	Model Comparisons, Impact of Federated Learning Strategies and classification	Accuracy of 94.58% for Resnet18 and 98.18% for ViT
Akcca et al. [18]	2024	Multiclass classification	OCT	84,495 images	ViT, T2T-ViT, and Mobile-ViT	Experimenting with larger datasets is missing.	Classification	Accuracy of 95.14%.
Jiang et al. [19]	2022	Multiclass classification	OCT	3,231 OCT images	ViT	the classifier needs further tuning.	Speed Classification and model comparison	Overall classification accuracy of 99.69
Hemalakshmi et al. [20]	2021	Multiclass classification	Fundus	Not specified	CNN-RBF	Requires further tuning	Promising results in classification	Overall accuracy of 97.22%

dose groups at each visit or as needed. Participants assigned to the variable dose groups underwent treatment assessments at each visit.

The dataset is balanced across four AMD subtypes, with 216 images per class. To maintain this balance, images are divided into two folders: one folder contains GA and Intermediate classes, and the other contains Normal and Wet AMD classes. In the ''CustomDataset'' class, each folder is treated as a distinct group label, with each containing 432 images (216 images per subtype). The 'random_split' function ensures balanced splits for training, validation, and testing.

Using a cascade approach, the first classifier distinguishes between the two groups (GA/Intermediate vs. Normal/Wet), and a second classifier differentiates between the subtypes within each group. This method preserves class balance and allows for efficient classification of all AMD subtypes.

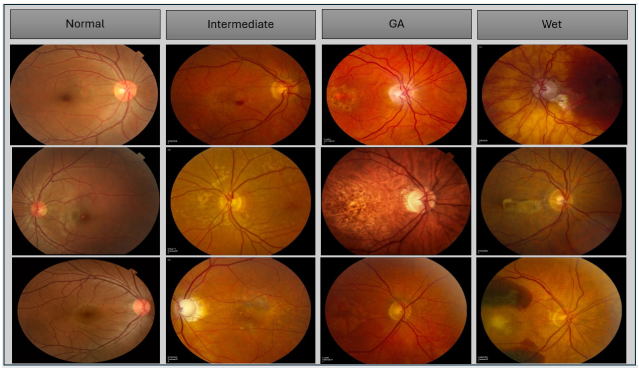


FIGURE 3. Samples form the utilized AMD dataset.

IV. METHODOLOGY

The proposed framework, presented in Figure 4, starts by employing a pre-input ViT model to extract numerical features after preprocessing fundus images to highlight regions of interest. A cascade technique of classifier models is then used to examine these features and Bayesian optimization is performed for hyperparameter tuning. Research has been conducted on three different types of classifiers: CNN, MLP, and ViT classifiers.

A. PREPROCESSING

The preprocessing procedure for this study is divided into three main stages to ensure that the input data is well-organized and full of information, several procedures are crucial: splitting and preparing the dataset for model training and evaluation; bitwise AND operation and image cropping; and numerical feature extraction using ViT. The structure of the preprocessing for the dataset is presented in Figure 5

1) BITWISE AND OPERATION AND IMAGE CROPPING

A two-phase preprocessing procedure was used. First, only the areas of interest (ROIs) were extracted by bitwise ANDing each fundus image with its corresponding mask. The images were scaled to ensure proper proportions and allow for accurate bitwise operations. Anatomical features in the images were given prominence.

After the first phase, a cropping process was used to remove the black border surrounding the ROIs to determine the bounding box around non-black pixels. Only the material of interest was kept. The cropped images were saved with a standard file name and format.

By focusing on the most relevant features, this integrated preprocessing method refines the dataset.

2) FEATURE EXTRACTION WITH ViT

Using ViTs for the first step in preprocessing the image data is the extraction of numerical features. We employ the "vit_base_patch16_224 model" [32], which has been trained on extensive datasets (e.g., Imagenet database [33]) to provide robust feature representations. The ability of this model to recognize complex patterns in images makes it particularly suitable for our use.

The first step of the procedure is to load the images from the designated location. Each image goes through a series of adjustments, including center cropping to 224 pixels, scaling to 256 pixels, and normalization using the mean and standard deviation values. Scaling to 256 ensures that each image is of the same size. The center crop is to ensure that the model processes the main object while removing unnecessary parts, the cropping stage is essential as it focuses on the most relevant part of the image. Pixel values were adjusted by normalization such that the mean and standard deviation values were obtained from the ImageNet dataset, exactly [0.485, 0.456, 0.406] for mean and [0.229, 0.224, 0.225] for standard deviation) and were used to standardize the pixel intensities of the image.

The ViT model needs images in a specific format, so these procedures standardize the input and make it compatible with the model. After being altered, the image is run through the selected ViT model to extract the features.

3) DATASET SPLITTING AND PREPARATION

After extracting the numerical features, the dataset is split into training, validation, and testing subsets to facilitate the model development and evaluation process. 70% of the data was split for training, 20% for validation and 10% for testing into three batch sizes 16, 32 and 64 to investigate how they affect the performance of the models.

B. CLASSIFICATION AND TUNING

In the cascade classification framework presented in this study, three different possible classification models are used. A cascade model is an advanced ensemble learning approach whereby predictions are refined by successively applying models, increasing overall reliability and accuracy [34]. The models operate sequentially, with each model generating a prediction and, based on the response of the previous model, activating one of the two subsequent models. The cascade remains active until a choice is reached or until the process is terminated [35].

1) MLP

This forward-propagating neural network consists of multiple layers of connected neurons and individual neurons that process input signals, apply a nonlinear activation function, and generate an output. The computation performed by a neuron can be represented by Equation 1 where y is the output of the neuron, f is an activation function (e.g., ReLU, Sigmoid, TanH), w_i are the weights associated with each

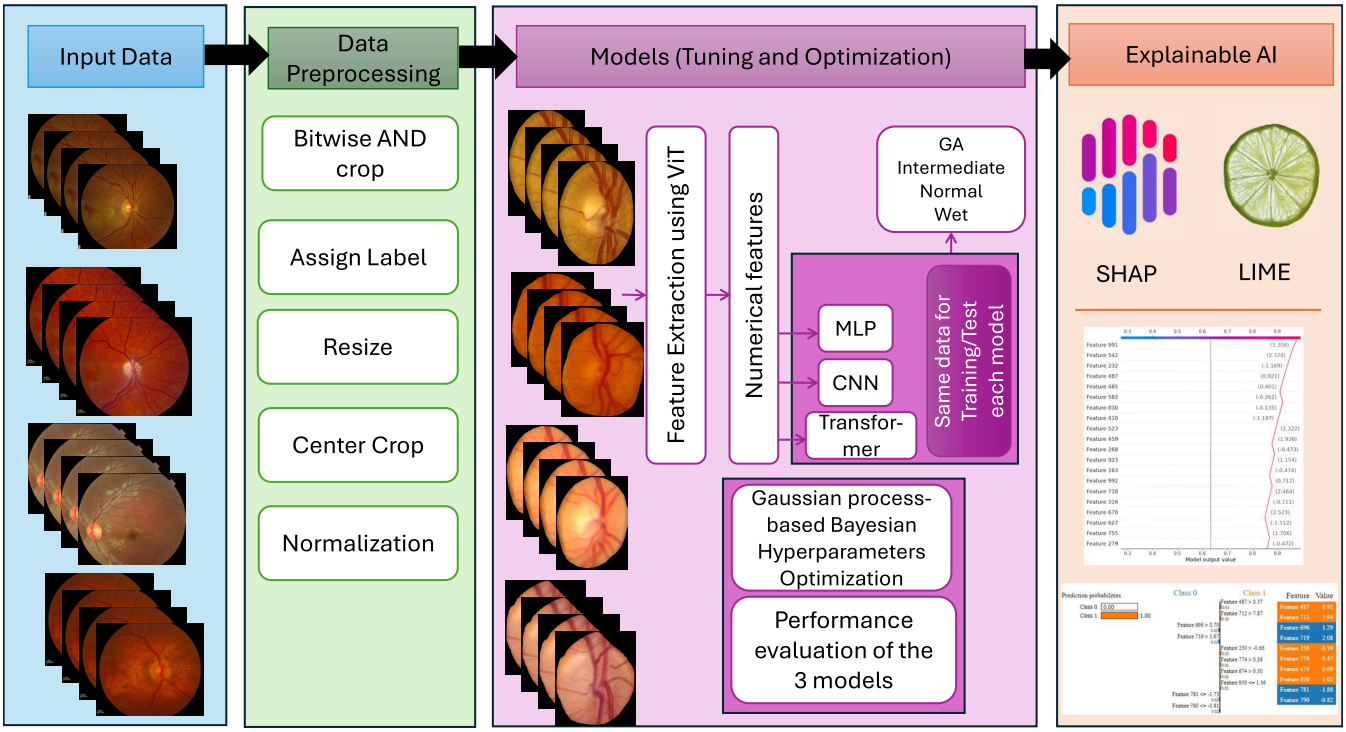


FIGURE 4. The proposed framework of the current study.

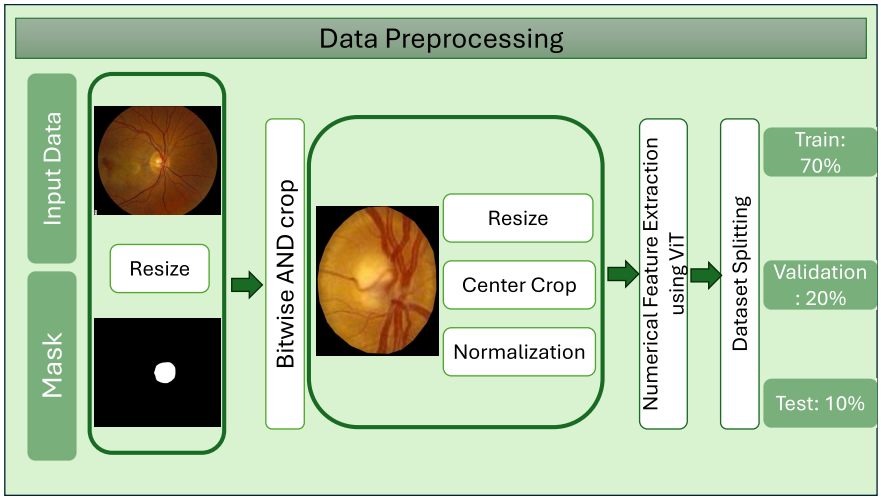


FIGURE 5. The utilized preprocessing approach.

input, x_i are the input signals, and b is the bias term.

$$y = f\left(\sum_{i=1}^{n} w_i \times x_i + b\right) \tag{1}$$

MLPs learn by iteratively changing weights and biases to minimize a loss function, which is known as backpropagation. For classification problems, the mean squared error (MSE) is a popular loss function, presented in Equation 2 where m is the number of data points, y_i is the true label, and $\hat{y}_i$ is the predicted label.

$$L = \frac{1}{m} \times \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{2}$$

2) CNN

It is a type of neural network used to analyze data that resembles a grid. Convolution is a concept that CNNs use to extract useful features from input data. The convolution process is a dot product of the input image region and the

filter, and creates a feature map by applying a series of filters, or kernels, to the input image using a convolutional layer. The resulting feature map captures local patterns and relationships within the image as presented in Equation 3 where $F_{i,j}$ is the element at row i and column j of the feature map, $K_{k,l}$ is the element at row k and column l of the filter, $I_{i+k-1,j+l-1}$ is the element at row $i+k-1$ and column $j+l-1$ of the input image, and b is the bias term.

$$F_{i,j} = f\left(\sum_{k=1}^{m}\sum_{l=1}^{n} K_{k,l} \times I_{(i+k-1,j+l-1)} + b\right) \quad (3)$$

Pooling layers are typically used to minimize the spatial dimensions of feature maps while still maintaining crucial information after the convolutional layer. Max pooling and average pooling are two popular pooling strategies.

Finally, the recovered features are combined with fully connected layers to create a classification or regression output.

3) TRANSFORMER CLASSIFIER

A transformer is a neural network design that can identify long-range dependencies in sequential input. Transformers handle input sequences in parallel. The self-attention mechanism is the core part of a transformer. Through this method, the model can determine which elements of the input sequence are the most important to process at a given time. The attention calculation is based on the dot product between the query, key, and value vectors, presented Equation 4 where Q, K, and V are matrices representing the query, key, and value vectors, respectively, and d_k is the dimension of the key vectors.

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V \quad (4)$$

Transformers can identify various patterns and correlations within the input sequence as numerous self-attention heads are employed. Additionally, positional encoding is used to provide the model with information about the relative position of each token in the sequence.

For classification tasks, a transformer can be used in conjunction with a classification head, such as a fully connected layer, to produce a probability distribution over possible class labels.

4) PHASE 1: INITIAL CLASSIFICATION

The initial stage of the $\mathcal{M}_1$ classifier involves processing the features and they are initially classified into two categories based on the indicators they present: those with GA or Intermediate conditions and those with normal or wet circumstances.

5) PHASE 2: SPECIALIZED DIAGNOSTICS

After the initial classification, depending on the result, $\mathcal{M}_2$ or $\mathcal{M}_3$ is activated:

- **GA or Intermediate Conditions**: For patients identified with GA or intermediate conditions, the $\mathcal{M}_2$ classifier is activated.
- **Normal or Wet Conditions**: For classifications of normal or wet conditions, the $\mathcal{M}_3$ classifier is activated.

The training and optimization of the classifier models were carried out using Bayesian optimization based on Gaussian processes. Unlike traditional grid search or random search, Bayesian optimization incorporates the previous evaluations, thus improving the efficiency and accuracy of hyperparameter tuning [36], as shown in Figure 6.

Bayesian optimization, as mathematically represented in Equation 5. Bayesian optimization differs from grid search or random search as it considers all historical evaluations.

$$\hat{x} \in \arg\max_{x \in X} f(x). \quad (5)$$

Key architectural factors were tuned for each of the classifier models. Each of these models was configured with unique features to maximize task performance. With a fixed input dimension and moving hidden dimensions, the MLP focused on collecting the data related to the retrieved features. The CNN's fully linked and convolutional layers were designed to detect spatial patterns in the data. Finally, the transformer classifier was tuned to handle the intricate interactions between data attributes.

C. ARCHITECTURE EXPLAINABILITY AND INTERPRETABILITY

For AI models to generate accurate and intelligible predictions in the context of AMD, it is imperative that they are understandable and interpretable. Applying techniques such as SHAP (Shap Additive Explanations) [37], [38] and LIME (Local Interpretable Model Explanations) [38], [39] could help achieve this.

SHAP values provide a unified measure of feature importance by attributing a model's prediction to its features using cooperative game theory. The SHAP value ϕ_i for a feature i is calculated as presented in Equation 6 where N represents the set of all features, S is a subset of N excluding feature i, and $f(S)$ denotes the model's prediction with the features in subset S.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \times [f(S \cup \{i\}) - f(S)] \quad (6)$$

LIME explains individual predictions by locally approximating the model with a simpler, interpretable model. The approach involves perturbing the input data and observing changes in predictions. The explanation model g is trained to minimize the following objective as presented in Equation 7 where $\mathcal{L}$ is the loss function (e.g., mean squared error), f is the original model, π_x is the locality measure around instance x, and $\Omega(g)$ denotes the complexity of the explanation model g.

$$\xi(x) = \arg\min_{g \in G} [\mathcal{L}(f, g, \pi_x) + \Omega(g)] \quad (7)$$

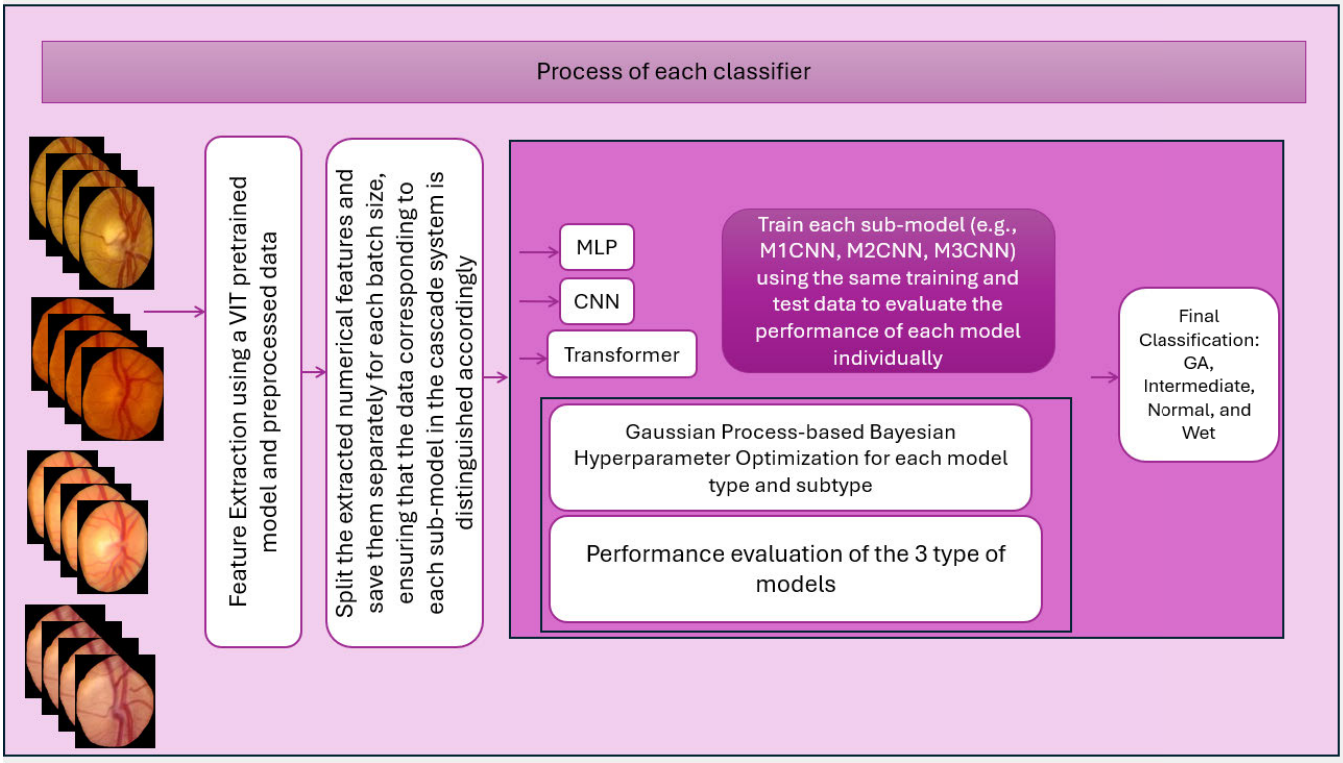


FIGURE 6. The process of each classifier for training.

In diagnosing and treating AMD, these interpretability methods can significantly aid clinicians by providing insights into how machine learning models make decisions based on retinal images. As an example:

- SHAP scores can show which features may be most important in predicting the course of AMD. SHAP values quantify the contribution of each feature to the model prediction for individual cases. For AMD, this might reveal that certain retinal layers or specific drusen features have a greater influence on disease progression. SHAP analysis can explain nonlinear connections and feature interactions, showing how multiple biomarker combinations affect classification.
- LIME offers local interpretations of each prediction's classification, helping clinicians understand the model's results. For each prediction, LIME builds a more understandable and simplified model that explains how the model reached its conclusion. LIME might identify the precise imaging features or clinical data points that were most important in determining a particular risk assessment for AMD patients. This patient-tailored explanation can help improve treatment strategies and patient communication about risk factors. LIME helps clinicians understand the outputs of models by providing local explanations for each patient's prognosis.

Having SHAP and LIME together offers complementary interpretation of the trained models. SHAP is ideal for identifying global trends and interactions between patient population characteristics, providing to clinicians a general understanding of key factors in AMD progression. LIME, on the other hand, allows clinicians to interpret individual patient cases in detail, tailoring explanations to specific diagnostic scenarios. By bringing the two methods together, clinicians can rely on a comprehensive and interpretable framework with SHAP to get general patterns and LIME for patient-specific information. This dual approach can improve the transparency and reliability of AI models, transforming this tools for diagnostic support in AMD.

D. PERFORMANCE ASSESSMENT

Some performance measures were selected to see and quantify the performance of each model on the classification tasks. These measures provide crucial details about the effectiveness and reliability of the suggested approach.

Accuracy is calculated dividing the total number of samples by the number of samples that were correctly classified. Higher accuracy shows that the classifier model is more able to accurately determine which cases are good and bad.

Sensitivity is the rate of true positives or recall and measures the ability of the classifier model to identify positive cases. It is calculated as the ratio of true positive samples to the sum of true positive and false negative samples. A higher sensitivity is important to reduce false negatives.

Specificity measures the ability of the classifier model to identify negative cases. It is calculated as the ratio of true negative samples to the sum of true negative samples and false positive samples. In order to reduce false positives, a greater specificity value is crucial.

Precision represents the proportion of correctly classified positive cases among all classified positive cases. It is calculated as the ratio of true positive samples to the sum of true positive samples and false positive samples. A lower rate of false positives is indicated by a higher accuracy value.

The F1 Score is the harmonic mean of sensitivity and accuracy and offers a compromise between sensitivity and accuracy, making it useful when working with skewed datasets. A higher F1 score indicates a better balance between these two parameters.

Balanced Accuracy is a metric that takes into account the true positive and true negative rates and is calculated by averaging the sensitivity and specificity. This metric is particularly useful for imbalanced datasets.

The Dice Coefficient, also known as the F1 score, measures the similarity between the predicted and actual segmentation cases. It is calculated as the ratio of the overlap between the two to the total number of pixels in both regions and provides a balanced evaluation, rewarding the precision and recall.

The false negative rate (FNR) measures the proportion of true positive cases that are incorrectly classified as negative. It is calculated as the ratio of false negatives to the total number of true positives. The FNR is good to see the model's inability to correctly identify positive cases, as a lower FNR indicates better performance in detecting positive cases [40].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{10}$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{13}$$

V. EXPERIMENTS AND DISCUSSION

Experiments Configurations: The classification models were trained using a dedicated local workstation with an NVIDIA Quadro M4000 graphics card [41]. It offers sufficient processing power for demanding deep learning applications. The purpose of this hardware configuration was to specifically utilize the parallel processing power of the GPU, which is essential for accelerating model training and enabling efficient manipulation of system files.

Python 3.9 [42] and PyTorch [43], a deep learning framework well known for its adaptability and effectiveness in research settings, were used to run the model building, training, validation, and testing processes.

Scikit-Learn [44] is a Python AI package that was used to generate evaluation metrics such as specificity, confusion matrix, F1, accuracy, and sensitivity. Two Python libraries were used to produce publication-quality figures and analyze the data, Matplotlib [45] and SeaBorn [46]. These tools improved the interpretability and clarity of the study data by facilitating the creation of visual representations such as plots of performance metrics, confusion matrices showing classification results, feature extraction results, and activation maps.

The target of the experiments is to answer the current study research questions. This is accomplished by applying machine learning algorithms such as CNN, MLP, and Transformer Classifier to examine the numerical features collected from retinal images. To ensure that each model is optimized, these models are trained and evaluated using a set of experiments including early stopping, loss, and precision tracking. Performance metrics of precision, accuracy, recall, and sensitivity are calculated through test runs. Using the Optuna library, hyperparameters of each model are tuned and the results of each tuning are saved to evaluate how well the models solve the key research problems.

A. HYPERPARAMETER TUNING

To maximize computational effectiveness and training stability across several models, we tested batch sizes of 16, 32, and 64. Early stopping was employed to avoid overfitting, and learning rates were constantly adjusted using the Adam optimizer [47], which was based on suggestions from Optuna. The best hyperparameters with the best accuracy were found after 300 attempts.

1) MLP

The MLP was designed with an input dimension of 1000, reflecting the number of extracted features. Hidden dimensions were optimized between 64 and 1,024, and the output dimension was set to 2 for binary classification.

For batch size 16, $\mathcal{M}_1$ was optimized with a hidden dimension of 856, a learning rate of 1.596×10^{-4}, and 146 epochs, while $\mathcal{M}_2$ had a hidden dimension of 322, a learning rate of 6.46×10^{-5}, and 191 epochs. $\mathcal{M}_3$, for the same batch size, featured a hidden dimension of 245, a learning rate of 2.93×10^{-5}, and 93 epochs.

With a batch size 32, $\mathcal{M}_1$ was configured with a hidden dimension of 116, a learning rate of 9.06×10^{-5}, and trained for 184 epochs. $\mathcal{M}_2$, under the same batch size, had a hidden dimension of 384, a learning rate of 5.291×10^{-4}, and trained for 137 epochs, while $\mathcal{M}_3$ had a hidden dimension of 471, a learning rate of 4.603×10^{-4}, and completed 81 epochs.

Finally, for batch size 64, $\mathcal{M}_1$ had a hidden dimension of 826, a learning rate of 3.258×10^{-4}, and trained for 97 epochs. $\mathcal{M}_2$ featured a hidden dimension of 663, a learning rate of 9.723×10^{-4}, and trained for 187 epochs,

TABLE 2. Quantitative metrics for evaluating AMD detection performance.

Performance	Equation	Definition
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	The proportion of observations successfully classified
Specificity	$\frac{TN}{FP+TN}$	The proportion of correctly predicted negative cases
Recall (or Sensitivity)	$\frac{TP}{TP+FN}$	The proportion of correctly predicted positive cases
Precision	$\frac{TP}{TP+FP}$	The proportion of true-positives to the overall number of positively identified samples
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$	The proportion of true negatives to the overall number of negative predictions
F1-score	$\frac{2\times(Precision\times Recall)}{Precision+Recall}$	The harmonic mean of the precision and recall of a classifier
Balanced Accuracy (BAC)	$\frac{Recall+Specificity}{2}$	The arithmetic mean of the sensitivity and specificity (more suitable for imbalanced data)
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$	The proportion of actual positive cases that are incorrectly classified as negative by the model (lower FNR indicates better performance in detecting positive instances)

while $\mathcal{M}_3$ was optimized with a hidden dimension of 900, a learning rate of 1.32×10^{-4}, and trained for 83 epochs.

2) CNN

The CNN optimization design included a first convolutional layer with output channels ranging from 16 to 64. The fully connected layer was optimized with features ranging from 128 to 512, while the second convolutional layer had output channels ranging from 32 to 128.

For Batch Size 16, $\mathcal{M}_1$ was configured with 52 output channels in the first convolutional layer and 58 in the second, along with 275 output features in the fully connected layer. The learning rate was set to 5.496×10^{-4}, and the model was trained for 20 epochs. $\mathcal{M}_2$, on the other hand, had 50 and 123 output channels in the first and second convolutional layers, respectively, with 421 output features in the fully connected layer, a learning rate of 8.052×10^{-4}, and was trained for 50 epochs. $\mathcal{M}_3$ featured 32 and 106 output channels, with 495 output features, a lower learning rate of 1.972×10^{-4}, and a considerably longer training duration of 148 epochs.

For Batch Size 32, $\mathcal{M}_1$ was optimized with 54 output channels in the first convolutional layer and 110 in the second, along with 475 output features in the fully connected layer, a learning rate of 4.464×10^{-4}, and 20 epochs of training. $\mathcal{M}_2$ included 38 and 117 output channels, with 324 output features, a learning rate of 7.197×10^{-4}, and was trained for 91 epochs. $\mathcal{M}_3$ was set with 34 and 106 output channels, 230 output features, a learning rate of 3.109×10^{-4}, and 26 training epochs.

For Batch Size 64, $\mathcal{M}_1$ consisted of 30 output channels in the first convolutional layer and 81 in the second, with 134 output features in the fully connected layer, a learning rate of 1.406×10^{-4}, and 45 epochs of training. $\mathcal{M}_2$ had 56 and 120 output channels, with 300 output features, a learning rate of 5.652×10^{-4}, and was trained for 128 epochs. $\mathcal{M}_3$ used 32 and 120 output channels, 420 output features, a learning rate of 3.352×10^{-4}, and 42 training epochs.

3) TRANSFORMER CLASSIFIER

The Transformer model was fit to an input dimension of 1000 and the number of layers was varied from 1 to 6. The number of attention heads was optimized between 1 and 8, and the model dimension between 64 and 1024 to ensure divisibility by the number of attention heads for efficient multi-head attention.

For a batch size of 16, $\mathcal{M}_1$ achieved optimal performance with a single layer, six attention heads, and a model dimension of 640. It utilized a learning rate of approximately $1.256 \times 10^{-5}$and was trained for 23 epochs. $\mathcal{M}_2$ also used one layer and six attention heads but had a larger model dimension of 940, the learning rate was 4.593×10^{-5}, with 17 epochs. $\mathcal{M}_3$ was configured with a setup with four layers, eight attention heads, and a model dimension of 377 with a learning rate of 4.22×10^{-5} and was trained for 20 epochs.

For a batch size of 32, the optimal configuration for $\mathcal{M}_1$ included four layers, four attention heads, a model dimension of 559, a learning rate of $5.066 \times 10^{-5}$and was trained for 17 epochs. $\mathcal{M}_2$ had two layers, four attention heads, a model dimension of 554, a higher learning rate of 1.73×10^{-4} and was trained for 12 epochs. $\mathcal{M}_3$ in this batch was optimized for six layers, six attention heads, model dimension of 866, learning rate of 2.57×10^{-5}, and it was trained for 24 epochs.

With a batch size of 64, $\mathcal{M}_1$ was configured with one layer, two attention heads, a model dimension of 245, a learning rate of $2.19 \times 10^{-5}$and was trained for 44 epochs. $\mathcal{M}_2$ was optimized with three layers, four attention heads, a large model dimension of 1,200, a learning rate of $8.83 \times 10^{-5}$and was trained for 26 epochs. Finally, $\mathcal{M}_3$ had one layer, seven attention heads, a learning rate of 12.6×10^{-5} and a model dimension of 547.

B. REPORTED RESULTS

1) MLP

We present the testing metrics for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ across different batch sizes as follows:

For a batch size of 16, $\mathcal{M}_1$ got a test loss of 0.8627, an accuracy of 82.56%, a F1-Score of 81.02%, a FNR score of 0.1795 and a sensitivity of 82.05%. The BAC and NPV were both 82.05%, and the confusion matrix got 39 true positives, 8 false positives, 7 false negatives, and 32 true negatives. $\mathcal{M}_2$ had a test loss of 0.7575, an accuracy of 85.71%, a F1-Score of 87.89%, a FNR score of 0.1852 and a sensitivity of 81.48%, with BAC and NPV values also at 81.48%,

TABLE 3. The utilized hyperparameters for the Transformer, MLP, and CNN models.

Transformer Classifier		
Hyperparameter	**Range**	**Step**
Number of layers (L)	1 to 6	1
Number of attention heads (H)	1 to 8	1
Model dimension (dim)	64 to 1024	64
Learning rate (γ)	10^{-5} to 10^{-3}	log
Batch size (BS)	16, 32, 64	-
MLP		
Hyperparameter	**Range**	**Step**
Input dimension (d_{input})	1000	-
Hidden dimension (d_{hidden})	64 to 1024	64
Output dimension (d_{output})	2	-
Learning rate (γ)	10^{-5} to 10^{-3}	log
Batch size (BS)	16, 32, 64	-
CNN		
Hyperparameter	**Range**	**Step**
Convolution layer 1 out channels (ϕ_1)	16 to 64	8
Convolution layer 2 out channels (ϕ_1)	32 to 128	16
Fully connected layer 1 out features (ψ_{fc1})	128 to 512	32
Learning rate (γ)	10^{-5} to 10^{-3}	log
Batch size (BS)	16, 32, 64	-

and its confusion matrix showed 14 true positives, 1 false positive, 5 false negatives, and 22 true negatives. $\mathcal{M}_3$ got the highest performance with a test loss of 0.2132, an accuracy of 95.45%, a F1-Score of 95.00%, a FNR score of 0.0.05 and a sensitivity of 95.00%. Both BAC and NPV were 95.00%, supported by its confusion matrix of 23 true positives, 1 false positive, 1 false negative, and 19 true negatives.

With a batch size of 32, $\mathcal{M}_1$ had a test loss of 0.2639, an accuracy of 88.37%, a F1-Score of 87.29%, a FNR score of 0.0811 and a sensitivity of 91.89%, with corresponding BAC and NPV values of 91.89%. This was reflected in its confusion matrix of 42 true positives, 7 false positives, 3 false negatives, and 34 true negatives. $\mathcal{M}_2$ reported a test loss of 0.6390, an accuracy of 85.71%, a F1-Score of 85.00%, a FNR score of 0.1053 and a sensitivity of 89.47%, with BAC and NPV at 89.47%, derived from its confusion matrix of 19 true positives, 4 false positives, 2 false negatives, and 17 true negatives. $\mathcal{M}_3$ achieved a test loss of 0.1177, an accuracy of 95.45%, a F1-Score of 94.12%, a FNR score of 0.0588 and a sensitivity of 94.12%, with BAC and NPV values of 94.12%, supported by its confusion matrix of 26 true positives, 1 false positive, 1 false negative, and 16 true negatives.

For a batch size of 64, $\mathcal{M}_1$ recorded a test loss of 0.3656, an accuracy of 91.86%, a F1-Score of 90.91%, a FNR score of 0.1026 and a sensitivity of 89.74%, with BAC and NPV also at 89.74%, based on its confusion matrix with 44 true positives, 3 false positives, 4 false negatives, and 35 true negatives. $\mathcal{M}_2$ had a test loss of 0.5158, an accuracy of 92.86%, a F1-Score of 91.30%, a FNR score of 0.0526 and a sensitivity of 94.74%, with BAC and NPV values of 94.74%, as shown by its confusion matrix of 21 true positives, 2 false positives, 1 false negative, and 18 true negatives. $\mathcal{M}_3$ achieved a test loss of 0.4102, an accuracy of 90.91%, a F1-Score of 92.31%, a FNR score of 0.1429 and a sensitivity of 85.71%, with BAC and NPV at 85.71%, reflected in its confusion matrix of 16 true positives, 0 false positives, 4 false negatives, and 24 true negatives.

The cascade test results for different batch sizes are summarized in the following part:

For a batch size of 16, the cascade method achieved an accuracy of 89.53%, a F1-Score of 89.85% and a FNR score of 0.956. The confusion matrix showed 37 true positives, 2 false positives, 7 false negatives, and 40 true negatives. Sensitivity of the classes $\mathcal{M}_3$ was 85.11%, and of the classes $\mathcal{M}_2$ was 94.87%. The BAC was calculated as 90.09%, and the NPV was 85.11%.

With a batch size of 32, the accuracy improved to 91.86%. The confusion matrix indicated 45 true positives, 5 false positives, 2 false negatives, and 34 true negatives. Sensitivity of the classes $\mathcal{M}_3$ was 94.44%, and for classes of $\mathcal{M}_2$ was 90.00%. The BAC was 92.22%, F1-Score was 90.73%, FNR score was 0.0749 and the NPV was 94.44%.

For a batch size of 64, the accuracy was slightly reduced to 88.37%. The confusion matrix displayed 35 true positives, 4 false positives, 6 false negatives, and 41 true negatives. Sensitivity of the classes $\mathcal{M}_3$ was 87.23%, and of the classes $\mathcal{M}_2$ was 89.74%. The BAC was 88.49%, F1-Score was 90.42%, FNR score was 0.1148 and the NPV was 87.23%.

2) CNN

We present the testing metrics for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ across different batch sizes as follows:

For $\mathcal{M}_1$ with a batch size of 16, the model achieved a Test Loss of 0.8465, an Accuracy of 82.56%, a FNR score of 0.2051 and a Sensitivity of 79.49%. The F1-Score was 80.92%, the NPV was 85.11%, and the BAC was 82.75%. These metrics indicate that while the model performed reasonably well, there was room for improvement in sensitivity. $\mathcal{M}_2$ with the same batch size showed a Test Loss of 1.2267, an Accuracy of 80.95%, a Sensitivity of 77.78%, a F1- Score of 76.92%, a FNR score of 0.2222, the NPV of 81.25%, and the BAC was 78.57%, the performance in the accuracy and sensitivity was lower compared to $\mathcal{M}_1$. $\mathcal{M}_3$, also with a batch size of 16, demonstrated superior performance with a Test Loss of 0.4867, an Accuracy of 90.91%, a Sensitivity of 90.00%, a F1-Score of 90.32%, a NPV of 91.67%, a FNR score of 0.1 and a BAC of 90.91%, this is the most effective model in this batch size category.

For $\mathcal{M}_1$ with a batch size of 32, the model recorded a Test Loss of 0.3203, an Accuracy of 91.86%, a FNR score of 0.0270 and a Sensitivity of 97.30%. The F1-Score was 94.56%, the NPV was 97.67%, and the BAC was 93.45%, showing better results than the batch size of 16. $\mathcal{M}_2$ at this batch size achieved a Test Loss of 0.5914, an Accuracy of 92.86%, a FNR score of 0.0526 and a Sensitivity of 94.74%. The F1-Score was 93.75%, the NPV was 95.45%, and the BAC was 92.77%, indicating better results but not as high as Model 1 in this batch size. $\mathcal{M}_3$ obtained the best performance with a batch size of 32, showing a Test Loss of 0.1907, an Accuracy of 95.45%, a Sensitivity of 88.24% a F1-Score

of 93.10%, a NPV of 96.43%, a FNR score of 0.1176 and a BAC of 91.93%, showing the best results among all models in this batch size.

For $\mathcal{M}_1$ with a batch size of 64, the model got a Test Loss of 0.3202, an Accuracy of 93.02%, a Sensitivity of 94.87%, a F1-Score of 93.86%, a NPV of 95.83%, a FNR score of 0.0513 and a BAC of 92.91%. $\mathcal{M}_2$ with this batch size had a Test Loss of 0.4620, an Accuracy of 90.48%, a Sensitivity of 94.74%, a F1-Score of 91.18%, a NPV of 95.24%, a FNR score of 0.0526 and a BAC of 92.14%, indicating lower metrics than $\mathcal{M}_1$. $\mathcal{M}_3$ got a Test Loss of 0.4979, an Accuracy of 90.91%, a Sensitivity of 85.71%, a F1-Score of 91.43%, a NPV of 94.12%, a FNR score of 0.1277 and a BAC of 90.25%, reflecting a good performance with minor tradeoffs compared to the other models in this batch size category.

The cascade method was applied to each batch size to check the overall performance of the model. For the batch size of 16, the method achieved an accuracy of 86.05%. The confusion matrix showed a balanced performance across classes, with a sensitivity of 82.98% for the classes $\mathcal{M}_3$ and 89.74% for the classes $\mathcal{M}_2$. The F1-Score was 86.33%, the NPV was 89.77%, the FNR score was 0.1702 and the BAC was 86.36%. When applied to the models trained with a batch size of 32, the cascade method achieved the best results, with an accuracy of 94.19%, a F1-Score of 93.98%, a NPV of 96.08%, a FNR score of 0.0.0833 and a BAC of 93.83%. The sensitivity was 91.67% for the classes $\mathcal{M}_3$ and 96.00% for classes $\mathcal{M}_2$.

For the largest batch size of 64, the cascade method obtained an accuracy of 88.37%, a sensitivity of 87.23% for the $\mathcal{M}_3$ classes and 89.74% for the $\mathcal{M}_2$ classes, with an F1-Score of 88.39%, an NPV of 89.74%, a FNR score of 0.1429 and a BAC of 88.48%.

The cascade approach got the best metrics with the batch size of 32, as it was the most accurate and balanced results.

3) TRANSFORMER CLASSIFIER

We present the testing metrics for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ across different batch sizes as follows:

For a batch size of 16, $\mathcal{M}_1$ got a test loss of 0.7770, an accuracy of 83.72%, a sensitivity of 84.62%, which translates to a BAC of approximately 82.05% a F1-Score of 82.50%, a FNR score of 0.1538 and an NPV of 86.75%. $\mathcal{M}_2$, with a test loss of 0.7456 and an accuracy of 85.71%, showed a sensitivity of 81.48%, a BAC of around 81.48% a F1-Score of 87.93%, a FNR score of 0.1852 and an NPV of 87.38%. $\mathcal{M}_3$ achieved a lower test loss of 0.2652, an accuracy of 93.18%, perfect sensitivity of 100%, a F1-Score of 93.02%, a FNR score of 0.0, and both BAC and NPV were also 100%.

When evaluated with a batch size of 32, $\mathcal{M}_1$ recorded a test loss of 0.3222 a F1-Score of 90.93%, a FNR score of 0.0541 and an accuracy of 91.86%. It had a sensitivity of 94.59%, corresponding to a BAC and NPV of approximately 94.59%. $\mathcal{M}_2$'s performance included a test loss of 0.3193 and an accuracy of 88.10%, with a sensitivity of 84.21%, a F1-Score of 86.49%, a FNR score of 0.1579,leading to a BAC and NPV of around 84.21%. $\mathcal{M}_3$ achieved a test loss of 0.2164, a F1-Score of 90.90%, a FNR score of 0.1176 and an accuracy of 93.18%. Its sensitivity was 88.24%, with the BAC and NPV both approximating 88.24%.

For a batch size of 64,$\mathcal{M}_1$ got a test loss of 0.4669, an accuracy of 90.70%, with a sensitivity of 89.74%, a F1-Score of 89.74%, a FNR score of 0.1026, a BAC and NPV of about 89.74%. $\mathcal{M}_2$ had a test loss of 0.5705 and an accuracy of 90.48%, with a sensitivity of 94.74%, a F1-Score of 89.95%, a FNR score of 0.0526, corresponding to a BAC and NPV of 94.74%. $\mathcal{M}_3$, got higher test loss of 0.4297 and an accuracy of 90.91%, a sensitivity of 85.71%, a F1-Score of 92.31%, a FNR score of 0.1429, with the BAC and NPV both around 85.71%.

The following findings provide an overview of the performance of the cascaded transformer model for various batch sizes.

For a batch size of 16, the cascade model achieved an accuracy of 83.72%, a F1-Score of 85.31%, a FNR score of 0.1277 with a sensitivity for $\mathcal{M}_3$ classes of 87.23%, while for $\mathcal{M}_2$ classes it was 79.49%. The BAC for this batch size is approximately 83.36% and the NPV is 86.21%.

For a batch size of 32, the cascade model recorded an accuracy of 77.91%, a F1-Score of 73.24% and a FNR score of 0.2778. The sensitivity for $\mathcal{M}_3$ class was 72.22%, while for $\mathcal{M}_2$ classes it was 82.00%. This model got a BAC of approximately 77.11% and a NPV of 72.22%.

With a batch size of 64, the cascade model achieved an accuracy of 81.40%, a F1-Score of 81.46% and a FNR score of 0.2553. The sensitivity for$\mathcal{M}_3$ classes was 74.47% and for $\mathcal{M}_2$ classes, it was 89.74%. The BAC for this batch size was approximately 82.11% and the NPV was 81.98%.

The findings presented provide an analysis of the three models used to categorize AMD. The models were evaluated with different batch sizes and using a number of criteria. Overall, in terms of the different metrics, the MLP and CNN consistently performed better than the Transformer classifier. The optimal batch size varied between the different models, but a batch size of 32 yielded the best metrics in most cases.

Research using MLP, CNN, and Transformer classifier models provides evidence to answer RQ1, results show that across a range of evaluation metrics, MLP and CNN models outperformed the Transformer classifier.It is evident that for a batch size of 16, the MLP model, namely $\mathcal{M}_3$, obtained an accuracy of 95.45% and a sensitivity of 95.00%. The observed enhancement in performance implies that artificial intelligence and machine learning algorithms greatly enhance the precision of AMD early detection and staging, signifying a breakthrough in comparison to conventional diagnostic techniques.

Regarding RQ2, the findings of the study demonstrate the influence of AI-assisted models on diagnostic accuracy. The MLP model proved to be valid for accurately recognizing AMD cases, as evidenced by its balanced accuracy (BAC) of 95.00% and a perfect sensitivity for $\mathcal{M}_3$.

TABLE 4. Summarization of the obtained results for the different configurations and classification models. **Bold** results are for the cascaded suggested approach. **Blue** results are the best reported metrics.

Model	Loss	ACC (%)	REC (%)	SPEC (%)	F1 (%)	NPV (%)	BAC (%)	FNR (%	Hyperparameters
$\mathcal{M}_1MLP$	0.8627	82.56	82.05	82.98	81.02	84.78	82.52	0.1795	d_{input}: 1000, d_{hidden}: 856, d_{output}: 2, γ: 1.596×10^{-4}, ϵ: 146, BS: 16
$\mathcal{M}_2MLP$	0.7575	85.71	81.48	93.33	87.89	73.68	87.41	0.1852	d_{input}: 1000, d_{hidden}: 322, d_{output}: 2, γ: 6.46×10^{-5}, ϵ: 191, BS: 16
$\mathcal{M}_3MLP$	0.2132	95.45	95.00	95.83	95.00	95.83	95.42	0.05	d_{input}: 1000, d_{hidden}: 245, d_{output}: 2, γ: 2.93×10^{-5}, ϵ: 93, BS: 16
Cascaded	-	**89.53**	**89.99**	**84.09**	**89.85**	**84.09**	**89.99**	**0.956**	BS: 16
$\mathcal{M}_1MLP$	0.2639	88.37	91.89	85.71	87.29	93.33	88.80	0.0811	d_{input}: 1000, d_{hidden}: 116, d_{output}: 2, γ: 9.06×10^{-5}, ϵ: 184, BS: 32
$\mathcal{M}_2MLP$	0.6390	85.71	89.71	82.61	85.00	90.48	86.04	0.1053	d_{input}: 1000, d_{hidden}: 384, d_{output}: 2, γ: 5.291×10^{-4}, ϵ: 137, BS: 32
$\mathcal{M}_3MLP$	0.1177	95.45	94.12	96.29	94.12	96.29	95.21	0.0588	d_{input}: 1000, d_{hidden}: 471, d_{output}: 2, γ: 4.603×10^{-4}, ϵ: 81, BS: 32
Cascaded	-	**91.86**	**92.22**	**95.74**	**90.73**	**95.53**	**92.22**	**0.0749**	BS: 32
$\mathcal{M}_1MLP$	0.3656	91.86	89.74	93.62	90.91	91.67	91.68	0.1026	d_{input}: 1000, d_{hidden}: 826, d_{output}: 2, γ: 3.258×10^{-4}, ϵ: 97, BS: 64
$\mathcal{M}_2MLP$	0.5158	92.86	94.74	91.30	91.30	65.45	93.02	0.0526	d_{input}: 1000, d_{hidden}: 663, d_{output}: 2, γ: 9.723×10^{-4}, ϵ: 187, BS: 64
$\mathcal{M}_3MLP$	0.4102	90.91	85.71	100	92.31	80.00	92.86	0.1429	d_{input}: 1000, d_{hidden}: 900, d_{output}: 2, γ: 1.32×10^{-4}, ϵ: 83, BS: 64
Cascaded	-	**88.37**	**88.49**	**85.37**	**90.42**	**85.37**	**89.74**	**0.1148**	BS: 64
$\mathcal{M}_1CNN$	0.8465	82.56	79.49	85.11	80.52	83.33	82.30	0.2051	ϕ_1: 52, ϕ_2: 58, ψ_{fc1}: 275, γ: 5.496×10^{-4}, ϵ: 20, BS: 16
$\mathcal{M}_2CNN$	1.2267	80.95	77.78	86.67	84.09	83.33	82.30	0.2222	ϕ_1: 50, ϕ_2: 123, ψ_{fc1}: 421, γ: 8.052×10^{-4}, ϵ: 50, BS: 16
$\mathcal{M}_3CNN$	0.4867	90.91	90.00	91.67	90.00	91.67	90.83	0.1	ϕ_1: 32, ϕ_2: 106, ψ_{fc1}: 495, γ: 1.972×10^{-4}, ϵ: 148, BS: 16
Cascaded	-	**86.05**	**86.36**	**89.74**	**92.30**	**81.40**	**92.90**	**0.1702**	BS: 16
$\mathcal{M}_1CNN$	0.3203	91.86	97.30	87.76	91.15	97.73	92.53	0.0270	ϕ_1: 54, ϕ_2: 110, ψ_{fc1}: 475, γ: 4.464×10^{-4}, ϵ:20, BS: 32
$\mathcal{M}_2CNN$	0.5914	92.86	94.74	91.30	92.30	95.45	93.02	0.0526	ϕ_1: 38, ϕ_2: 117, ψ_{fc1}: 324, γ: 7.197×10^{-4}, ϵ: 91, BS: 32
$\mathcal{M}_3CNN$	0.1907	95.45	88.24	100	93.75	93.10	94.12	0.1176	ϕ_1: 34, ϕ_2: 106, ψ_{fc1}: 230, γ: 3.109×10^{-4}, ϵ: 26, BS: 32
Cascaded	-	**94.19**	**93.84**	**96.00**	**93.00**	**94.10**	**93.80**	**0.0833**	BS: 32
$\mathcal{M}_1CNN$	0.3202	93.02	94.87	91.49	92.49	95.56	93.18	0.0513	ϕ_1: 30, ϕ_2: 81, ψ_{fc1}: 134, γ: 1.406×10^{-4}, ϵ: 45, BS: 64
$\mathcal{M}_2CNN$	0.4620	90.48	94.74	86.96	89.99	86.96	90.85	0.0526	ϕ_1: 56, ϕ_2: 120, ψ_{fc1}: 300, γ: 5.652×10^{-4}, ϵ: 128, BS: 64
$\mathcal{M}_3CNN$	0.4979	90.91	85.71	100	92.31	80.00	92.86	0.1277	ϕ_1: 32, ϕ_2: 120, ψ_{fc1}: 420, γ: 3.352×10^{-4}, ϵ: 42, BS: 64
Cascaded	-	**88.37**	**88.49**	**89.74**	**89.10**	**85.40**	**88.50**	**0.1429**	BS: 64
$\mathcal{M}_1TF$	0.7770	83.72	84.62	82.98	82.50	86.67	83.80	0.1538	L: 1, H: 8, dim: 925, γ: 1.256×10^{-5}, ϵ: 98, BS: 16
$\mathcal{M}_2TF$	0.7456	85.71	81.48	93.33	87.93	86.67	83.80	0.1852	L: 1, H: 8, dim: 330, γ: 4.593×10^{-5}, ϵ: 61, BS: 16
$\mathcal{M}_3TF$	0.2652	93.18	100	87.50	93.02	100	93.75	0.0	L: 4, H: 3, dim: 253, γ: 4.22×10^{-5}, ϵ: 20, BS: 16
Cascaded	-	**83.72**	**83.86**	**83.72**	**85.31**	**83.78**	**83.36**	**0.1277**	BS: 16
$\mathcal{M}_1TF$	0.3222	91.86	94.59	89.80	90.93	95.65	92.19	0.0541	L: 4, H: 3, dim: 253, γ: 5.066×10^{-5}, ϵ: 79, BS: 32
$\mathcal{M}_2TF$	0.3193	88.10	84.21	91.30	86.49	87.50	87.76	0.1579	L: 2, H: 8, dim: 554, γ: 1.73×10^{-4}, ϵ: 12, BS: 32
$\mathcal{M}_3TF$	0.2164	93.18	88.24	96.30	90.90	92.86	92.27	0.1176	L: 6, H: 6, dim: 866, γ: 2.57×10^{-5}, ϵ: 24, BS: 32
Cascaded	-	**77.91**	**77.11**	**77.11**	**73.24**	**80.39**	**77.11**	**0.2778**	BS: 32
$\mathcal{M}_1TF$	0.4469	90.70	89.74	91.49	89.74	91.49	90.61	0.1026	L: 3, H: 3, dim: 664, γ: 2.19×10^{-5}, ϵ: 67, BS: 64
$\mathcal{M}_2TF$	0.5705	90.48	94.74	86.96	89.95	95.24	90.85	0.0526	L: 5, H: 7, dim: 130, γ: 8.83×10^{-5}, ϵ: 83, BS: 64
$\mathcal{M}_3TF$	0.4297	90.91	85.71	100	92.31	80.00	92.86	0.1429	L: 2, H: 4, dim: 318, γ: 1.26×10^{-5}, ϵ: 81, BS: 64
Cascaded	-	**81.40**	**80.60**	**80.60**	**81.46**	**74.47**	**82.11**	**0.2553**	BS: 64

ACC: Accuracy, REC: Recall, and SPEC: Specificity.

Even the Transformer classifier was not lacking in potential, as $\mathcal{M}_3$ achieved a perfect sensitivity of 1 and a test loss of 0.2652. However, overall, the MLP and CNN models performed better. This high degree of accuracy highlights how AI has the potential to improve the efficacy and accuracy of AMD diagnosis, resulting in better treatment options and ultimately better patient outcomes.

C. EXPLAINABILITY AND INTERPRETABILITY

Explainability and interpretability have been applied to the $\mathcal{M}_3$ model with a batch size of 32 to the three different

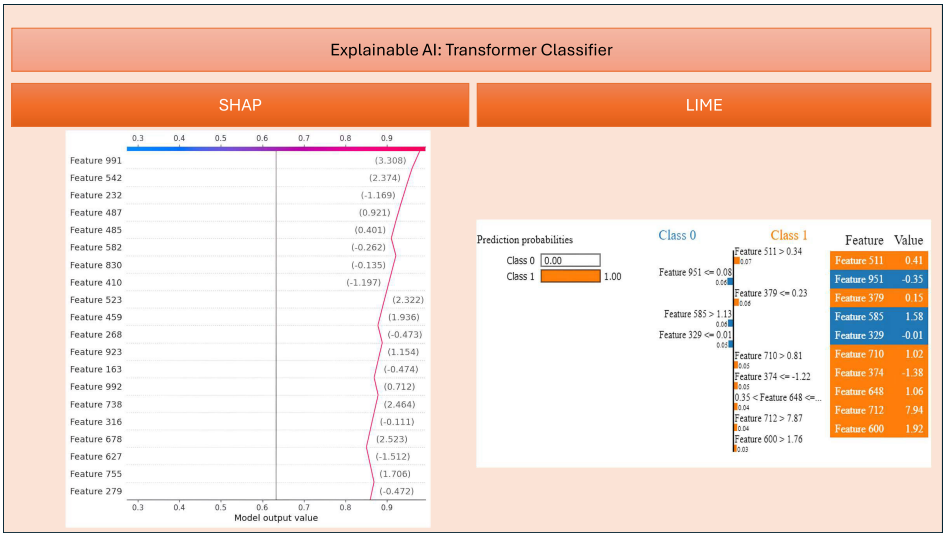


FIGURE 7. Explainability and interpretability of the transformer classifier.

models. SHAP analysis revealed that several features played a pivotal role in shaping the Transformer model, which is shown in Figure 7. The most significant feature was found to be feature number 678 as it continuously increased the output of the model.

On the other hand, feature number 627 had a very detrimental effect. Although not as noticeable as feature number 678, features 542, 523, 738, and 991 also showed notable positive impacts. These findings provide information on the factors affecting the behavior of the model, information that may be used to enhance the model's performance.

The high probability of 1 associated with the GA class in the Lime plot suggests that the Transformer model has a high degree of confidence in its prediction of that class for the specific occurrence. Significant features supporting the prediction include feature number 585, which has a strong positive value, and feature number 951 and 379, where a negative value appears to be significant. Even though they are less significant, features such as 374, 710, and 648 may still have some influence.

As for the CNN model, SHAP analysis showed (Figure 8) that feature number 5 is the most notable positive contributor, resulting in the best results. On the other hand, feature number 177 shows a significant negative influence that frequently results in lower predictions. Moreover, features with different effects, such as features 202 and 242, suggest that the interaction of other features may affect their influence.

The high probability of 1 associated with GA in the case of the LIME explanation indicates that the model has a high degree of confidence in its prediction of GA. Feature 487, feature 712, and potentially feature 696 are important features that support this prediction. They all have positive values that appear to be significant. Features such as 719, 250, 774, 674, 950, 781, and 790 may also affect the prediction to some extent, but are less significant. All in all, the Lime chart indicates that the combination of these particular feature values is what primarily drives the GA model's prediction.

Regarding the MLP model, the SHAP analysis demonstrated that, as shown in Figure 9, Feature 684 is the most significant positive contributor, continuously resulting in greater outputs. Feature 831, on the other hand, continuously lowers expectations, suggesting a negative association. Feature 362 and Feature 800, in particular, show varying impacts, indicating that their impact depends on how other characteristics interact.

The high probability of 1 associated with GA in the case of the LIME explanation indicates that the model has a high degree of confidence in its prediction of GA for the particular instance. Features 774, 695, and 618 are important features that have positive values and seem to have an impact on this prediction. Even though they are less significant, traits like 323, 645, 324, 784, 67, 617, and 329 can also have a small impact on the prediction.

D. RELATED STUDIES COMPARISONS

It is clear that our MLP model performs well, particularly with a batch size of 32 where it achieved an accuracy of 0.9186. The sensitivity for the classes $\mathcal{M}_3$ and $\mathcal{M}_2$ was 0.9444 and 0.9000, respectively, with a BAC of 92.22%. However, when we compared to state-of-the-art methods, such as those by Jiang et al., [19] which reported a classification accuracy of 99.69% with ViT, and Gholami et al., [15] with ViT encoder accuracy of 98.18% and 99.11%, your MLP results are lower.

These studies emphasize the superior performance of transformer-based models in medical imaging tasks, indicating that while our MLP shows strong performance, ViT-based approaches still outperform in terms of accuracy and sensitivity.

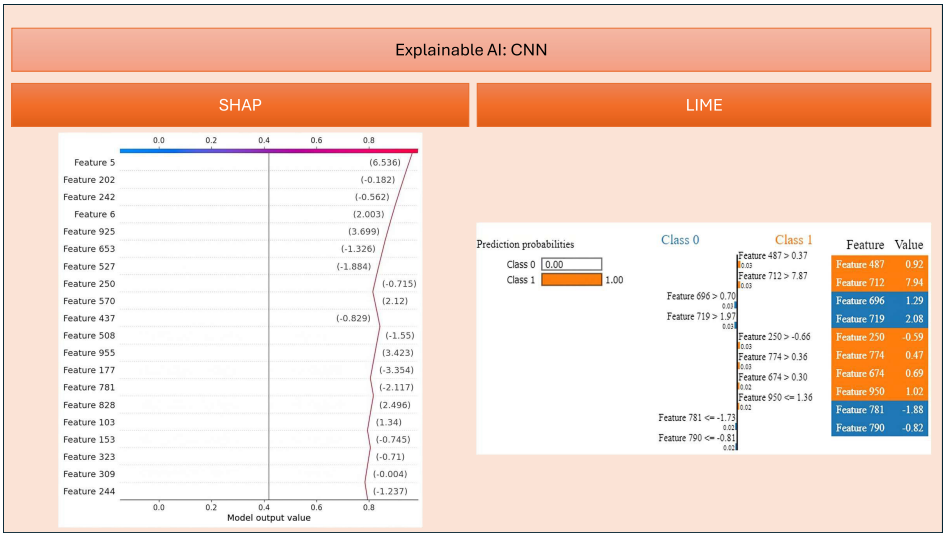


FIGURE 8. Explainability and interpretability of the CNN classifier.

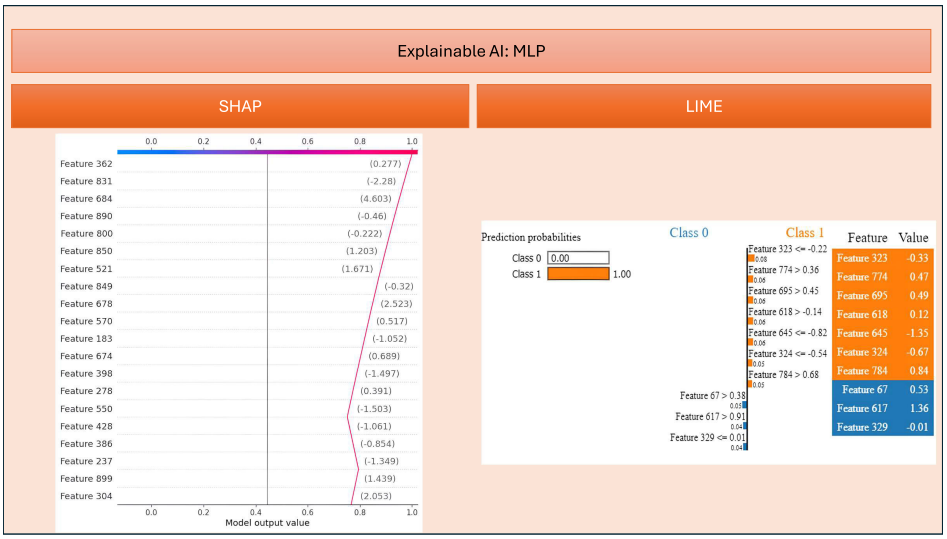


FIGURE 9. Explainability and interpretability of the MLP classifier.

Our CNN results, particularly with a batch size of 32, are quite competitive, achieving an accuracy of 94.19% and a sensitivity of 0.9600 for the $\mathcal{M}_2$ class. This outcome is quite similar to the 95.14% ViT model accuracy reported by Akcca et al. [18]. Even yet, our CNN model is not as accurate as the CNN-RBF model of Hemalakshmi et al., [20] which reached 97.22%. This suggests that transformer-based models, like ViT, may still be superior in extremely specialized datasets even though CNNs are quite effective in our framework, especially when using the cascade strategy.

Out of all our models, the Transformer classifier in our analysis performs the worst. It obtained an accuracy of just 0.7791 with a batch size of 32, and sensitivities of 0.7222 for $\mathcal{M}_3$ and 0.8200 for $\mathcal{M}_2$. In comparison to the transformer-based models published by Gholami et al., Jiang et al., and Akcca et al., this result is noticeably lower. ViT models achieved accuracy levels close to 95% in those studies, with some models reaching 99.69%.

Our results demonstrate that while the MLP and CNN models perform well, particularly with a batch size of 32, the performance gap between these models and ViT-based approaches remains notable. Specifically, transformer-based models such as those by Jiang et al. [19], Gholami et al., [15] and Akcca et al. [18] achieve accuracy levels between 95% and 99.69%, surpassing the sensitivity and accuracy of our CNN and MLP models. This suggests that ViT's feature extraction capabilities, which are utilized in our framework,

TABLE 5. Comparison of the performance of our models with state-of-the-art methods.

Model/Study	Architecture	Accuracy (%)	Sensitivity (Selected Classes)	Balanced Accuracy (BAC)	Interpretability	Computational Efficiency
Our MLP	MLP (Batch Size: 32)	91.86	$\mathcal{M}_3$: 0.9444, $\mathcal{M}_2$: 0.9000	92.22	Moderate (fewer parameters)	High (efficient)
Our CNN	CNN (Batch Size: 32)	94.19	$\mathcal{M}_2$: 0.9600	-	Moderate	High (fast for smaller data)
Our Transformer	Transformer (Batch Size: 32)	77.91	$\mathcal{M}_3$: 0.7222, $\mathcal{M}_2$: 0.8200	-	Low	Moderate
Jiang et al. (2022)[19]	ViT	99.69	Not specified	-	High (self-attention)	Lower (high memory usage)
Gholami et al. (2023) [15]	ViT Encoder	98.18, 99.11	Not specified	-	High	Lower
Akcca et al. (2024)[18]	ViT	95.14	Not specified	-	High	Lower
Hemalakshmi et al. [20]	CNN-RBF	97.22	Not specified	-	Moderate	High

still provide substantial improvements when applied to AMD detection.

The ViT-based feature extraction step plays a crucial role in identifying fine-grained patterns in the data, particularly for subtle abnormalities associated with AMD. While our CNN and MLP models perform well in the cascading framework, integrating ViT's features could be further optimized to improve their performance, especially in detecting challenging cases.

In terms of sensitivity, the ViT models show superior ability to identify true positives, which is crucial for early detection in medical applications. The interpretability of ViT also provides transparency in understanding the regions of the image that contribute to the model's decisions, enhancing trust in clinical settings.

In conclusion, our CNN and MLP models perform well, particularly when batch size is 32, but they are not as good as the most recent ViT models without numerical features. Studies have shown that transformer-based methods have higher sensitivity and accuracy, thus including such models in your framework may improve performance even more.

VI. LIMITATIONS

The limited amount and diversity data of the dataset selected to train and test the models may have limited the work and results. In addition, much work remains to be done on the interpretability and explainability of the models, which is critical for clinical decision making. Gaining trust and ensuring ethical use requires an understanding of model thinking.

Technically speaking, the bitwise AND operation and cropping steps rely on bounding box calculation and image mask correction. These procedures are prone to errors that provide unreliable results. Resizing the images can also cause distortions that degrade the quality of its features. Another area of concern is hyperparameter tuning. BThere is some possibility that the Bayesian optimization has not fully explored the parameter space and that the number of tests or the preset ranges may result in a less than ideal fit.

VII. CONCLUSION AND FUTURE DIRECTIONS

AMD is a degenerative eye disease affecting the macula and primarily affects older adults. AMD arises from age-related damage to the macula, which houses a high density of photoreceptors essential for sharp central vision. Drusen, or a buildup of debris, between the retinal pigment epithelium and Bruch's membrane, are an early sign of AMD. The research provides compelling evidence to address two research questions. The first is to how AI/ML algorithms can improve early diagnosis and staging of AMD compared to conventional techniques. In particular, the MLP model $\mathcal{M}_3$ achieved an impressive accuracy of 0.9545 and sensitivity of 0.95 with a batch size of 16. This improvement demonstrates the significant role of AI and ML algorithms in increasing the accuracy of early detection and staging of AMD. The second research question is about how AI-assisted retinal image interpretation impacts the efficacy and accuracy of AMD diagnosis and treatment options. The study highlights the substantial influence of AI-assisted models on diagnostic accuracy. The MLP model was shown to be able to identify AMD cases with good accuracy, with a BAC of 0.9500 and perfect sensitivity for the $\mathcal{M}_3$ model. While the Transformer classifier also showed potential, achieving a perfect sensitivity of 1 and a test loss of 0.2652, the MLP and CNN models generally delivered superior performance.

A major challenge is early detection of AMD, which can be difficult due to its insidious progression and the limitations of current diagnostic techniques to identify subtle changes in the retina. To address this, our work employed Bitwise AND Operation and Image Cropping techniques with AI algorithms, which improved the identification of these alterations and facilitated intervention. Another challenge is the need for personalized care due to the variability of AMD. Our framework incorporated AI techniques for both image classification and preprocessing to improve the accuracy of AMD classification and provide personalized treatment recommendations. The challenge of case interpretation poses difficulties due to the subjective nature of the data, considering the metrics. Our study utilized AI models (e.g., ViT, CNN, and MLP) to mitigate this subjectivity and added the explainability studies of the different models.

Future research is encouraged to examine whether these models can be applied to diverse clinical settings and demographic groups, as well as whether remote monitoring systems can be implemented to promote treatment adherence. Moreover, to provide a more comprehensive context for the diagnosis and classification of AMD, the integration of multimodal data (e.g., clinical, demographic, and genetic information) together with retinal images will be investigated. The creation of AI-based remote monitoring systems for continuous patient follow-up is another important area of focus. These systems allow early changes in patients'

condition to be detected and improve medication adherence, which in turn optimizes disease management. Furthermore, involving ophthalmologists in a feedback loop will be positive to improve AI models with the help of their clinical expertise. By incorporating their experience into the feedback, the model can be iteratively fine-tuned to improve both performance and explainability, better aligning with real-world diagnostic needs and facilitating broader uptake into clinical practice.

DECLARATIONS

- **CONFLICT OF INTEREST**: No conflict of interest exists.
- **INTELLECTUAL PROPERTY**: We confirm that, we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, concerning intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.
- **AUTHORSHIP**: We confirm that the manuscript has been read and approved by all named authors. We confirm that the order of authors listed in the manuscript has been approved by all named authors.
- **DATA AVAILABILITY**: Data information is provided in the Materials Section.
- **COMPLIANCE WITH ETHICAL STANDARDS**: The current study does not contain any studies with human participants and/or animals performed by any of the authors.
- **CONSENT TO PARTICIPATE**: There is no informed consent for the current study.
- **CONSENT FOR PUBLICATION**: Not applicable.

REFERENCES

[1] M. Chaudhuri, Y. Hassan, P. P. S. B. Vemana, M. S. B. Pattanashetty, Z. U. Abdin, and H. F. Siddiqui, "Age-related macular degeneration: An exponentially emerging imminent threat of visual impairment and irreversible blindness," *Cureus*, vol. 15, no. 5, May 2023, Art. no. e39624.

[2] A. D. Abdin, M. Devenijn, R. Fulga, A. Langenbucher, B. Seitz, and H. Kaymak, "Prevalence of geographic atrophy in advanced age-related macular degeneration (AMD) in daily practice," *J. Clin. Med.*, vol. 12, no. 14, p. 4862, Jul. 2023.

[3] J. Ling, H. Zhang, D. Glujovsky, Z.-L. Xie, Y. Wang, G. Esoimeme, and X.-X. Luo, "Risk factors and global prevalence of neovascular age-related macular degeneration: Protocol for a systematic review and meta-analysis," *Res. Square*, Mar. 2024, doi: 10.21203/rs.3.rs-4079413/v1.

[4] D. B. Rein, J. S. Wittenborn, Z. Burke-Conte, R. Gulia, T. Robalik, J. R. Ehrlich, E. A. Lundeen, and A. D. Flaxman, "Prevalence of age-related macular degeneration in the U.S. in 2019," *JAMA Ophthalmol.*, vol. 140, no. 12, p. 1202, Nov. 2022.

[5] M. Fleckenstein, S. Schmitz-Valckenberg, and U. Chakravarthy, "Age-related macular degeneration: A review," *JAMA*, vol. 331, no. 2, pp. 147–157, 2024.

[6] M. Sahoo, M. Mitra, and S. Pal, "Improved detection of dry age-related macular degeneration from optical coherence tomography images using adaptive window based feature extraction and weighted ensemble based classification approach," *Photodiagnosis Photodynamic Therapy*, vol. 42, Jun. 2023, Art. no. 103629.

[7] R. D. Jager, W. F. Mieler, and J. W. Miller, "Age-related macular degeneration," *New England J. Med.*, vol. 358, no. 24, pp. 2606–2617, 2008.

[8] L. M. Khachigian, G. Liew, K. Y. C. Teo, T. Y. Wong, and P. Mitchell, "Emerging therapeutic strategies for unmet need in neovascular age-related macular degeneration," *J. Transl. Med.*, vol. 21, no. 1, p. 133, Feb. 2023.

[9] J. Blasiak, E. Pawlowska, J. Ciupińska, M. Derwich, J. Szczepanska, and K. Kaarniranta, "A new generation of gene therapies as the future of wet AMD treatment," *Int. J. Mol. Sci.*, vol. 25, no. 4, p. 2386, Feb. 2024.

[10] N. M. Schultz, S. Bhardwaj, C. Barclay, L. Gaspar, and J. Schwartz, "Global burden of dry age-related macular degeneration: A targeted literature review," *Clin. Therapeutics*, vol. 43, no. 10, pp. 1792–1818, Oct. 2021.

[11] G. A. Saleh, N. M. Batouty, S. Haggag, A. Elnakib, F. Khalifa, F. Taher, M. A. Mohamed, R. Farag, H. Sandhu, A. Sewelam, and A. El-Baz, "The role of medical image modalities and AI in the early detection, diagnosis and grading of retinal diseases: A survey," *Bioengineering*, vol. 9, no. 8, p. 366, Aug. 2022.

[12] M. Fabre, L. Mateo, D. Lamaa, S. Baillif, G. Pagès, L. Demange, C. Ronco, and R. Benhida, "Recent advances in age-related macular degeneration therapies," *Molecules*, vol. 27, no. 16, p. 5089, Aug. 2022.

[13] L. Hamid, A. Elnokrashy, E. H. Abdelhay, and M. M. Abdelsalam, "A deep learning LSTM-based approach for AMD classification using OCT images," *Neural Comput. Appl.*, vol. 36, no. 31, pp. 19531–19547, Nov. 2024.

[14] P. K. Karn and W. H. Abdulla, "On machine learning in clinical interpretation of retinal diseases using OCT images," *Bioengineering*, vol. 10, no. 4, p. 407, Mar. 2023.

[15] S. Gholami, J. I. Lim, T. Leng, S. S. Y. Ong, A. C. Thompson, and M. N. Alam, "Federated learning for diagnosis of age-related macular degeneration," *Frontiers Med.*, vol. 10, 2023, Art. no. 1259017.

[16] Y. Kihara, M. Shen, Y. Shi, X. Jiang, L. Wang, R. Laiginhas, C. Lyu, J. Yang, J. Liu, R. Morin, R. Lu, H. Fujiyoshi, W. J. Feuer, G. Gregori, P. J. Rosenfeld, and A. Y. Lee, "Detection of nonexudative macular neovascularization on structural OCT images using vision transformers," *Ophthalmol. Sci.*, vol. 2, no. 4, Dec. 2022, Art. no. 100197.

[17] K. Xu, S. Huang, Z. Yang, Y. Zhang, Y. Fang, G. Zheng, B. Lin, M. Zhou, and J. Sun, "Automatic detection and differential diagnosis of age-related macular degeneration from color fundus photographs using deep learning with hierarchical vision transformer," *Comput. Biol. Med.*, vol. 167, Dec. 2023, Art. no. 107616.

[18] S. Akça, Z. Garip, E. Ekinci, and F. Atban, "Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal OCT images using vision transformers: A comparative study," *Lasers Med. Sci.*, vol. 39, no. 1, p. 140, May 2024.

[19] Z. Jiang, L. Wang, Q. Wu, Y. Shao, M. Shen, W. Jiang, and C. Dai, "Computer-aided diagnosis of retinopathy based on vision transformer," *J. Innov. Opt. Health Sci.*, vol. 15, no. 2, Mar. 2022, Art. no. 2250009.

[20] G. R. Hemalakshmi, D. Santhi, V. R. S. Mani, A. Geetha, and N. B. Prakash, "Classification of retinal fundus image using MS-DRLBP features and CNN-RBF classifier," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 9, pp. 8747–8762, Sep. 2021.

[21] Z. Yao, Y. Yuan, Z. Shi, W. Mao, G. Zhu, G. Zhang, and Z. Wang, "FunSwin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images," *Frontiers Physiol.*, vol. 13, Jul. 2022, Art. no. 961386.

[22] F. D. Marco, L. D. Biasi, A. A. Citarella, and G. Tortora, "Improving PVC detection in ECG signals: A recurrent neural network approach," in *Proc. Italian Workshop Artif. Life Evol. Comput.* Cham, Switzerland: Springer, Jan. 2024, pp. 256–267.

[23] F. De Marco, F. Ferrucci, M. Risi, and G. Tortora, "Classification of QRS complexes to detect premature ventricular contraction using machine learning techniques," *PLoS ONE*, vol. 17, no. 8, Aug. 2022, Art. no. e0268555.

[24] A. A. A. El-Khalek, H. M. Balaha, A. Mahmoud, N. S. Alghamdi, M. Ghazal, A. T. Khalil, M. E. A. Abo-Elsoud, and A. El-Baz, "A novel machine learning-based classification framework for age-related macular degeneration (AMD) diagnosis from fundus images," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, May 2024, pp. 1–4.

[25] A. A. Abd El-Khalek, H. M. Balaha, N. S. Alghamdi, M. Ghazal, A. T. Khalil, M. E. A. Abo-Elsoud, and A. El-Baz, "A concentrated machine learning-based classification system for age-related macular degeneration (AMD) diagnosis using fundus images," *Sci. Rep.*, vol. 14, no. 1, p. 2434, Jan. 2024.

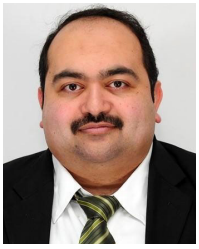
[26] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—A contemplative retrospection," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106126.
[27] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 103000.
[28] Y. Shisu, S. Mingwin, Y. Wanwag, Z. Chenso, and S. Huing, "Improved EATFormer: A vision transformer for medical image classification," 2024, *arXiv:2403.13167*.
[29] A. A. Abd El-Khalek, H. M. Balaha, A. Sewelam, M. Ghazal, A. T. Khalil, M. E. A. Abo-Elsoud, and A. El-Baz, "A comprehensive review of AI diagnosis strategies for age-related macular degeneration (AMD)," *Bioengineering*, vol. 11, no. 7, p. 711, Jul. 2024.
[30] N. Kushwah, K. Bora, M. Maurya, M. C. Pavlovich, and J. Chen, "Oxidative stress and antioxidants in age-related macular degeneration," *Antioxidants*, vol. 12, no. 7, p. 1379, Jul. 2023.
[31] *CATT | Center for Preventive Ophthalmology and Biostatistics (CPOB) | Perelman School of Medicine at the University of Pennsylvania*. Accessed: Jun. 18, 2024. [Online]. Available: https://www.med.upenn.edu/cpob/catt.html
[32] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.
[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[34] I. de Zarzà, J. de Curtò, E. Hernández-Orallo, and C. T. Calafate, "Cascading and ensemble techniques in deep learning," *Electronics*, vol. 12, no. 15, p. 3354, Aug. 2023.
[35] G. Lee and J. Kim, "Parallel vs. sequential cascading MEP coordination strategies: A pharmaceutical building case study," *Autom. Construct.*, vol. 43, pp. 170–179, Jul. 2014.
[36] T. Bai, Y. Li, Y. Shen, X. Zhang, W. Zhang, and B. Cui, "Transfer learning for Bayesian optimization: A survey," 2023, *arXiv:2302.05927*.
[37] M. S. Bandstra, J. C. Curtis, J. M. Ghawaly, A. C. Jones, and T. H. Y. Joshi, "Explaining machine-learning models for gamma-ray detection and identification," *PLoS ONE*, vol. 18, no. 6, Jun. 2023, Art. no. e0286829.
[38] A. Sathyan, A. I. Weinberg, and K. Cohen, "Interpretable AI for bio-medical applications," *Complex Eng. Syst.*, vol. 2, no. 4, p. 18, 2022.
[39] H. Hakkoum, A. Idri, and I. Abnane, "Global and local interpretability techniques of supervised machine learning black box models for numerical medical data," *Eng. Appl. Artif. Intell.*, vol. 131, May 2024, Art. no. 107829.
[40] L. Di Biasi, F. De Marco, A. Auriemma Citarella, M. Castrillón-Santana, P. Barra, and G. Tortora, "Refactoring and performance analysis of the main CNN architectures: Using false negative rate minimization to solve the clinical images melanoma detection problem," *BMC Bioinf.*, vol. 24, no. 1, p. 386, Oct. 2023.
[41] *NVIDIA Quadro M4000 Features and Specifications*. Accessed: Jun. 18, 2024. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/documents/75509_DS_NV_Quadro_M4000_US_NV_HR.pdf
[42] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA, USA: CreateSpace, 2009.
[43] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, Jan. 2019, pp. 8024–8035.
[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. J. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2012.
[45] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.
[46] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.
[47] Z. Zhang, "Improved Adam optimizer for deep neural networks," in *Proc. IEEE/ACM 26th Int. Symp. Quality Service (IWQoS)*, Jun. 2018, pp. 1–2.

AINHOA OSA-SANCHEZ received the degree in industrial electronics and automation engineering from the University of Deusto, in 2021, and the master's degree in industry 4.0 from the International University of La Rioja, in 2022. Currently, she is pursuing the Ph.D. degree with the University of Deusto and remains part of the eVida research group. During her bachelor's and master's studies, she undertook internships with the eVida research group with the University of Deusto. Her master's thesis focused on telemonitoring vital signs in nursing homes using IoT and 3D design, employing a serverless architecture with Amazon Web Services for data storage and visualization. Her Ph.D. thesis focuses on using EEG and NIR wearable sensor signals for pain detection, applying artificial intelligence models. Since June 2024, she has been engaged in an International Research Stay with the Department of Bioengineering, University of Louisville to continue the International Capstone project titled "Age-related macular degeneration."

HOSSAM MAGDY BALAHA was born in Egypt, in 1993. He received the B.Sc. and M.Sc. degrees from the Computers Engineering and Systems Department, Mansoura University. He is currently pursuing the Ph.D. degree. He is a Graduate Teaching Assistant (GTA) with the Bioengineering Department, Speed School of Engineering, University of Louisville (UofL), USA. He also serves as an Assistant Lecturer with the Computers Engineering and Systems Department, Faculty of Engineering, Mansoura University, Egypt. In addition to his academic roles, he is a Senior Full Stack Laravel Web Developer, an AI Engineer, and a Freelancer. He has created a range of educational courses on YouTube, covering topics, such as artificial intelligence (AI) techniques in digital histopathology and radiomics, web development, and the Internet of Things (IoT). His work includes the development of various web and mobile applications. He has actively contributed to numerous projects and competitions aligned with his fields of interest. He has published several research papers, presented at conferences, and contributed to book chapters on these topics. His primary research interests include medical image analysis, AI, deep learning (DL), machine learning (ML), computer vision (CV), radiomics, soft computing, web development, and system development. He has been recognized by the University of Louisville with several awards, including the Outstanding International Student Award. Additionally, he serves as a Reviewer for several prestigious journals, including IEEE Access, IEEE Transactions on Cybernetics, *Journal of Ambient Intelligence and Humanized Computing* (JAIHC), and *Multimedia Tools and Applications* (MTAP).

ALI MAHMOUD received the B.S. and M.S. degrees in electrical engineering from Alexandria University, Egypt, in 2005 and 2009, respectively, and the Ph.D. degree in electrical engineering from the University of Louisville, Louisville, KY, USA, in 2014. He is currently a Researcher with the Bioengineering Department, University of Louisville. His research interests include computer vision and image processing, object detection and tracking, medical imaging, and facial biometrics.

ASHRAF SEWELAM, photograph and biography not available at the time of publication.



MOHAMMED GHAZAL (Senior Member, IEEE) received the B.Sc. degree in computer engineering from the American University of Sharjah, and the M.Sc. and Ph.D. degrees in electrical and computer engineering (ECE) from Concordia University, Canada, in 2006 and 2010, respectively. He is currently a Professor and the Chairperson of the Department of Electrical, Computer, and Biomedical Engineering, College of Engineering, Abu Dhabi University. He is the author of more than 200 publications in journals and international conferences, including *Scientific Reports*, Elsevier's *Renewable and Sustainable Energy Reviews*, IEEE Transactions on Biomedical Engineering, IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Consumer Electronics. His current research interests include AI in bioimaging, genomics, and smart systems. He is a member of ACM, BMES, and an ABET program evaluator for the Engineering Accreditation Commission. He has received several awards, including Dhs1,000,000 from the UAE Prime Minister's Office for an innovative mobile app during the 2015 Government Summit, the distinguished Faculty Award from Abu Dhabi University, in 2014 and 2016, and Abu Dhabi University Research Fellow, in 2018. He was awarded an Alexander Graham Bell Fellowship from NSERC and QFRNT from the Government of Quebec. He received the President's Cup and the Ministry of Education's Shield for Creative Thinking. He was also inducted into the American University of Sharjah Hall of Fame by Sheikh Sultan Al Qasimi, in 2018.

BEGONYA GARCIA-ZAPIRAIN was born in San Sebastian, Spain. She received the degree in communications engineering, specializing in telematics from Basque Country University, Bilbao, Spain, in 1994. In 2004, she defended the Ph.D. thesis in the biomedical signal processing research area. After four years of working for ZIV Company, in 1997, she joined the University of Deusto as a Lecturer in signal theory and electronics with the Engineering School. She was the Head of the Telecommunication Department, University of Deusto, from 2002 to 2008, and also the Head of the DeustoTech-LIFE Department, from 2008 to 2018. In 2001, she created the eVIDA Research Group (evida.deusto.es). She has participated in more than 100 research projects on international, national, and regional levels, published more than 70 articles in international scientific ISI indexed journals, and presented more than 160 papers in international and national scientific conferences. She has been supervising more than ten theses and has ongoing five Ph.D. students. She has strong collaborations with research laboratories in USA, France, Ireland, and The Netherlands. She has five research awards at national level.

AYMAN EL-BAZ (Fellow, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering, in 1997 and 2001, respectively, and the Ph.D. degree in electrical engineering from the University of Louisville, in 2006. He is currently a Professor with University Scholar, and the Chair of the Department of Bioengineering, University of Louisville. He is also a Distinguished Professor with the University of Louisville at Alamein International University (UofL-AIU), New Alamein City, Egypt. He was named as a fellow of Coulter, AIMBE, and NAI for his contributions to the field of artificial intelligence in medicine and biomedical translational research. He has two decades of hands-on experience in the fields of bio-imaging modeling and non-invasive computer-assisted diagnosis systems. He has authored or co-authored more than 700 technical articles (168 journals, 44 books, 85 book chapters, 255 refereed conference papers, 196 abstracts, and 36 U.S. patents).