DOI: 10.65731/ama/2026-0059

# AN EXPLAINABLE AI ENSEMBLE MODEL FOR ACQUIRED VITELLIFORM LESION IDENTIFICATION

**Pawel POWROŹNIK[*], Maria SKUBLEWSKA-PASZKOWSKA[*], Katarzyna NOWOMIEJSKA[**], Robert REJDAK[**], Marcin DERLATKA[***]**

[*]Faculty of Electrical Engineering and Computer Science, Department of Computer Science, Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland
[**]Faculty of Medicine, Chair and Department of General and Pediatric Ophthalmology, Medical University of Lublin, Chmielna 1, 20-079, Lublin, Poland
[***]Institute of Biomedical Engineering, Faculty of Mechanical Engineering, Bialystok University of Technology, Bialystok, Poland,

p.powroznik@pollub.pl, maria.paszkowska@pollub.pl, katarzyna.nowomiejska@umlub.pl, robertrejdak@yahoo.com, m.derlatka@pb.edu.pl

*received 07 April 2026, revised 30 May 2026, accepted 13 June 2026*

**Abstract:** Retinal diseases gradually weaken eyesight and may even potentially lead to blindness. Recognizing changes in the retina based on OCT imaging allows for the detection of diseases at their early stages and thus for making an appropriate diagnosis. In this study, acquired vitelliform lesions (AVL), drusen, and healthy cases are identified utilizing various CNN-based architectures, such as the Convolutional Gated Recurrent Units U-Net, Deep CNN-GRU Network, Residual Attention CNN. The dataset consisting of OCT images was created using photographs gathered from two research centres and publicly available OCT database. The single models obtained to be very effective in recognition the retinal diseases, obtaining the accuracy of 93.80%, 94.03%, and 95.18% for the Convolutional Gated Recurrent Units U-Net, Deep CNN-GRU Network, Residual Attention CNN, respectively. These models outperformed the pre-trained deep learning architectures, VGG-16, ResNet-18, InceptionV3, and DenseNet-121. In order to further enhance the performance of AVL, drusen, and normal cases identification up to 97.09% accuracy, bagging, boosting, and stacking of ensemble learning methods for all models are applied. Moreover, in order to eliminate the black box effect and indicate on what basis the classifier makes conclusions, two interpretability techniques are employed: Gradient Weighted Class Activation Maps (Grad-CAM) and Shap values. This study gives the in-depth insight into AVL, Drusen, and normal cases identification. Moreover, it provides the most effective CNN-based architecture with high accuracy to support ophthalmologists.

**Key words:** retinal diseases, Acquired Vitelliform Lesions, DRUSEN, Convolutional Gated Recurrent Units U-Net, Deep CNN-GRU Network, Residual Attention CNN, ensemble learning, XAI

## 1. INTRODUCTION

Abnormalities that occur in retina may have a severe impact on human vision and even may lead to blindness. These changes in retina structure are allied with various pathologies, such as diabetic macular edema (DME), choroidal neovascularization (CNV), age-related macular degeneration (AMD), or acquired vitelliform lesions (AVL) [1–7].

Age-related macular degeneration is a severe retinal disease that affects central part of vision. It usually appears in the elderly due to a combination of genetic and behavioural factors. It is estimated that until 2040 AMD will expand to 288 million cases [8–10]. Small yellow deposits beneath the retina, known as drusen, are hallmark features of dry this distortion [11].

Acquired vitelliform lesions encompass a broad spectrum of retinal diseases [6,12]. In addition to hereditary Best disease, which occurs in younger patients, there is AVL, which often occurs in older individuals, such as AMD [13]. The vitelliform lesion represents an accumulation of lipofuscin, melanosomes, melanolipofuscin, and outer segment debris in the subretinal space, between the photoreceptors layer and retinal pigment epithelium [14].The AVL and drusen may be easily confused that is why accurate method of their identification is of great importance.

Optical Coherence Tomography (OCT) is a groundbreaking and non-invasive imaging method often applied to monitor retinal distortions and detect all abnormalities [4]. This technique supports ophthalmologists by providing very accurate and high-resolution of retinal structure.

Convolutional Neural Network (CNN) architectures play a pivotal role in retinal diseases imaging identification [6,15–18]. The convolutional operations provide feature extraction and thus leverage depth-wise and spatial information for enhancing prediction process. Many studies have proposed custom-based models, creating from scratch, to obtain high effectiveness for various retinal pathologies. A great number of papers concern on pre-trained, deep learning state-of-the-art architectures, where the final layer is training to tune the model for specific tasks identification. Due to the fact that each model leverages the unique features, various types of ensemble learning are widely employed [14,19]. They main goal

is to improve the overall accuracy of the final model. Moreover, this approach possesses a great number of advantages, like reducing overfitting, reducing their dependency on noisy data.

To eliminate the black box effect and indicate on what basis the classifier makes conclusions, various interpretability techniques are employed [20]. The Gradient Weighted Class Activation Maps (Grad-CAMs), Testing with Concept Activation Vectors (TCAVs), and Shapley values determine the human-interpretable context for imaging with great success [21–23]. These visualizations create a simple interpretation how model works.

Due to the fact that the majority of retinal disorders are identified in their advanced stages, there is a low probability of recovery because of the lack of adequate healthcare infrastructure [3]. Moreover, there are no sufficient professionals in the rural areas. Another aspect that should be emphasized is that retinal diseases are usually diagnosed employing manual procedures that extremely rely on clinicians' expertise. Thus, the main motivation of this study focuses on proposing a new, automated OCT-image based model that supports ophthalmologists with high reliability. This solution will provide significant assistant in quickly diagnosis of AVL and Drusen disorders. The architecture should also include healthy cases and distinguish them from the pathological images.

The main aim of this study to develop an accurate automatic tool to identify AVL and Drusen distortions, and to distinguish them from healthy cases. For this purpose, the CNN-based models are proposed. Moreover, three ensemble learning methods, bagging, boosting, and stacking, are applied to create the most effective architecture.

The main contributions of this study are as follows:

- Gathering the dataset providing OCT imaging of AVL, Drusen, and healthy cases. The dataset is obtained from two research centres and Kaggle OCT dataset;
- Employing the Deep CNN-GRU Network [24], Convolutional Gated Recurrent Units U-Net [18], and Residual Attention to AVL, Drusen, and healthy cases identification. Their performance is verified using accuracy, precision, recall, and F1-score metrics. These CNN-based architectures obtained hight reliability with average accuracy of 90.50%, 89.59%, and 92.12%, respectively. The precision for AVL class extends 60%, which outperforms the state-of-the-art models such as VGG-16, ResNet-18, InceptionV3, and DenseNet-121;
- Applying the Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the most relevant image areas based on which the above-mentioned CNN models provide their decision-making;
- Employing bagging, boosting, and stacking ensemble learning techniques to provide the most effective model for AVL, Drusen, and healthy cases recognition. The stacking method states to be the most effective, achieving 97.93% average accuracy. It should be emphasized that ensemble learning models provide precision at a level exceeding 94.44%;
- Comparing the proposed models with the state-of-the-art pretrained architectures to evaluate which architecture is the most convenient for the gathered dataset. The Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention as well as the models obtain with bagging, boosting, and stacking ensemble learning outperform the VGG-16, ResNet-18, InceptionV3, and DenseNet-121.

This study gives an in-depth insight into AVL, Drusen, and normal cases identification.

## 2. RELATED WORKS

CNN architectures have gained a lot of interest recently for retinal eyes diseases detection. The well-known, pre-trained, deep learning models are often applied to identify various retina pathologies. There are a group of studies presenting new architectures. A branch of the field concerns the models combining various architectures to enhance classifiers performance.

A lightweight CNN model was proposed for distinguishing heathy cases from these with various pathology, such as: cataract, diabetic retinopathy, and others [25]. The study utilized two datasets: the Ocular Disease Intelligence Recognition and Retinal Fundus Multi-Disease Image Dataset. For feature extractions different types of family wavelet were selected. The developed architectures provided high performance, reaching 96.21% accuracy using 3-level decomposition biorthogonal wavelet. Another lightweight CNN-based model, namely RetNet, for accurate CNV, DME, Drusen, and normal cases with Retinal Diseases Classification photos dataset was proposed in [15]. The architecture consists of 22 layers, including six 2D convolutional, and three dense ones. It achieved 95.41% accuracy, outperforming the state-of-the-art models, such as ResNet-50, InceptionV3, EfficientNet B0, Xception, and VGG-16.

Diabetic retinopathy pathologies, such as CNV, DME, Drusen, together with healthy cases were recognized using the model utilizing complex of 1-Dimensional convolutional neural networks [1]. The study was performed using Mendeley database containing OCT images. All data were pre-processed with constrained normalized subband adaptive filter. The proposed architecture provided to be extremely effective, reaching up to 98.12% accuracy. This model outperformed convolutional and deep convolutional neural networks. CNV, DME, Drusen, and normal cases were identified with the proposed lightweight architectures utilizing deep learning approach, a judicious fusion of convolutional and identity blocks [2]. The study was performed using the pre-processed and augmented data gathered from three datasets: OCT2017, Retinal OCT-C8, and Database-101. This model achieved 92.97% accuracy outperforming the pre-trained ResNet50v2. The same retinal disorders from OCT images were recognized utilizing the 5-layer custom CNN architecture achieving 98% accuracy [3]. This model surpassed the networks such as VGG-16, ResNet-50, and AlexNet. The study was performed utilizing Kaggle dataset. Moreover, the authors emphasized that standard machine learning methods like Support Vector Machine and XGBoost are not efficiently enough to obtain high performance of CNV, DME, Drusen, and normal cases detection. TezNet, CNN-based architecture, was proposed for CNV, DME, Drusen, and normal cases from Kaggle OCT dataset [26]. This model utilized residual block to eliminate vanishing gradient issues, and leverage Inception module technique to enhance network width and thus improved fine-detail detection. This solution achieved 96.20% accuracy, outperforming Con-Vit model. It should be emphasized that the TezNet model had half parameter of deep learning architectures. 60-layer CNN-based architecture, namely OCTm, with attention mechanism is another model proposed for CNV, DME, Drusen, and healthy cases identification [7]. This solution achieved 99.79% and 99.69% accuracy for UCSD-v2 and Duke datasets, respectively. The residual self-attention vision transformer, RS-A ViT , for AVL, Drusen, and healthy cases identification using OCT imaging provided to be highly effective, achieving 93.33% accuracy [6]. This model outperformed the standard deep learning architectures, such as EfficientNet, InceptionV3, ResNet-50, and VGG-16.

A majority of studies about diabetic retinopathy were performed utilizing the pre-trained deep learning models [27]. Recognition the cases with and without DME from fundus and OCT images was studied in [28]. The Messidor-2 and Retinal OCT datasets were involved for training and testing the new model based on two pre-trained VGG-19 and ResNet-50 architectures. The macula regions were segmented utilizing DeeplabV3+ approach. This high-performance model achieved 98.79% and 98.81% accuracy results for VGG-19, and ResNet-50, respectively. The VGG-16, VGG-19, ResNet-50, MobileNet, and the proposed LayerV, extending the VGG-16, architectures were employed to DME, Drusen, CNV, and normal cases identification from OCT imaging [29]. The LayerV outperformed all above-mentioned models with respect to precision, recall, and F1-score metrics. In case of these all measures, it achieved values exceeding 90%. Early diagnosis of CNV, DME, Drusen together with normal cases recognition were performed with high success utilizing VGG16, VGG19, ResNet50, and InceptionV3 using OCT imaging [30]. Three experiments, with and without preprocessing and data augmentation, as well as class imbalance handling using feature space SMOTE were done. The results indicated the VGG-based architectures obtained the best competitive performance for automatic identification, extending 97%.

Retinal diseases such as CNV, Drusen, DME, and normal cases were identified with three methods utilizing the OCT Kaggle dataset [31]. The pre-trained ResNet-18 obtained 86.12% average accuracy. Slightly lower performance, 82.45%, was reached using 4-deep layer CNN. An improved dynamic-layered classification, utilizing ResNet model, outperformed these two networks, achieving 99.45%. Three choroid disorders, namely central serous chorioretinopathy, CNV, and polypoid choroidal vasculopathy together with normal cases, and drusen were recognized using the 15-layer CNN with the modified VGG-16 architecture [32]. This model achieved 98.5% accuracy, outperforming other networks, such as ResNet-50 and VGG-16. Moreover, the DME degenerations were segmented utilizing the fine-tuned UNet architecture. That model provided severity analysis of retinal and choroidal diseases. The study was performed utilizing OCT datasets. The custom CNN, VGG-19, and DenseNet-121 were compared in order to identify CNV, DME, Drusen, and normal cases based on OCT imaging with high performance [33]. The CNN architecture achieved 92.59% accuracy for balanced dataset, while the pre-trained models obtained 86.78% and 84.38% accuracy values for VGG-19 and DenseNet-121, respectively. Another study [34] employed the custom OCT-CNN architecture and compared it with popular models, ResNet-50, VGG-19, MobileNet, and InceptionV3. Also in this case, the proposed solution based on the CNN architecture achieved higher classification accuracy for CNV, DME, Drusen, and healthy cases, 99.23%, compared to standard pre-trained architectures. The 3-, 4-, and 5-layer CNN models were employed for the same retinal diseases achieving 53.50%, 78.50%, and 90.00% accuracy, respectively [35]. The comparison with the pre-trained architectures proved that the VGG-16 obtained the highest accuracy of identification, equal to 96.00%, while the ShuffleNet did not reach high performance with the accuracy of 76.12%. In [4] the custom-based CNN, the modified VGG-19, InceptionV3, the standard ResNet-101, and ResNet-101 with Gaussian median wavelet filters were compared for CNV, DME, Drusen, and normal cases OCT identification. The CNN model obtained 99% accuracy, outperforming other architectures. The hybrid multilayered classification for the same above-mentioned retinal distortions employing CNN and VGG-19 models [36]. This architectures combination provided to be extremely effective, achieving 97.76%, 98.12%, and 98.12% accuracy for CNV, DME, and Drusen, respectively.

Trust of AI-based architectures for medical purposes are of great importance. Thus, in order to provide transparent model's decision making SHAP explainable method was applied for DME, CNV, and Drusen identification [37]. The visual explanation of VGG16 architecture helped to improve model reliability. A new explainable approach combining ResNet50 and Vision Transformer layers based on pixel-based explanation was introduced to provide diabetic retinopathy, AMD, and glaucoma diseases [38]. This solution shows clinically significant area. The custom OCT scan-based explainable method was proposed for clinically meaningful diagnostic features for AMD identification [39]. The grading medical landmarks are indicated taking into consideration subretinal and sub-RPE tissue, retinal fluids, and detecting discontinuities of layers.

In order to enhance the performance of the eye diseases identification, the ensemble learning methods have been widely applied. They leverage the strength of individual models and as a result may improve overall recognition effectiveness [40]. DenseNet-201, NaSNetMobile, and ResNet-101v2 were combined utilizing average ensemble learning to improve the CNV, DME, drusen, and normal cases identification by 1.76% accuracy according to the best single architecture [40]. The proposed model obtained 91.51% accuracy. In [5] the new 21-layer CNN-based model, Ret-Detect, was suggested for CNV, DME, Drusen, and normal cases recognition. This architecture outperformed the pre-trained networks, such as: MobileNet, DenseNet-121, VGG-19, InceptionV3, VGG-16, ResNet-50. Moreover, the Ret-Detect obtained better performance than ensemble learning model combining all above-mentioned solutions, reaching 96% accuracy. In [41] two various architectures, ResNetRS and RETFound, based on autoencoder paradigm, were stacked in order to enhance the AMD identification utilizing fundus imaging. This architecture achieved a quadratic weighted kappa of 0.7364 and an accuracy of 66.03%.

The majority of current research largely focuses on diagnosing CNV, AMD, Drusen distortions, and healthy images using OCT imaging. Modern architectures based on convolutional neural networks, deep learning, and the residual-based solutions have enormous potential for diagnosing AVL and AMD.

## 3. MATERIALS AND METHODS

Fig. 1 presents the architecture-level overview of the ensemble model. The figure highlights the complementary character of the three base learners: recurrent spatial-context modelling in Deep CNN-GRU, encoder-decoder recurrent feature fusion in Convolutional GRU U-Net, and residual attention-based feature selection in RAN. Their predictions are then combined using bagging, boosting, and stacking, with XGBoost used as the stacking meta-model, to obtain the final classification of OCT images into AVL, Drusen, and normal classes.

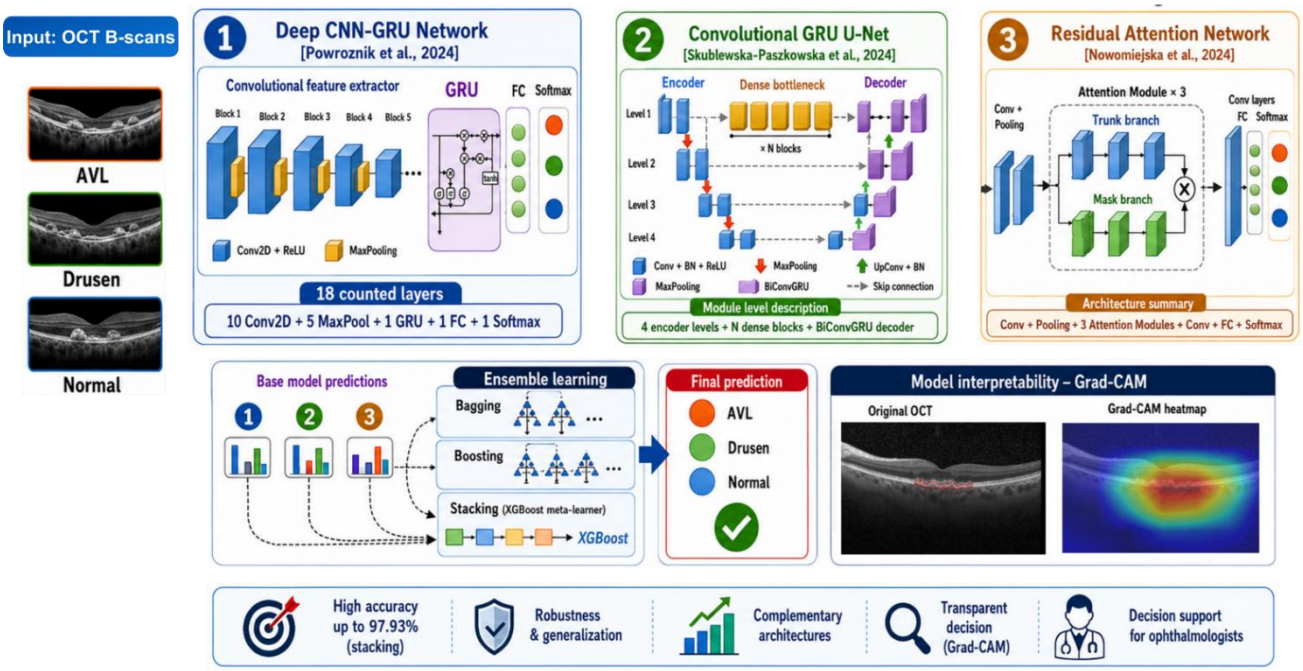



**Fig. 1.** The proposed methodology for ensemble learning model

### 3.1. Dataset

For the purpose of this study, the OCT dataset was gathered from two research centres and Labeled Optical Coherence Tomograph dataset [42]. AVL cases were collected from Eye Clinic, Department of Public Health, Federico II University in Italy and Chair and Department of General and Pediatric Ophthalmology at the Medical University of Lublin in Poland. These data were thoroughly verified. All illegible images were excluded. In total, 125 images were prepared for further study. From the OCT dataset [42] Drusen and normal cases were gathered. As a result, 577 images of Drusen, and 578 normal cases were assembled (Fig. 2).

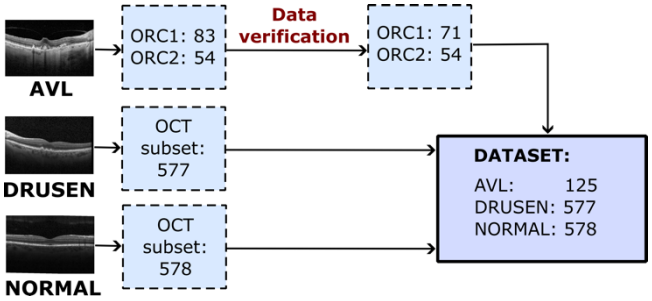



**Fig. 2.** The OCT dataset structure (ORC1 stands for Eye Clinic, Department of Public Health, Federico II University in Italy, ORC2 corresponds to Chair and Department of General and Pediatric Ophthalmology at the Medical University of Lublin in Poland) [6]

The study was carried in accordance with the relevant guidelines and regulations, as well as, the Helsinki Declaration. The scope and method of conducted study were approved by the Ethics Committee of the Medical University of Lublin (no. KE-0254/260/12/2022). An informed consent was obtained from all study participants or their legal guardians. All data available as part of the conducted studies were anonymized to prevent obtaining personal patient data.

The gaussian and median filters were applied to decrease speckle noise and to keep essential structures of all OCT images. Gaussian filtering was used to suppress high-frequency speckle noise by locally smoothing pixel intensities. In contrast, median filtering was used to suppress isolated intensity outliers without blurring diagnostically significant retinal layer boundaries. Both filters were applied with small local kernels before resizing and normalisation.

In case of AVL distortions, the region with the macula was manually selected in order to indicate the relevant area of this disease. Next, all images were resized to a fixed resolution of 224x224 pixels to be appropriate to input layer of CNN architectures. The intensities of each image were scaled to range between 0 and 1 in order to normalize pixel intensities and reduce variability of various imaging sessions. The dataset involved in this study consisted of 125 AVL images, 577 Drusen images and 578 images of normal cases. The whole dataset was divided into training, validation, and testing subsets, with a ratio of 70%, 15%, and 15%, respectively. The proportions of individual classes were maintained (Tab. 1).

**Tab. 1.** Structure of training, validation, and testing subsets

| Subset | Class | No. of data before augmentation | No. of data after augmentation |
|---|---|---|---|
| Training | AVL | 87 | 400 |
| | DRUSEN | 404 | 450 |
| | HEALTHY | 404 | 450 |
| Validation | AVL | 19 | - |
| | DRUSEN | 87 | - |
| | NORMAL | 86 | - |
| Testing | AVL | 19 | - |
| | DRUSEN | 87 | - |
| | NORMAL | 87 | - |

In order to improve the generalization capabilities of machine learning architecture and its performance, the augmentation

process was involved. It was utilized in model training to increase samples diversity and to reduce the risk of overfitting. The following image transformations were involved [43,44]: rotation, scaling, randomly cropping and resizing, colour jittering, gaussian noise, elastic deformation, random erasing, random masking by square regions, horizontal, and vertical flipping.

### 3.2. Deep CNN-GRU

Deep CNN-GRU architecture [17] integrates a deep convolutional backbone with a recurrent module so that both fine-scale structural cues and wider contextual dependencies present in optical-coherence-tomography (OCT) B-scans are captured within a single end-to-end trainable network. The convolutional front-end is organised as five consecutive blocks, each composed of two $(3 \times 3)$ convolutions employing zero padding and unit stride, rectified linear unit (ReLU) activations, and a $(2 \times 2)$ max-pooling operation that halves the spatial resolution. Filter counts grow in a doubling scheme: (64, 128, 256, 512, 512), to enlarge the receptive field while preserving computational tractability. This design mirrors the philosophy of CNN-style architectures, allowing deeper representational hierarchies without excessively large kernels. After the last pooling stage, the feature tensor has spatial dimensions $(7 \times 7)$ with 512 channels. Instead of flattening these maps directly, the tensor is first reshaped into a sequence of 49 feature vectors (one per spatial position), each 512-dimensional, and forwarded to a gated recurrent unit (GRU) layer.

The GRU processes the sequence row-by-row, using its update and reset gates to decide how much of the previously aggregated contextual information to retain and how much to overwrite with the current spatial sample. The network effectively models long-range co-occurrences between distant retinal regions that may be diagnostically linked yet spatially disjoint on a single B-scan. The recurrent block therefore complements the strictly local modelling of the convolutional filters. The final hidden state emitted by the GRU serves as a global representation of the entire image. This vector is fed to a 64-unit fully connected layer that applies dropout regularisation during training to mitigate overfitting. A terminal Softmax layer with four logits transforms the network output into posterior probabilities over the target classes allowing straightforward optimisation via categorical cross-entropy. Information therefore flows through the model as: $[I_{224\times224\times3} \xrightarrow{\text{Conv-Block}_{1\ldots5}} F_{7\times7\times512} \xrightarrow{\text{reshape}} \{f_t\}_{t=1}^{49} \xrightarrow{\text{GRU}} h \xrightarrow{FC_{64}} \text{z} \xrightarrow{\text{Softmax}} \text{p}$ where *I* denotes the input OCT slice, *F* the deepest feature maps, $f_t$ the sequential embeddings, *h* the GRU state, *z* the fully connected activations, and p the predicted class distribution. Each stage is fully differentiable, enabling simultaneous tuning of convolutional kernels, recurrent weights, and classification parameters through back-propagation and stochastic-gradient descent.

### 3.3. Convolutional GRU U-Net

Convolutional GRU U-Net [18] fuses the symmetrical encoder–decoder topology of the original U-Net with bidirectional convolutional gated-recurrent units (BiConvGRU). Thereby enabling the model to exploit both multi-scale spatial context and long-range feature dependencies when analysing ultra-wide-field fundus images.

In the contracting branch, the network ingests a $300 \times 375$ pixel image that has been intensity-normalised to the range $[0,1]$. Four successive resolution levels each apply a pair of $3 \times 3$ convolutions with zero padding and rectified-linear-unit activations, immediately followed by a $2 \times 2$ max-pooling operation that halves the spatial dimensions while doubling the number of feature maps, thus generating a hierarchy of increasingly abstract representations.

At the bottleneck, the plain convolutions of the vanilla U-Net are replaced by a stack of densely connected convolutional blocks in the spirit of DenseNet. Every block receives, as additional channels, the concatenation of all feature maps produced by its predecessors, so that gradient signals propagate unimpeded and redundant filters are suppressed through feature reuse. This design widens the effective receptive field without inflating the parameter count and provides rich semantic tensors for the subsequent decoding path.

The expansive branch begins by up-sampling the encoded tensor through a transposed convolution that doubles the spatial resolution and halves the channel depth. Concurrently, the feature maps cropped from the mirror level of the encoder are resized by an identical up-convolution to ensure perfect alignment.

Rather than concatenating the two streams outright, the architecture interposes a bidirectional convolutional GRU that processes the aligned tensors as a spatio-sequential signal scanned in both raster directions. Inside each GRU cell, the reset and update gates, as well as the candidate activation, are implemented with $3 \times 3$ convolutions so that state transitions preserve the spatial layout. The forward and backward hidden states are finally concatenated, giving a context-enriched feature block that is passed to the next decoding stage.

A batch-normalisation layer follows every up-convolution to stabilise the distribution of activations and accelerate convergence. Decoding proceeds through four such up-sampling BiConvGRU stacks, progressively restoring full image resolution while blending coarse semantic cues from the bottleneck with fine localisation cues from shallow encoder layers.

The last decoder output feeds a $1 \times 1$ convolution that compresses the channel dimension to three logits, one per target class. A spatial softmax converts these logits into pixel-wise posterior maps, and averaging these maps over the field of view yields an image-level class distribution that is optimised end-to-end with categorical cross-entropy.

Overall information flow can be summarised as:

$$[I_{300\times375\times3} \xrightarrow{Enc_{1\ldots4}} E_{H\times W\times C} \xrightarrow{Dense} B \xrightarrow{\bigoplus_{i=1}^{4} UpConv+\text{BiConvGRU}} D_{300\times375\times C'} \xrightarrow{1\times1} Z \xrightarrow{Softmax} \hat{y},$$

where *I* is the input image, *E* the multiscale encoder output, *B* the densely connected bottleneck representation, *D* the reconstructed feature map, *Z* the class-logit tensor, and $\hat{y}$ the predicted class distribution. All operations are differentiable, allowing joint training of convolutional, recurrent, and classification parameters via back-propagation and stochastic-gradient descent.

### 3.4. Residual Attention Convolutional Neural Network

The Residual Attention Network (RAN) proposed in [45] combines deep residual learning with stacked soft–attention blocks so that diagnostically salient structures in OCT-angiography B-scans

of the optic nerve head are selectively amplified while irrelevant background is suppressed. Each Attention Module (AM) contains a trunk branch, implemented as a pre-activation residual unit and a mask branch that follows a bottom-up / top-down scheme to generate a spatially and channel-wise adaptive weight map, which is applied multiplicatively to the trunk features before the residual skip connection is added back.

Input images of $352 \times 352 \times 3$ are first processed by a $3 \times 3$ convolution (stride 1, padding 1) followed by a $2 \times 2$ max-pooling layer, yielding feature maps of size $176 \times 176$. The core of the encoder then unfolds as four resolution stages. Each stage consists of one residual unit that halves the spatial resolution, immediately succeeded by an AM that refines the down-sampled representation. Consequently, the feature tensor dimensions progress as $176 \times 176 \rightarrow 88 \times 88 \rightarrow 44 \times 44 \rightarrow 22 \times 22 \rightarrow 11 \times 11$. Filter depths grow in proportion to preserve capacity, while identity shortcuts inside every residual unit guarantee unobstructed gradient flow.

Within an AM, the mask branch performs successive max-pooling operations to aggregate global context and then upsamples via bilinear interpolation, inserting skip links from matching encoder levels so that fine localisation cues are reinstated. Two $1 \times 1$ convolutions and a sigmoid activation restrict the soft mask $M(z)$ to $[0,1]$. The final output of the module is $H_{i,c} = M_{i,c}\, T_{i,c}$, where T denotes the trunk features and the indices *i,c* enumerate spatial positions and channels, respectively. To prevent attenuation of residual signals when many AMs are stacked, the authors employ an augmented formulation $H_{i,c} = \left(1 + M_{i,c}\right) F_{i,c}$ that preserves identity mapping when the mask goes to zero.

The mask branch itself is a fully convolutional encoder–decoder whose receptive field expands rapidly through repeated pooling, enabling capture of long-range dependencies; symmetric upsampling ensures that the mask aligns pixel-wise with the trunk output. A pair of skip connections between corresponding bottom-up and top-down levels further stabilises training and sharpens the attention contours. Three such AMs, parameterised by $(x, m, n) = (1,2,1)$, denoting one pre-processing residual unit, two trunk residual units and one residual unit between each pooling layer in the mask path, are cascaded, allowing progressive refinement of attention as depth increases.

After the final residual–attention pair at $11 \times 11$, global average pooling collapses the spatial dimensions, producing a feature vector that feeds a fully connected. A softmax transforms outputs into class posterior probabilities and the entire model is optimised end-to-end with categorical cross-entropy.

The complete forward mapping can be formalised as:

$$I_{352\times352\times3} \xrightarrow{\text{Conv } 3\times3+\text{Pool}} F_{176} \xrightarrow{\prod_{j=1}^{4}(\text{Res}_j+\text{AM}_j)} F_{11} \xrightarrow{GAP} h \xrightarrow{FC} z \xrightarrow{Softmax} p,$$

where $F_k$ denotes feature maps of spatial size $k \times k$, *h* the global descriptor generated by *GAP*, *z* the pre-activation logits and *p* the predicted class distribution, GAP stands for global average pooling, *FC* denotes fully connected layers. All convolutional, residual, and attention parameters are updated simultaneously by back-propagation, ensuring that spatial attention masks, residual filters, and classifier weights co-adapt to maximise discriminative power.

### 3.5. Ensemble models

Bagging generates diversity by repeatedly resampling the original training set with replacement, so every bootstrap subset contains a unique mosaic of duplicated and omitted instances. A separate copy of the same base algorithm is fitted on each subset. Because the individual learners are trained on overlapping yet distinct views of the data, their errors tend to be weakly correlated. Final predictions are obtained by simple majority vote, in case of classification. The ensemble therefore drives down variance without inflating bias, yielding a model that is markedly less prone to over-fitting than any single constituent learner. In ophthalmic scenarios a bagged ensemble ensures that no single high-prevalence pattern determines the decision boundary, leading to more stable detection of rare findings.

Boosting takes an almost opposite route: instead of sampling the data, it reweights it. A first "weak" learner is trained on the full dataset; misclassified samples then receive higher weights, forcing the next learner to concentrate on precisely those difficult cases. This sequential reweighting repeats for a user-set number of rounds. The final output is an error-weighted vote of all intermediate models, so that better learners exert greater influence. By systematically focusing attention on outliers and borderline instances, boosting reduces bias while maintaining low variance. When confronted with subtle phenotypic variations, for example, faint AVL dystrophy patterns that can be overlooked by conventional classifiers, boosting iteratively sharpens the ensemble's sensitivity, often correcting images that earlier rounds mislabelled.

Stacking introduces a meta-learning layer on top of a heterogeneous collection of first-level models. Each base learner, often drawn from fundamentally different families, first produces predictions for every instance. Those predictions, typically expressed as class probabilities, are then assembled into a new dataset on which a second-level "meta-learner" is trained. Because the meta-learner observes where individual models agree or diverge, it can discover non-linear relationships among their outputs and learn how to weight them optimally. Unlike bagging or boosting, stacking is free to exploit complementary inductive biases: an attention networks might excel at local texture cues, whereas a GRU based methods could leverage global colour statistics. The meta-learner reconciles both. This flexibility is especially advantageous in multimodal or highly heterogeneous eye-disease datasets (fundus photographs, autofluorescence, OCT), where no single algorithm consistently dominates across all imaging modalities or disease stages.

### 3.6. Explainable

Gradient-weighted Class Activation Mapping is an attribution technique that produces coarse, human-interpretable heatmaps showing which image regions most strongly influence a convolutional neural network's prediction for a given class.

Given a trained model, let $A^k$ denotes the $k^{\text{th}}$ feature map of a selected convolutional layer and $y^c$ the logit associated with class $c$. Grad-CAM computes the class-specific importance coefficient $\alpha_k^c = \frac{1}{Z}\sum_{i,j} \frac{\partial y^c}{\partial A_{ij}^k}$, where $i, j$ stands for index spatial locations and $Z$ is the total number of pixels in the map.

A weighted combination of the feature maps is formed, $L_{\text{Grad-CAM}}^c = ReLU(\sum_k \alpha_k^c\, A^k)$, and then up-sampled to input resolution to yield the final localization map. Because gradients are taken with respect to activations that preserve spatial structure, the heatmap highlights regions that provide positive evidence for the target class, while the ReLU operation suppresses negative contributions [23].

### 3.7. Metrics

The CNN architectures are evaluated using the commonly applied metric assessing the performance of machine learning models, such as: accuracy (1), precision (2), recall (3), and F1-score (4). These measures provide deepen understanding of model's predictive power and its ability to correctly classify samples. In eq. (1)-(4) TP, FP, TN, and FN, correspond to true positive, false positive, true negative, and false negative, respectively.

Accuracy measures the number of correctly predicted samples across all samples (1). Precision assesses the model accuracy of the positive predictions (2). This metric is of great importance in medical diagnosis. The higher the value is, the more correct the model prediction is obtained. Recall provides information about all positive cases, which is extremely important in medical diagnosis where all distortions should be recognized (3). F1-score combines precision and recall to obtain one measure indicating the model's performance (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (1)$$

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (2)$$

$$Recall = \frac{TP}{TP+FN} * 100\% \quad (3)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} * 100\% \quad (4)$$

## 4. RESULTS AND DISCUSSION

### 4.1. Single Convolutional Neural Networks

**Tab. 2.** Accuracy, precision, recall and F1-score results for CNN architectures. All values are in %

| Model | Acc. | Class | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Deep CNN-GRU | 90.50 | AVL | 64.29 | 97.74 | 76.61 |
| | | Drusen | 93.67 | 85.06 | 89.08 |
| | | Healty | 91.86 | 90.80 | 91.32 |
| CGRU U-Net | 89.59 | AVL | 60.71 | 89.47 | 72.43 |
| | | Drusen | 92.94 | 90.80 | 91.83 |
| | | Healty | 96.25 | 88.51 | 92.36 |
| RAN | 92.12 | AVL | 64.29 | 94.74 | 76.54 |
| | | Drusen | 96.47 | 94.25 | 95.38 |
| | | Healty | 95.00 | 87.36 | 91.00 |

The performance for each class for the Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net (CGRU U-Net), and Residual Attention Network (RAN) are gathered in Tab.2. The 5-fold cross validation is utilized in order to enhance the reproducibility and stability of the obtained results. The confuses matrices are presented in Fig. 3-5. The training, validation accuracies and loss functions for three CNN architectures are depicted in Fig. 6-8. These results prove that the proposed architectures have the ability to distinguish normal cases from these affected by Drusen or AVL. The high values of recall, exceeding 85%, and moderate precision results for AVL detection state that the model catches most positive cases, but it also incorrectly labels many negative cases as positive. For Drusen and heatlty images, the models are more effective.

The Deep CNN-GRU Network achieved 90.20% average model accuracy.

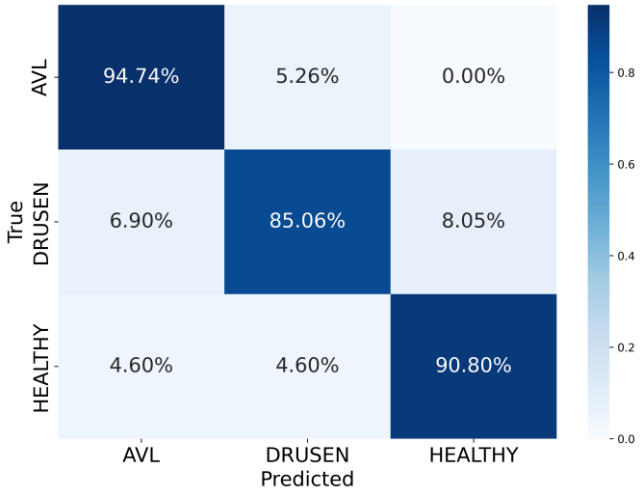



**Fig. 3.** Confusion matrix for the Deep CNN-GRU Network

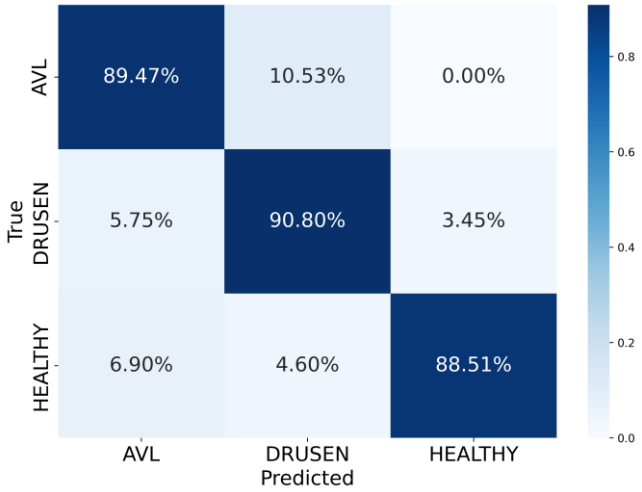



**Fig. 4.** Confusion matrix for the Convolutional Gated recurrent Units U-Net

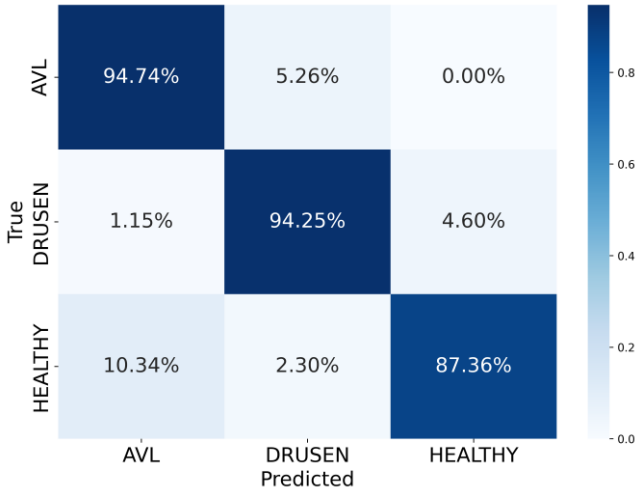



**Fig. 5.** Confusion matrix for the Residual Attention Network

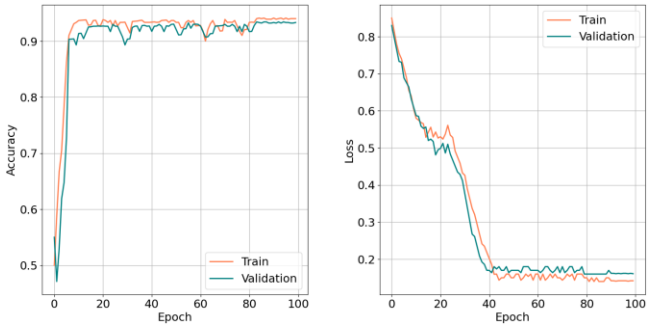



**Fig. 6.** Training and validation accuracies (left) and loss function (right) for the Deep CNN-GRU model

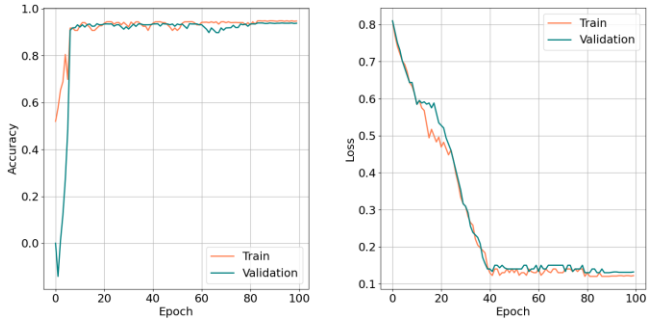



**Fig. 7.** Training and validation accuracies (left) and loss function (right) for the Convolutional Gated recurrent Units U-Net model

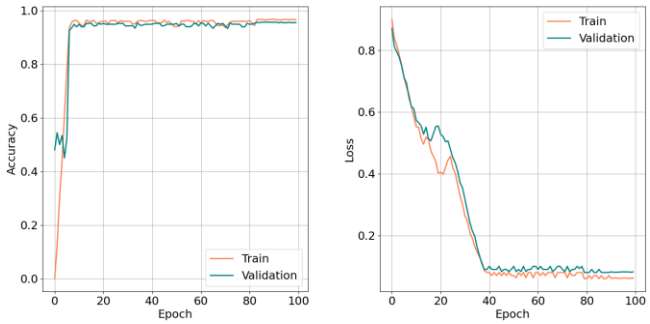



**Fig. 8.** Training and validation accuracies (left) and loss function (right) for the Residual Attention Network model

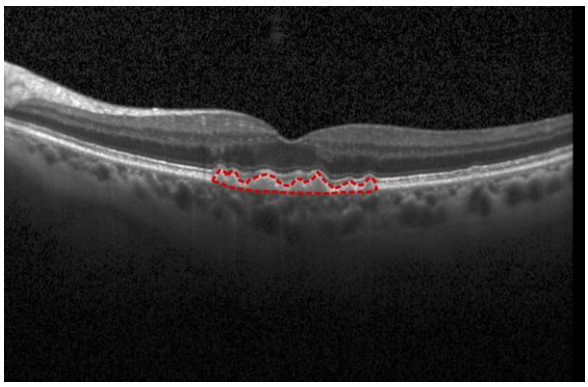
The analysed CNN architectures have high average accuracy for AVL, Drusen, and normal cases detection, above 89% (Fig. 3-5). The Deep CNN-GRU Network identifies AVL with 94.74% accuracy. This disease is rarely misclassified with Drusen, up to 5.26%. Drusen images are classified with the accuracy of 85.06%. They are confused with other classes only up to 8.06%. In case of healthy cases, they are very rarely misclassified with images presenting retinal diseases. The Convolutional Gated Recurrent Units U-Net detects all classes with high accuracy, ranging between 88.51% for normal cases and 90.90% for Drusen. Residual Attention Network identifies AVL and Drusen with the highest accuracy, exceeding 94%. Slightly worse result is obtained for normal images. As in case of Deep CNN-GRU, the Convolutional Gated Recurrent Units U-Net and Residual Attention Network confuse AVL only with Drusen, while Drusen and healthy images are rarely misleading with other classes.

The accuracies obtained while training and validation and loss function provide the effectiveness of three CNN-based architectures (Fig. 6-8). The learning dynamics are similar for all models. The architectures reach a high level of accuracy, reaching 20 epochs or even earlier. The difference between the training and validation curves remains low. In the context of the difficult, sparse AVL class, high validation accuracy suggests that the network effectively recognizes even relatively rare disease patterns. At the same time, the loss function decreases smoothly for both training and validation, stabilizing within the range of 0.1-0.2.

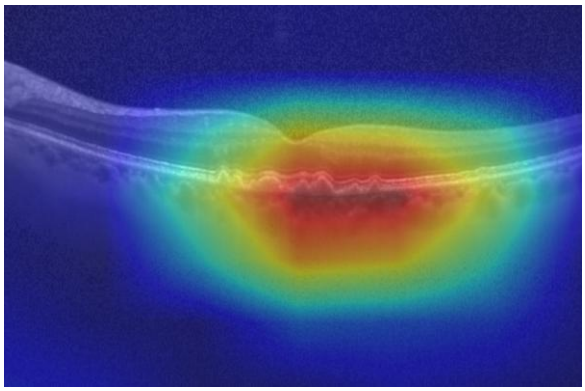
All obtained results confirm the effectiveness of the deep learning approach for AVL, Drusen, and normal cases identification.

### 4.2. Explainability of CNN-based architectures

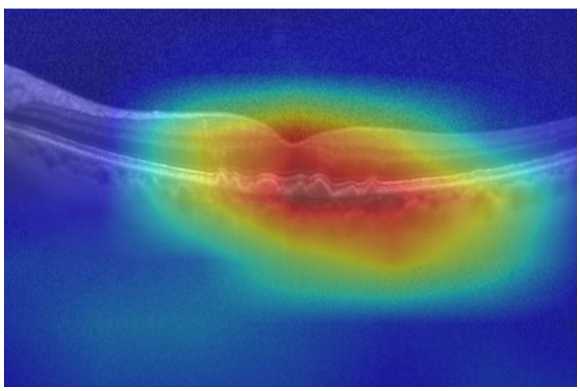
The Grad-CAM heatmaps are applied to highlight the image regions that are the most relevant for AVL, Drusen, and healthy cases identification. The warmer colour is the part of the image of the greatest importnt for the model's decision is. The colder tones indicate the lower significante of the particular regiion.



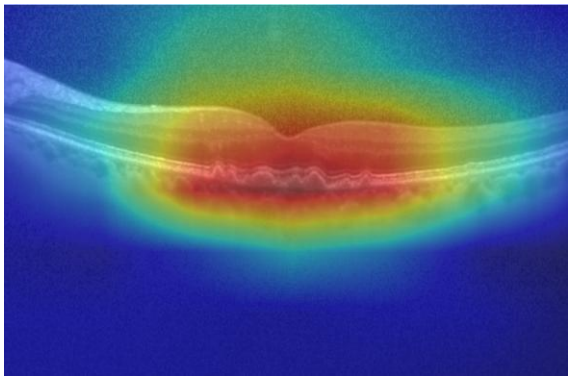
a) Ground truth DRUSEN



b) Deep CNN-GRU



c) CGRU U-Net



d) RAN

**Fig. 9.** An exmaple ground truth image (a) and Grad-CAMs for Deep CNN-GRU (b), CGRU U-Net (c) and RAN (d) models

The Grad-CAM heatmaps for Drusen identification by the single CNN-based models are depicted in Fig. 9 b)-d). The symptoms of this disease are shown in the Fig. 9 a), marked with a red border by a professional ophthalmologist. It should be emphasized that all models considered in this study focus on the image region containing retinal distortions in the form of Drusen. However, in the case of the Deep CNN-GRU Network and Convolutional Gated Recurrent Units U-Net, it can be observed that the leftmost part of the distortions is omitted as important for the diagnosis of this disease. The Residual Attention Network model, on the other hand, increases the image area, based on which it makes decisions about classifying Drusen. As a result, the most significant region, dark red, concentrates all the distortions associated with this retinal disease. Therefore, RAN is the most suitable model for identification of this type of retinal lesions. The Grad-CAM heatmaps play a crucial role in improving the explainability of the models, which is especially important in diseases identification. This technique proves the models involved in this study to make the decision based on relevant regions. Moreover, this technique highlights the distortions and stay as an additional a tool for ophthalmologists to indicate the relevant areas on OCT imaging.

### 4.3. Ensemble learning

The experiments involving bagging, boosting, and stacking ensemble learning techniques are performed in order to enhance the performance of the analysed architectures in this study. The new models are created utilizing all three single CNN-based networks. The final response of the bootstrap model is determined by majority voting. In stacking technique, the XGBoost was chosen as a meta-model. In case of all types of ensemble learning the performance of retinal diseases identification is improved. Bagging enhanced the detection rate of AVL, Drusen, and normal cases by 6.91%, 7.82%, and 5.29% compared to the Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention Network, respectively. Boosting method increased the mean accuracy by 6.39%, 7.30%, and 4.77% as opposed to the Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention Network, respectively. The stacking model achieved the highest accuracy, enhancing the model's performance by 7.43%, 8.34%, and 5.81% of the Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention Network, respectively. The obtained results in Tab. 3 clearly prove that bagging, boosting, and stacking ensemble learning methods are suitable techniques to improve the model's ability to recognize AVL, Drusen and that discriminates them from healthy cases.

**Tab. 3.** Precision, recall and Fi-score results for different ensemble methods. All values are in %

| Model | Acc. | Class | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Bagging | 97.41 | AVL | 94.74 | 94.74 | 94.74 |
| | | Drusen | 96.59 | 97.70 | 97.14 |
| | | Healty | 98.84 | 97.70 | 98.27 |
| Boosting | 96.89 | AVL | 94.44 | 89.47 | 91.89 |
| | | Drusen | 97.67 | 96.55 | 97.11 |
| | | Healty | 96.63 | 98.85 | 97.73 |
| Stacking | 97.93 | AVL | 94.74 | 94.74 | 94.74 |
| | | Drusen | 97.73 | 98.85 | 98.29 |
| | | Healty | 98.84 | 97.70 | 98.27 |

### 4.4. Comparison with the state-of-the-art

In order to compare the results obtained in this study for the proposed single CNN-based architectures and their combinations utilizing ensemble learning techniques, the state-of-the-art deep learning models, which are often applied in disease identification based on OCT imaging, were selected. The VGG-16, ResNet-18, InceptionV3, and DenseNet-121 were chosen due to their high-performance achieved. Their ability to identify the AVL, Drusen, and healthy cases using the dataset from this study, are computed (Tab. 4). The ResNet-18, InceptionV3, and DenseNet-121 demonstrates good model performance, obtaining average model accuracy of 82.86%, 82.64%, and 82.80%, respectively. Slightly worse accuracy, 78.64%, was obtained for the VGG-16 model. It should be noted that these models achieved the lowest precision for AVL identification, which also resulted in the lowest F1-score values for this disease.

**Tab. 4.** Comparison for Accuracy, Precision, recall and Fi-score results for common CNN architectures. All values are in %

| Model | Acc. | Class | Precision | Recall | F1 |
|---|---|---|---|---|---|
| VGG-16 [46] | 78.64 | AVL | 48.00 | 84.00 | 62.00 |
| | | Drusen | 81.00 | 76.00 | 79.00 |
| | | Healty | 84.00 | 76.00 | 80.00 |
| ResNet-18 [47] | 82.86 | AVL | 48.00 | 84.00 | 62.00 |
| | | Drusen | 87.00 | 82.00 | 84.00 |
| | | Healty | 92.00 | 86.00 | 87.00 |
| InceptionV3 [48] | 82.64 | AVL | 52.00 | 79.00 | 62.00 |
| | | Drusen | 90.00 | 83.00 | 86.00 |
| | | Healty | 98.00 | 86.00 | 88.00 |
| DenseNet-121 [49] | 82.80 | AVL | 54.00 | 74.00 | 62.00 |
| | | Drusen | 95.00 | 83.00 | 88.00 |
| | | Healty | 88.00 | 92.00 | 90.00 |

All studies were conducted using a computer equipped with a 24-core AMD Ryzen Threadripper PRO 7965WX processor clocked at 4.20 GHz. The computer uses 512 GB of RAM. The system also benefits from the support of a NVIDIA GeForce RTX 4090 graphics card. The equipment runs in a 64-bit operating system

environment. The comparison of computational complexity and average processing time is presented in Tab. 5.

**Tab. 5.** Comparison of inference complexity and speed for the baseline CNN and Deep CNN-GRU variants. Timing measurements for RTX 4090 and single-core Ryzen 7950X

| Model | Params [M] | GFLOPs | RTX 4090 [ms] | Ryzen 7950X [ms] |
|---|---|---|---|---|
| VGG-16 [46] | 134.3 | 15.5 | 42.5 | 332.0 |
| ResNet-18 [47] | 11.2 | 1.8 | 5.0 | 39.0 |
| InceptionV3 [48] | 21.8 | 32.2 | 88.8 | 691.0 |
| DenseNet-121 [49] | 7.0 | 2.9 | 7.9 | 62.0 |
| Deep CNN-GRU [17] | 11.1 | 22.4 | 63.7 | 582.0 |
| GRU U-Net [18] | 9.2 | 39.8 | 24.9 | 382.0 |
| RAN [45] | 14.8 | 26.3 | 19.0 | 225.0 |
| Bagging | 30.5 | 47.6 | 134.2 | 1201.0 |
| Boosting | 29.9 | 48.3 | 129.8 | 989.0 |
| Stacking | 30.7 | 46.7 | 119.9 | 1007.0 |

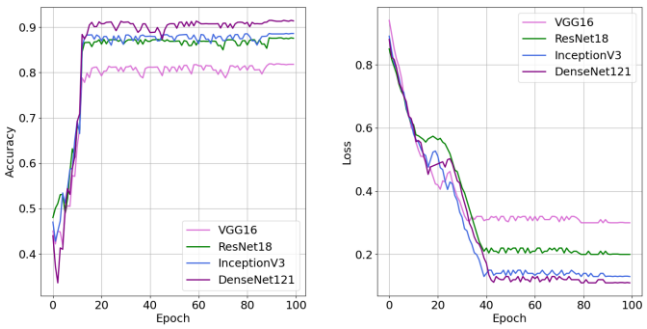



**Fig. 10.** Validation accuracies (left) and loss function (right) for the state-of-the-art models

At the current level of development of retinal diseases detection, it can be seen that the Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention Network provide significantly better ability to recognize AVL, Drusen, and healthy cases based on OCT imaging, achieving average models accuracy in the range of 89.59%-92.12%. None of the state-of-the-art architectures supplies these distortions identification at a similar level using the same dataset. The VGG-16, ResNet-18, InceptionV3, and DenseNet-121 achieved the highest ability to recognize healthy cases. The same situation occurs with the Deep CNN-GRU Network. However, the Convolutional Gated Recurrent Units U-Net, Residual Attention Network, as well as boosting and stacking models identify retinal distortions with the highest performance, which is extremely important in medical diagnosis. Moreover, the ensemble learning methods prove to enhance ability to recognize all retinal diseases and healthy cases. Due to overcoming individual model limitations and delivering more reliable results, the new models are characterized by better generalization to new samples.

The performance of all architectures, involved in this study, were verified using accuracy and loss curves (Fig. 8, 10). All accuracy curves increase over time gradually, enhancing the model's performance. All networks learn from the training data properly, in stable and consistent manner up to reach convergence. It means that all models understood important patterns to make accurate predictions. The loss curves indicate that the models improve their performance while the loss gradually decreases.

The Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention Network, as well as proposed new ensemble learning models outperform the state-of-the-art architectures.

## 5. CONCLUSIONS AND FUTURE WORKS

Acquired vitelliform lesions encompass a broad spectrum of retinal diseases. may be easily confused with drusen that is why accurate method of their identification is of great importance. This study focuses on developing an accurate automatic tool to identify AVL and Drusen distortions, as well as to distinguish them from healthy cases based on OCT imaging. In this study, three CNN-based models, the Deep CNN-GRU Network, Convolutional Gated Recurrent Units U-Net, and Residual Attention Network were applied for this purpose. Their performance outperform the state-of-the-art models, such as the VGG-16, ResNet-18, InceptionV3, and DenseNet-121. The proposed models are distinguished by their ability to recognize all retinal distortions, which is of great importance in medical diagnosis. Moreover, this study presents that ensemble learning by aggregating single CNN-based architectures results in enhancing the model's performance. Bagging, boosting, and stacking proved to be adequate ensemble learning methods for retinal diseases identification. The stacking technique obtained the highest accuracy of 97.93% with the precision for alll classes exceeding 94%. These results allow us to assume that this automatic tool is a suitable to assist ophthalmologists in recognizing AVL, Drusen, and healthy cases. This kind of software allows these cases to be identified quickly and with high reliability.

Although the results of this study are promising, some limitations should be addressed in future works. More CNN-based architectures and their combination utilizing ensemble learning techniques should be verified for AVL, Drusen, and healthy cases identification. Transformer vision-based models may be applied and compared to CNN-based solutions for retinal diseases. Moreover, the AVL dataset should be extended to obtain more balanced dataset.

## REFERENCES

1. Subasree S, KSakthivel N., Baby Kalpana Y, Manikandan M. A deep learning–based method for diabetic retinopathy classification using 1-dimensional complex-valued convolutional neural networks to improve early diagnosis and treatment outcomes. Int J Diabetes Dev Ctries. 2025 Jul 3. https://doi.org/10.1007/s13410-025-01517-7
2. Muntaqim MdZ, Smrity TA, Miah ASM, Kafi HM, Tamanna T, Farid FA, Rahim MA, Karim HA, Mansor S. Eye Disease Detection Enhancement Using a Multi-Stage Deep Learning Approach. IEEE Access. 2024;1–1. https://doi.org/10.1109/access.2024.3476412
3. Baba S, Kumari P, Saxena P. Retinal Disease Classification Using Custom CNN Model From OCT Images. Procedia Computer Science. 2024;235:3142–52. https://doi.org/10.1016/j.procs.2024.04.297
4. Shahzadi Z, Zubair M. Multiclass Classification of Retinal Disorders Using Optical Coherence Tomography Images. In: 2024 Horizons of Information Technology and Engineering (HITE) [Internet]. Lahore, Pakistan: IEEE; 2024;1–6. Available from: https://ieeexplore.ieee.org/document/10777173/ https://doi.org/10.1109/hite63532.2024.10777173

5. Rahman MM, Roy A, Karim DZ. Ret-Detect: Deep Learning-Driven Automated Detection of Retinal Diseases. In: 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) [Internet]. Chittagong, Bangladesh: IEEE; 2025; 1–6. Available from: https://ieeexplore.ieee.org/document/11013820/ https://doi.org/10.1109/ecce64574.2025.11013820
6. Powroznik P, Skublewska-Paszkowska M, Nowomiejska K, Gajda-Deryło B, Brinkmann M, Concilio M, Toro MD, Rejdak R. Residual self-attention vision transformer for detecting acquired vitelliform lesions and age-related macular drusen. Sci Rep. 2025 May 16;15(1). https://doi.org/10.1038/s41598-025-02299-y
7. Al-Amiry SH, Al-Juboori AM. OCTm: Custom Deep Learning Model for Diagnosing Retinal Diseases. In: 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA) [Internet]. Balikpapan, Indonesia: IEEE; 2024; 639–43. Available from: https://ieeexplore.ieee.org/document/10747941/ https://doi.org/10.1109/icsintesa62455.2024.10747941
8. Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng CY, Wong TY. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. The Lancet Global Health. 2014 Feb;2(2):e106–16. https://doi.org/10.1016/s2214-109x(13)70145-1
9. Elsharkawy M, Sharafeldeen A, Khalifa F, Soliman A, Elnakib A, Ghazal M, Sewelam A, Thanos A, Sandhu HS, El-Baz A. A Clinically Explainable AI-Based Grading System for Age-Related Macular Degeneration Using Optical Coherence Tomography. IEEE J Biomed Health Inform. 2024 Apr;28(4):2079–90. https://doi.org/10.1109/jbhi.2024.3355329
10. Li F, Chen H, Liu Z, Zhang X dian, Jiang M shan, Wu Z zheng, Zhou K qian. Deep learning-based automated detection of retinal diseases using optical coherence tomography images. Biomed Opt Express. 2019 Dec 1;10(12):6204. https://doi.org/10.1364/boe.10.006204
11. Gehrs KM, Anderson DH, Johnson LV, Hageman GS. Age-related macular degeneration—emerging pathogenetic and therapeutic concepts. Annals of Medicine. 2006 Jan;38(7):450–71. https://doi.org/10.1080/07853890600946724
12. Querques G, Forte R, Querques L, Massamba N, Souied EH. Natural Course of Adult-Onset Foveomacular Vitelliform Dystrophy: A Spectral-Domain Optical Coherence Tomography Analysis. American Journal of Ophthalmology. 2011 Aug;152(2):304–13. https://doi.org/10.1016/j.ajo.2011.01.047
13. Dalvin LA, Pulido JS, Marmorstein AD. Vitelliform dystrophies: Prevalence in Olmsted County, Minnesota, United States. Ophthalmic Genetics. 2017 Mar 4;38(2):143–7. https://doi.org/10.1080/13816810.2016.1175645
14. Balaratnasingam C, Hoang QV, Inoue M, Curcio CA, Dolz-Marco R, Yannuzzi NA, Dhrami-Gavazi E, Yannuzzi LA, Freund KB. Clinical Characteristics, Choroidal Neovascularization, and Predictors of Visual Outcomes in Acquired Vitelliform Lesions. American Journal of Ophthalmology. 2016 Dec;172:28–38. https://doi.org/10.1016/j.ajo.2016.09.008
15. Roy A, Abdullah R, Ahmed F, Mashfi S, Khan SH, Karim DZ. RetNet: Retinal Disease Detection using Convolutional Neural Network. In: 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE) [Internet]. Chittagong, Bangladesh: IEEE; 2023; 1–6. Available from: https://ieeexplore.ieee.org/document/10101661/. https://doi.org/10.1109/ecce57851.2023.10101661
16. Powroźnik P, Skublewska-Paszkowska M, Nowomiejska K, Aristidou A, Panayides A, Rejdak R. Deep convolutional generative adversarial networks in retinitis pigmentosa disease images augmentation and detection. Adv Sci Technol Res J. 2024 Nov 20;19(2):321–40. https://doi.org/10.12913/22998624/196179
17. Powroznik P, Skublewska-Paszkowska M, Rejdak R, Nowomiejska K. Automatic Method of Macular Diseases Detection Using Deep CNN-GRU Network in OCT Images. Acta Mechanica et Automatica. 2024;18(4):697–706. https://doi.org/10.2478/ama-2024-0074
18. Skublewska-Paszkowska M, Powroznik P, Rejdak R, Nowomiejska K. Application of Convolutional Gated Recurrent Units U-Net for Distinguishing between Retinitis Pigmentosa and Cone–Rod Dystrophy. Acta Mechanica et Automatica. 2024;18(3):505–13. https://doi.org/10.2478/ama-2024-0054
19. Kumar Y, Gupta S. Deep Transfer Learning Approaches to Predict Glaucoma, Cataract, Choroidal Neovascularization, Diabetic Macular Edema, DRUSEN and Healthy Eyes: An Experimental Review. Arch Computat Methods Eng. 2023 Jan;30(1):521–41. https://doi.org/10.1007/s11831-022-09807-7
20. Thakoor KA, Koorathota SC, Hood DC, Sajda P. Robust and Interpretable Convolutional Neural Networks to Detect Glaucoma in Optical Coherence Tomography Images. IEEE Trans Biomed Eng. 2021;68(8):2456–66. https://doi.org/10.1109/tbme.2020.3043215
21. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) [Internet]. 2017. https://doi.org/10.48550/ARXIV.1711.11279
22. An S, Teo K, McConnell MV, Marshall J, Galloway C, Squirrell D. AI explainability in oculomics: How it works, its role in establishing trust, and what still needs to be addressed. Progress in Retinal and Eye Research. 2025;106:101352. https://doi.org/10.1016/j.preteyeres.2025.101352
23. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis. 2020;128(2):336–59. https://doi.org/10.1007/s11263-019-01228-7
24. Powroznik P, Skublewska-Paszkowska M, Rejdak R, Nowomiejska K. Automatic Method of Macular Diseases Detection Using Deep CNN-GRU Network in OCT Images. Acta Mechanica et Automatica. 2024;18(4):697–706. https://doi.org/10.2478/ama-2024-0074
25. Ghislain F, Beaudelaire ST, Atangana R, Daniel T. An improved deep learning approach for automated detection of multiclass eye diseases. Array. 2025;27:100452. https://doi.org/10.1016/j.array.2025.100452
26. Qaiser M, Abbas MS, Muhammad A, Haider K, Khan H. TezNet: An Efficient Convolutional Neural Network for Optical Coherence Tomography Classification. In: 2024 International Conference on Robotics and Automation in Industry (ICRAI) [Internet]. Rawalpindi, Pakistan: IEEE; 2024; 1–6. Available from: https://ieeexplore.ieee.org/document/10894682/. https://doi.org/10.1109/icrai62391.2024.10894682
27. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
28. Asim T, Akbar S, Shahzore U, Rehman A, Saba T, Ayesha N. Retinal Image Analysis for Detection of Diabetic Macular Edema Using Transfer Learning. In: 2025 8th International Conference on Data Science and Machine Learning Applications (CDMA) [Internet]. Riyadh, Saudi Arabia: IEEE; 2025; 144–9. Available from: https://ieeexplore.ieee.org/document/10908769/ https://doi.org/10.1109/cdma61895.2025.00030
29. Detection of Retinal Disease from Optical Coherence Tomography Images Using CNN Models. In: Lecture Notes in Networks and Systems [Internet]. Cham: Springer Nature Switzerland;14–21. Available from: https://link.springer.com/10.1007/978-3-031-73318-5_2. https://doi.org/10.1007/978-3-031-73318-5_2
30. Ajmal MM, Mumtaz R, Mumtaz S, Temdee P, Asif MDA, Khan MA, Khan S. Deep learning based eye disease classification using Optical Coherence Tomography (OCT) images. Experimental Eye Research. 2026;268:111017. https://doi.org/10.1016/j.exer.2026.111017
31. Nagamani GM, Sudhakar T. An improved dynamic-layered classification of retinal diseases. IJ-AI. 2024;13(1):417. https://doi.org/10.11591/ijai.v13.i1.pp417-429
32. George N, Shine L, N A, Abraham B, Ramachandran S. A two-stage CNN model for the classification and severity analysis of retinal and choroidal diseases in OCT images. International Journal of Intelligent

Networks. 2024;5:10–8. https://doi.org/10.1016/j.ijin.2024.01.002
33. G S, Yaligar N, Kumar KR, Padaki A, Dawe R, Katagi S. State-of-the-Art Deep Learning Strategies for Multi-Class Classification of Retinal OCT Images. In: 2024 Second International Conference on Advances in Information Technology (ICAIT) [Internet]. Chikkamagaluru, Karnataka, India: IEEE; 2024; 1–7. Available from: https://ieeexplore.ieee.org/document/10690676/ https://doi.org/10.1109/icait61638.2024.10690676
34. Bansal P, Harjai N, Saif M, Mugloo SH, Kaur P. Utilization of big data classification models in digitally enhanced optical coherence tomography for medical diagnostics. Neural Comput & Applic. 2024;36(1): 225–39. https://doi.org/10.1007/s00521-022-07973-0
35. Noor F, Nusaiba FT, Sami TA, Rahman MdA. Retinal Disease Detection by Optical Coherence Tomography (OCT)-Based Deep Learning. In: 2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI) [Internet]. Narayanganj, Bangladesh: IEEE. 2024; 1–6. Available from: https://ieeexplore.ieee.org/document/10951065/ https://doi.org/10.1109/sti64222.2024.10951065
36. Udayaraju P, Jeyanthi P, Sekhar BVDS. A hybrid multilayered classification model with VGG-19 net for retinal diseases using optical coherence tomography images. Soft Comput. 2023;27(17):12559–70. https://doi.org/10.1007/s00500-023-08928-w
37. KSD, TIM, Nataraj NK, Venu A, Prasad AK, Paul AM. CARE: A Comprehensive AI Retinal Expert for Disease Detection Using OCT Imaging and Explainable AI. In: 2025 IEEE International Conference on Advances in Computing Research On Science Engineering and Technology (ACROSET) [Internet]. INDORE, India: IEEE. 2025;1–6. Available from: https://ieeexplore.ieee.org/document/11281222/ https://doi.org/10.1109/ACROSET66531.2025.11281222
38. Sabitha R, Raja MAM, Manikandan K, R BB, Shilpa K, V P. Explainable Artificial Intelligence (AI) Powered Deep Learning Model for Interpretable Retinal Fundus Diagnosis using Biomedical Images. In: 2026 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) [Internet]. Chennai, India: IEEE. 2026; 1–9. Available from: https://ieeexplore.ieee.org/document/11479180/ https://doi.org/10.1109/ICSES66558.2026.11479180
39. Elsharkawy M, Sharafeldeen A, Khalifa F, Soliman A, Elnakib A, Ghazal M, Sewelam A, Thanos A, Sandhu HS, El-Baz A. A Clinically Explainable AI-Based Grading System for Age-Related Macular Degeneration Using Optical Coherence Tomography. IEEE J Biomed Health Inform. 2024;28(4):2079–90. https://doi.org/10.1109/JBHI.2024.3355329
40. Masum AH, Rony SH, Arpita HD, Al Ryan A. Comprehensive Approach to Retinal Disease Classification from Optical Coherence Tomography Images Using Average Ensemble Learning. In: 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) [Internet]. Chittagong, Bangladesh: IEEE. 2025; 1–6. Available from: https://ieeexplore.ieee.org/document/11013984/ https://doi.org/10.1109/ecce64574.2025.11013984
41. Romero-Oraá R, Herrero-Tudela M, López MI, Hornero R, Romero-Aroca P, García M. An Ensemble Model for Fundus Images to Aid in Age-Related Macular Degeneration Grading. Diagnostics. 2025;15(20):2644. https://doi.org/10.3390/diagnostics15202644
42. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VAN, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
43. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imag Rad Onc. 2021;65(5):545–63. https://doi.org/10.1111/1754-9485.13261
44. Liang W, Liang Y, Jia J. MiAMix: Enhancing Image Classification through a Multi-stage Augmented Mixed Sample Data Augmentation Method [Internet]. arXiv; 2023. Available from: https://arxiv.org/abs/2308.02804. https://doi.org/10.48550/ARXIV.2308.02804
45. Nowomiejska K, Powroźnik P, Skublewska-Paszkowska M, Adamczyk K, Concilio M, Sereikaite L, Zemaitiene R, Toro MD, Rejdak R. Residual Attention Network for distinction between visible optic disc drusen and healthy optic discs. Optics and Lasers in Engineering. 2024;176:108056. https://doi.org/10.1016/j.optlaseng.2024.108056
46. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [Internet]. arXiv; 2014. Available from: https://arxiv.org/abs/1409.1556. https://doi.org/10.48550/ARXIV.1409.1556
47. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Las Vegas NV USA: IEEE. 2016; 770–8. Available from: http://ieeexplore.ieee.org/document/7780459/ https://doi.org/10.1109/CVPR.2016.90
48. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Las Vegas NV USA: IEEE. 2016; 2818–26. Available from: http://ieeexplore.ieee.org/document/7780677/. https://doi.org/10.1109/CVPR.2016.308
49. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Honolulu, HI: IEEE. 2017; 2261–9. Available from: https://ieeexplore.ieee.org/document/8099726/. https://doi.org/10.1109/CVPR.2017.243

The work has been accomplished under the research project „Lubelska Unia Cyfrowa - Wykorzystanie rozwiązań cyfrowych i sztucznej inteligencji w medycynie - projekt badawczy", no. MEiN/2023/DPI/2194.

Pawel Powroźnik: https://orcid.org/0000-0002-5705-4785

Maria Skublewska-Paszkowska: https://orcid.org/0000-0002-0760-7126

Katarzyna Nowomiejska: https://orcid.org/0000-0002-5805-8761

Robert Rejdak: https://orcid.org/0000-0003-3321-2723

Marcin Derlatka: https://orcid.org/0000-0002-6736-5106