Multimedia Tools and Applications (2024) 83:17921–17939
https://doi.org/10.1007/s11042-023-16129-8

A CNN-based hybrid model to detect glaucoma disease

Cinare Oguz[1] · Tolga Aydin[1] · Mete Yaganoglu[1]

Received: 3 June 2022 / Revised: 6 May 2023 / Accepted: 26 June 2023 /
Published online: 17 July 2023

Abstract
Glaucoma is an eye disease caused by damage to the optic nerves and is a common cause of incurable blindness worldwide. If glaucoma is diagnosed early, vision loss can be prevented with regular exams and treatment. If diagnosed too late, the disease can cause severe damage to the optic nerve that cannot be reversed, leading to loss of central vision and blindness. Therefore, early diagnosis of the disease is critical. Most of the studies conducted in recent years have presented Deep Learning (DL) based architectures that use an automatic computerized system based on segmentation of hand-made features in fundus images. In this study, we seek to help experts detect glaucoma using a model that combines Deep Learning and Machine Learning using raw fundus images. Deep features are extracted using a new Convolutional Neural Networks (CNN) model. Deep features have been used in popular traditional Machine Learning methods (ML) for classification such as Adaboost, k Nearest Neighbor (kNN), Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Naive Bayes (NB). The performances of the hybrid models were evaluated using the ACRIMA dataset of 705 images. The dataset is reserved for 80% training and 20% testing data. Experimental results show that the hybrid model of CNN and Adaboost has the highest success rate with 92.96% accuracy, 93.75% F1 score and an AUC value of 0.928.

Keywords Glaucoma · Deep learning · CNN · AdaBoost · ACRIMA · Hybrid models

1 Introduction

Glaucoma is an eye disease caused by damage to the optic nerves [32].The optic nerves receive nerve signals from the retina and transmit these electrical signals to the brain. Therefore, damage to the optic nerves causes the brain to not perceive vision properly. Generally, the optic nerves are destroyed by high eye pressure.

If the disease is detected early, loss of vision can be prevented by regular examinations and treatment. Treatment can only help preserve the remaining vision. Lost nerve

✉ Mete Yaganoglu
yaganoglu@atauni.edu.tr

Cinare Oguz
cinare.oguz.91@gmail.com

Tolga Aydin
atolga@atauni.edu.tr

[1] Department of Computer Engineering, Faculty of Engineering, Ataturk University, Erzurum, Turkey

tissue cannot be regenerated. If the disease is diagnosed too late, it can cause severe damage to the optic nerve that cannot be repaired, leading to loss of central vision and blindness. Therefore, early diagnosis of the disease is essential [26].

Diagnosis of glaucoma includes examination of structural deterioration of the optic disk, macula, and Retinal Nerve Fiber Layer (RNFL) before the onset of vision loss. Fundoscopy, Optical Coherence Tomography (OCT), Confocal Scanning Laser Ophthalmoscopy (CSLO), and Scanning Laser Polarimetry (SLP) are used as retinal imaging techniques in glaucoma detection. OCT is used to examine the RNFL and optic disc. CSLO is used for analysis of the optic nerve and retina. It depends on an observer. SLP is used to measure RNFL thickness. These imaging techniques are bulky and costly. Fundoscopy is used for intraocular examination with a portable fundus camera. Color Fundus Imaging (CFI) provides analysis of the optic disk, macula, and retina. It is also inexpensive compared to other techniques [23]. Figure 1 shows the retinal image of a person with glaucoma and the retinal image of a healthy person.

In the clinical detection of glaucoma, experts manually analyze the morphological features of the optic disk, RNFL, and macular structures using retinal images. However, this process is time-consuming due to patient density, and results vary from specialist to specialist. With automated glaucoma diagnostic systems, it may be possible to diagnose glaucoma quickly and reliably. In recent years, ML and DL methods have been used for image classification and segmentation in medical images (e.g., CT, CFI, etc.) and successful results have been obtained. Traditional ML methods provide fast results, but these methods require structured data. Manual processing of this data (feature extraction from the image) can lead to erroneous results due to lack of attention. Since this process (feature extraction) is automatic with the DL models, errors that can occur are avoided. The use of DL with unstructured data increases the accuracy of the result. However, because the number of parameters in DL models is very high, high computational resources and large data sets are required for training. This problem can be overcome by data augmentation and a hybrid model.

CNNs have been particularly used for tasks such as detecting retinal diseases like diabetic retinopathy and macular degeneration using retinal images [12, 20].

Most automated computerized systems developed for glaucoma detection have used segmentation-based manually generated features in fundus images. In this study, raw fundus images are used in a different way.

In this study, we present a hybrid model that integrates the methods of ML and Convolutional Neural Networks for glaucoma detection using CFIs to develop a Decision Support System (DSS) for specialists in glaucoma detection.

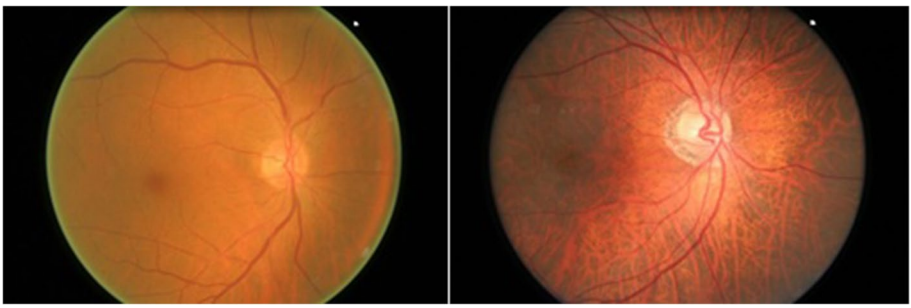

Fig. 1 (Left) Retinal image of a healthy individual (right) Retinal image of an individual with glaucoma

2 Related work

In this section, a summary of the literature based on Deep Learning-based automatic diagnostic systems for glaucoma disease detection is discussed. The works in the literature were reviewed sequentially based on the research methods, datasets used, and experimental results. These studies are grouped according to the retinal image types used, such as OCT and fundus.

2.1 Glucoma detection using OCT images

A summary of studies using OCT images in the literature, can be found in Table 1. Burgansky-Eliash et al. used OCT images from a total of 47 patients, 20 of whom were in early and 27 in advanced stages of glaucoma, and 42 healthy patients. Classification methods such as regression tree, recursive partitioning, SVM, and Linear Discriminant Analysis (LDA) were examined. SVM achieved the best performance with an accuracy of 0.966 [4]. Koh et al.. presented a ANN after RNFL thickness measurement for automatic diagnosis of glaucoma. Experiments were performed with 1795 OCT images obtained from 60 glaucoma and 1001 healthy subjects. The proposed model achieved 95% accuracy [18]. Devalla et al. presented a Deep Learning based model for a glaucoma diagnostic system. OCT Studies were conducted with 40 healthy and 60 glaucoma patients. The proposed model achieved 92% sensitivity and 94% accuracy [8]. Singh et al. used OCT images for the glaucoma diagnostic system. The data set was from the open source Mendeley. Studies were performed on 70 healthy eyes and 70 glaucomas. Forty-five important features were extracted from the images and used in different classifiers (LDA, Decision Trees, kNN, SVM and RF). The highest performance was obtained using kNN with 97% AUC value [28]. Sunija et al. proposed a deep-separable Convolutional Neural Network for glaucoma detection. The private databases they used included 1105 glaucoma and 1049 normal OCT images. The proposed model showed high performance with an accuracy of 0.9963 [31]. Raja et al. presented a systematic analysis of recent studies using OCT images for a glaucoma diagnostic system. They showed successful results using OCT images for glaucoma detection [23]. Juneja et al. proposed an unified framework for the ML and DL techniques. Experiments were performed using OCT images. The proposed model achieved an F1 score of 0.96 [16].

Table 1 Summary table: Glaucoma detection studies using OCT images

Works	Method	Dataset	Performance
[4]	SVM	Glaucoma: 47 Normal: 42	96.6% accuracy
[18]	ANN	Glaucoma: 86 Normal: 1709	95% accuracy
[8]	DL	Glaucoma: 60 Normal: 40	94% accuracy
[28]	kNN	Glaucoma: 70 Normal: 70	97% AUC
[31]	Deep-separable CNN	Glaucoma: 1105 Normal: 1049	99.63% accuracy
[16]	ML and DL hybrid model	Glaucoma: 847 Normal: 263	0.96 F1-score

2.2 Glucoma detection using fundus images

A summary of studies using CFI images in the literature is given in Table 2. Chen et al. presented a CNN model for automatic diagnosis of glaucoma. The proposed model includes 4 convolutional layers and 2 fully connected layers. The performance of this model was evaluated using two different databases such as ORIGA and SCES. According to the experimental results, the model achieved 0.831 and 0.887 AUC values in two databases (ORIGA, SCES), respectively [7]. Asaoka et al. used various features such as mean deviation and standard deviation for glaucoma detection. The DL-based Forward Neural Network (FNN) was used to classify CFIs. Using this method, an accuracy of 92.5% was achieved [1]. Raghavendra et al. proposed a new CNN-based model for glaucoma disease

Table 2 Summary table: Glaucoma detection studies using CFI images

Works	Method	Dataset	Performance
[7]	6 layer-CNN	ORIGA Glaucoma: 168 Normal: 482 / SCES Glaucoma: 46 Normal: 1630	0.831/0.887 AUC
[1]	FNN	University of Tokyo Hospital Glaucoma: 87 Normal: 108	92.5% accuracy
[22]	18 layer-CNN	Kasturba Medical College Glaucoma: 837 Normal: 589	98.13% accuracy
[5]	MB-NN	2554 CFI	91.51% accuracy
[9]	Xception	ACRIMA Glaucoma: 396 Normal: 309/ HRF Glaucoma: 27 Normal: 18/ Drishti-GS1 Glaucoma: 70 Normal: 31/ RIM-ONE Glaucoma: 194 Normal: 261/ sjchoi86-HRF Glaucoma:101 Normal: 300	70.21% accuracy/ 80% accuracy/ 75.25% accuracy/ 71.21% accuracy/ 70.82% accuracy
[25]	GoogLeNet	ACRIMA / HRF / Drishti-GS1 / RIM-ONE / sjchoi86-HRF	65% accuracy/ 76% accuracy/ 55% accuracy/ 70% accuracy/ 72% accuracy
[2]	Two-step DL	ORIGA	0.874 AUC
[21]	CNN	PRV-Glaucoma+ sjchoi86-HRF Glaucoma:280 Normal: 1058	98% accuracy 83% sensitivity
[11]	CNN and RNN	University of New South Wales CFIs and 295 fundus videos.	96.2% F1-score
[29]	Community-based CNN	LAG Glaucoma:4878 Normal: 6882/ RIM-ONE Glaucoma:200 Normal: 255	99.51% accuracy / 93% accuracy
[17]	CoG-NET	RIM-One Drishti, REFUGE and ACRIMA Glaucoma:661 Normal: 1511	93.5% accuracy

detection. The proposed model uses an 18-layer Convolutional Neural Network to classify 1426 CFIs as healthy or glaucoma. This proposed model achieved an accuracy of 98.13% [22]. Chai et al. proposed a multi-branch NN model for the diagnosis of glaucoma. Experimental results showed that the developed model achieved an accuracy, sensitivity, and specificity of 0.9151, 0.9233, and 0.9090, respectively [5]. Diaz-Pinto et al. investigated DL models such as ResNet50, VGG16, VGG19, Xception, and InceptionV3 used for automatic glaucoma assessment using CFIs. The performances of the models from DL were examined on 1707 images from 5 datasets. The performances of the DL models were examined on 1707 images from 5 datasets. The highest performance was obtained by the Xception architecture with a specificity of 0.8580 and a sensitivity of 0.9346 with an AUC value of 0.9605 [9]. Serte and Serener investigated three Deep Learning architectures such as ResNet, GoogLeNet, and ResNet-152 in identifying patients with glaucoma. A total of 5 datasets were used, with 4 datasets trained and 1 dataset tested. The result was that the proposed model was comparable or better than a previous study in 80% of cases [25]. Bajwa et al. used a two-step procedure to detect glaucoma [2].First, the optic disc was detected using the CFI and then classified as normal or having glaucoma. The performance of the developed model was evaluated using the ORIGA dataset. According to the experimental results, the model achieved a value of 0.874 AUC. Patil et al. [21] developed a DL-based model with a Deep Belief Network (DBN) for glaucoma detection. Feature optimization was performed with the DBN by extracting deep features with CNN. Finally, glaucoma and healthy fundus images were determined using the linear softmax classifier. The performance of the model was evaluated using 1338 retinal fundus images. The proposed model achieved 97% specificity, 83% sensitivity, 98% accuracy. Gheisari et al. proposed a combined system of CNN and RNN techniques. Experiments were performed with 1810 CFIs and 295 fundus videos. While the basic CNN achieved an F1 score of 79.2%, the proposed model achieved an F1-Score of 96.2% [11]. David presented a Community-based CNN model. He conducted experiments with the LAG and RIM-ONE datasets. While the proposed model achieved 99.51% accuracy with the LAG dataset, it achieved 93% accuracy with the RIM-ONE dataset [29]. Juneja et al. proposed a DL-based CoG-NET model for an automatic glaucoma diagnosis system. The proposed model showed higher performance than the most recently used technologies (InceptionV3, Xception, ResNet, VGG16, DenseNet) with an accuracy value of 93.5% [17].

CFI is widely used by professionals to diagnose glaucoma because it clearly shows the structural deterioration of the optic disc. OCT The images are also used by researchers to diagnose glaucoma because of their fast data acquisition rate and rapid diagnosis [27].

Fundus images are less expensive than OCT. Therefore, fundoscopy may facilitate glaucoma detection in situations where more expensive equipment is not available.

When the studies in the literature are examined, Fully (end-to-end) CNN models, Reccurrent Neural Network (RNN) based models, Transfer Learning based models and two-step frameworks based on segmentation are suggested. Improves overall performance for Optical Cup(OC) and Optical Disc(OD) segmentation on end-to-end CNN-based models. It provides the reuse of the necessary trainable parameters in the training process. However, the need for preprocessing and a long training time are limitations of such models. In RNN-based methods, multipath architectures help to obtain semantic features. RNN also supports capturing high-level information. The fact that it requires polar transformation as a preprocessing step and its evaluation using a single data set can be counted as the limitations of this method. It improves segmentation performance in glaucoma screening in transfer learning based models. However, pre-processing,

pre-trained weights and some transfer learning requirements are disadvantages of these methods. The obvious advantage of two-step frameworks is that the OP and OC boundaries are set and given to the deep learning model to classify over the particular region. Thus, an examination cannot be made on the whole figure. It can also achieve better success than single-step frameworks using small datasets. As a disadvantage, these models are prone to compound errors. An error from the first stage (segmentation) propagates throughout the entire process and can lead to errors in the second stage and incorrect predictions. In addition, developing these models requires more medical expertise and more time as more clinical explanations are needed in the training data.

Traditional ML methods provide faster results than DL-based classification, but these methods require structured data. Manual processing of this data (feature extraction from the image) can lead to erroneous results due to a lack of attention. Since this process (feature extraction) is automatic in DL models, possible errors are prevented. Using DL with unstructured data increases the accuracy of the result. However, DL-based models require high computational resources for training because the number of parameters is too large. By combining the traditional ML structure and DL structure using a hybrid model, while feature extraction was made from unstructured data with DL structure, the classification process can performed using these attributes with traditional ML structure. Thus, we can reduce the cost and training time. Unlike end-to-end CNN, Transfer Learning, and RNN-based methods, high computational resources and long training times can be avoided by using fewer parameters in a hybrid model. However, high performance can be achieved.

The contributions of the article are as follows:

- We proposed a new CNN-based hybrid model to complement the existing models.
- The proposed CNN-based hybrid model was compared with the end to end CNN model.
- Generally, classifications are performed with CNN-based pre-trained transfer learning models on retinal images. Comparisons were made between the proposed hybrid model and pre-trained deep learning models.

3 Methodology

The method proposed in this study to detect glaucoma patients is shown in Fig. 2. As can be seen in the figure, preprocessing of the retinal images was performed in the first stage to achieve more successful results with the data set. The dataset consists of 80% training data and 20% test data. Deep features were obtained using CNN. Then, the obtained features were used in different classification methods and the prediction process was performed. Finally, the performance of the classification methods was evaluated using the performance measures of accuracy, F1 score, AUC, specificity, sensitivity, and precision.

3.1 Dataset

ACRIMA datasets [9] were used for the experiments. The database consists of images from the ACRIMA project to develop automatic classification algorithms for retinal images. Images were acquired with the consent of glaucoma and normal patients and in accordance with the ethical standards of the 1964 Declaration of Helsinki. CFIs were acquired using the Topcon TRC retinal camera and the IMAGEnet imaging system with a 35-degree angle of view. Images were labelled by two glaucoma specialists with 8 years of experience

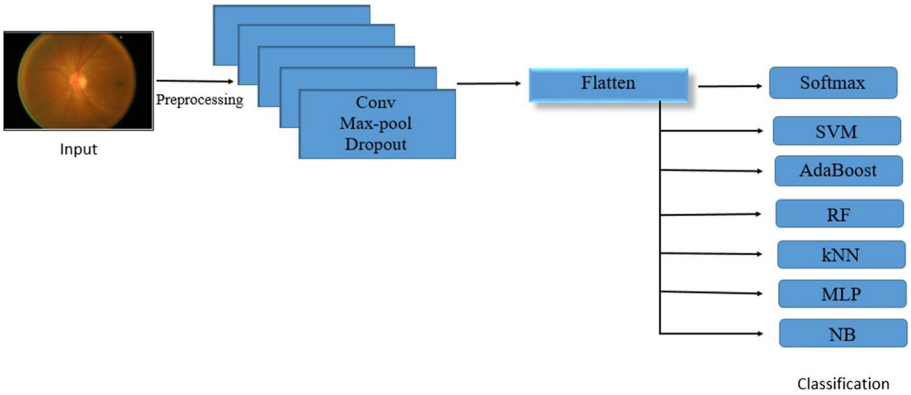

Fig. 2 Proposed model of glaucoma detection

independent of other clinical information. The data set included a total of 705 images, of which 396 were glaucoma-positive and 306 were glaucoma-negative. In the study, 142 (20%) of the colour fundus images were used for testing. Table 3 shows the details of the data set.

3.2 Preprocessing

Preprocessing is the process of preparing the inputs to the convolutional network. The input image was cropped to 1054 X 1054 pixels and the size of the image was reduced to 256 X 256 pixels. The cropped CFIs were then enlarged with 40-degree rotation, zoom, pan and flip to increase the number of CFIs and improve the performance of the convolutional networks.

For models based on Deep Learning, a value between 0 and 1 increases the performance of the model. Therefore, pixel values between 0 and 255 were scaled as 0–1 during normalization.

3.3 Feature extraction

A Convolutional Neural Network consists of multiple layers and convolutional layers, with input received in the input layer. The structure of the CNN model is shown in Fig. 3. As can be seen in the figure, the layers are the convolutional layer (conv), the maximum pool layer (max pool), the activation layer, and the flatten layer. The convolution layer creates the feature maps with the convolution filters contained in them. Feature maps are new images that highlight the unique features of the CFIs. Scaling is done by filtering the image. The size of the image is reduced in the pooling layer. This is done by converting certain adjacent pixels of the image to a maximum value.

Table 3 Details of the dataset

	Glaucoma Negative	Glaucoma Positive
Training set	247	316
Test set	62	80
Total	309	396

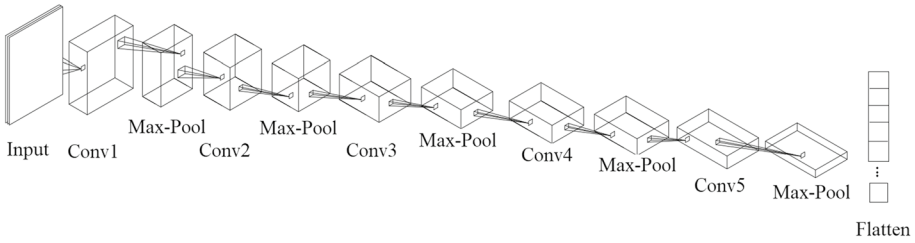

Fig. 3 CNN structure

The feature map created by the conv layer is processed by an activation function. Nonlinear features are introduced into the network. The result of the function is 1 for values greater than zero and 0 for other values [19]. The data obtained from the previous layers are transformed into a one-dimensional matrix in the flatten layer. The deep features that determine to which class the image belongs are obtained from this layer.

Detailed information about the layers of the CNN model can be found in Table 4. As shown in the Table 4, the proposed CNN model consists of 5 convolutional layers, 5 activation layers, and 5 max-pool layers after each convolutional layer. The first convolutional layer has 7×7 convolutional filters, and the other convolutional layers contain 3×3 convolutional filters. The RELU activation function (AF) was used as the activation function. Pooling takes place on 2×2 pixel windows.

3.4 Classification algorithm

3.4.1 Naive Bayes

Naive Bayes (NB) is a simple classification method that relates conditional probabilities [13]. In the NB classifier, attributes are considered independent of each other, given the class. It is a very fast algorithm compared to its counterparts. The conditional probabilities of the training data are calculated according to Eq. 1.

Table 4 Details of Layers Used in CNN

Layer	Filter Size	Pool Size	Activation	Data Depth	Stride	Padding	Number of Parameters
Input	–	–	–	3	–	–	0
Conv1	7×7	–	RELU	32	1	Same	4736
Maxpool	–	2×2	–	32	2		0
Conv2	3×3	–	RELU	64	1	Same	18,496
Maxpool	–	2×2	–	64	2		0
Conv3	3×3	–	RELU	128	1	Same	73,856
Maxpool	–	2×2	–	128	2		0
Conv4	3×3	–	RELU	512	1	Same	590,336
Maxpool	–	2×2	–	512	2		0
Conv5	3×3	–	RELU	1024	1	Same	4,719,616
Maxpool	–	2×2	–	1024	2		0
Flatten	–	–	–	1024	–	–	0

$$P(C_k|X) = \frac{P(C_k) * P(X|C_k)}{P(X)} \quad (1)$$

In Eq. 1, according to Bayes' theorem, a correlation is given between the probability of the presence of feature X under the condition of knowledge of class C_k and the probability of the presence of class C_k given knowledge of feature X. The class to which the test sample belongs is determined by the maximum conditional probability. Naive Bayes is a classification method widely used in fields such as pattern recognition and data mining.

3.4.2 Support vector machines

Support Vector Machines (SVM) is a powerful learning algorithm based on a very simple idea [14]. It is based on statistical learning and structural risk minimization. While SVM theoretically separates the examples belonging to the two classes, two narrow and parallel boundary lines are drawn in the two classes. The data are classified by drawing a common boundary line that is parallel to and equidistant from these two boundary lines. SVMs created using this idea are extremely successful classifiers.

When using the SVM algorithm, the hyperparameters were selected using the method of cross-validation by grid search. Accordingly, the classification process was performed by choosing the rbf kernel and the trade-off value 0.01.

3.4.3 Random Forest

The biggest problem with models that use Decision Trees (DT) is that overfitting occurs when data are scarce. Instead of creating a DT, Breiman [3] suggested using a model consisting of different decision trees, each using different data. The Random Forest (RF) algorithm is an algorithm that consists of a combination of DT. Among the decision trees used, the trees with the highest accuracy and independence are preferred.

When using the RF algorithm, the hyper-parameters were selected by the grid search cross-validation method. Accordingly, the number of estimators was set to 1000 and the maximum feature parameter was set to auto, and the classification process was performed.

3.4.4 K-nearest neighbors

The K-nearest Neighbor (kNN) algorithm is a simple algorithm that classifies data to be classified according to their proximity to previous data. The algorithm essentially considers the distance of the data to be classified from its neighbors and performs the classification using the most appropriate label. The algorithm, which uses a number of neighbors parameter to determine the closest class, predicts the results based on this number of neighbors. The number of neighbors is determined by how many nearest neighbors should be compared to the data [15]. The kNN classification algorithm selects the nearest neighbors according to the specified number of neighbors and decides that the unknown data belong to the class with the maximum number. Distance calculations between data are usually performed using 2 types of distance functions (Euclidean and Manhattan). When using the kNN algorithm, the hyperparameters were selected by a cross-validation method with grid search. Accordingly, the classification process was

performed by choosing the Euclidean distance function and the number of 17 neighbors. The Euclidean distance function is calculated as shown in Eq. 2.

$$d(j,i) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \tag{2}$$

3.4.5 Adaptative boosting

Boosting algorithms are machine learning algorithms called ensemble methods. Ensemble method algorithms are the creation of a stronger learning algorithm by combining many weak learning algorithms. Fast-derived versions of decision trees are chosen as weak algorithms because they are usually easy to construct. Boosting is similar to ensemble method algorithms using other weak classifiers. Similarly, a strong classifier is formed by combining many weak classifiers. But during this creation process, each weak classifier is assigned a weight so that the model can better classify the samples that it previously misclassified. Therefore, the weights of the learned models vary from one sample to the next. As the iterations progress, the hard-to-predict samples gain more and more influence. Thus, the model is forced to focus on examples that were overlooked in earlier iterations. Adaptive Boosting (AdaBoost) is one of the most commonly used boosting algorithms [10].

When using the AdaBoost algorithm, the hyper-parameters were selected by the cross-validation method using grid search. Accordingly, the number of estimators was set to 1000 and the learning rate was set to 0.1, and the classification process was performed.

3.4.6 Multilayer perceptron

The Multilayer Perceptron (MLP) is the most widely used artificial neural network model with a simple structure consisting of an input, an output, and one or more hidden layers. Unlike the creation of the input and output layers, the process begins without prior knowledge of the number of hidden layers when the network architecture is executed. The number of hidden layers in the MLP network architecture in this study is two and was determined by the trial-and-error method. When the network architecture was created, the number of neurons in the layers was determined using the trial-and-error method [6]. As a result of the tests performed, it was found that the input neurons of the network architecture were 512, the neurons of the 1st hidden layer were 256, the neurons of the 2nd hidden layer were 128, and the number of output neurons was 1. A 0.5 dropout procedure was applied between each layer. Relu AF is used in the input layer and the sigmoid AF is used in the output layer. In the MLP, each k-index neuron in the hidden layer adds after multiplying the i-index x input by the weight of the connection wki and calculates the k-index y output as a function of the sum according to Eq. 3.

$$y_k = f\left(\sum_{i=1}^{n} w_{ki} x_i\right) \tag{3}$$

4 Experimental results

In this study, 309 normal-class fundus images and 396 glaucoma-class fundus images from the ACRIMA dataset were used. Experimental studies were performed on a computer with NVIDIAÂ® GeForce graphics card, Intel(R) Core(TM) i7-8550U 1.8 GHz CPU, and

16 GB memory. The models were developed with the Python programming language using the Keras library.

In this study, the classification was performed using the metrics of Accuracy, Sensitivity, Specificity, Precision, and F1-score, and the performance of the models used was evaluated. These metrics, given in Eqs. 4–8, are obtained from the confusion matrix.

True Positives (TP) are values predicted to have glaucoma and actually have glaucoma; False Negatives (FN) are values predicted to have normal glaucoma but actually have glaucoma; False Positives (FP) are values predicted to have glaucoma but actually have normal glaucoma; whereas True Negatives (TN) denote values estimated to be normal and actually have normal glaucoma.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$F1 - score = \frac{Precision * Sensitivity}{Precision + Sensitivity} * 2 \tag{8}$$

In addition, the Area Under Curve - Receiver Operating Characteristic (AUC-ROC) curve, one of the popular metrics, has also been used to evaluate classification success. ROC Curve is a graph that shows the true-positive and false-positive rates. The area under the ROC curve is expressed as the Area Under the Curve (AUC). The larger the AUC, the better the model is at predicting disease.

The confusion matrix created to evaluate the performance of the hybrid models is shown in Fig. 4. As can be seen in Fig. 4, the models are generally better at detecting images with glaucoma than normal images. The values obtained by all models for accuracy, precision, sensitivity, specificity, AUC value, and F1-score are shown in Table 5. Looking at the sensitivity value in Table 5, we find that CNN and kNN achieve the highest value (97.5%). But the specificity value (59.68%) was too low, so normal images could not be detected. The model that recognised normal images best was the hybrid model of CNN and AdaBoost. However, it showed the same success in recognising images with glaucoma. The hybrid model of CNN and AdaBoost had the highest accuracy value of 92.96%. The hybrid model of CNN and NB provided the lowest accuracy value of 80%. The results of the experimental studies show that CNN and Adaboost achieve the highest values in the performance metrics of F1-score, specificity, precision AUC, and accuracy. As can be seen in Table 5, it can be said that CNN-based hybrid models (except CNN + NB and CNN + kNN hybrid models) outperform the basic CNN model.

The success of all models used in the study was also evaluated with (AUC-ROC) curves. ROC Curve and AUC values are shown in Fig. 5. As can be seen in the figure, the model that makes the best distinction between classes is the hybrid model of CNN and Adaboost.

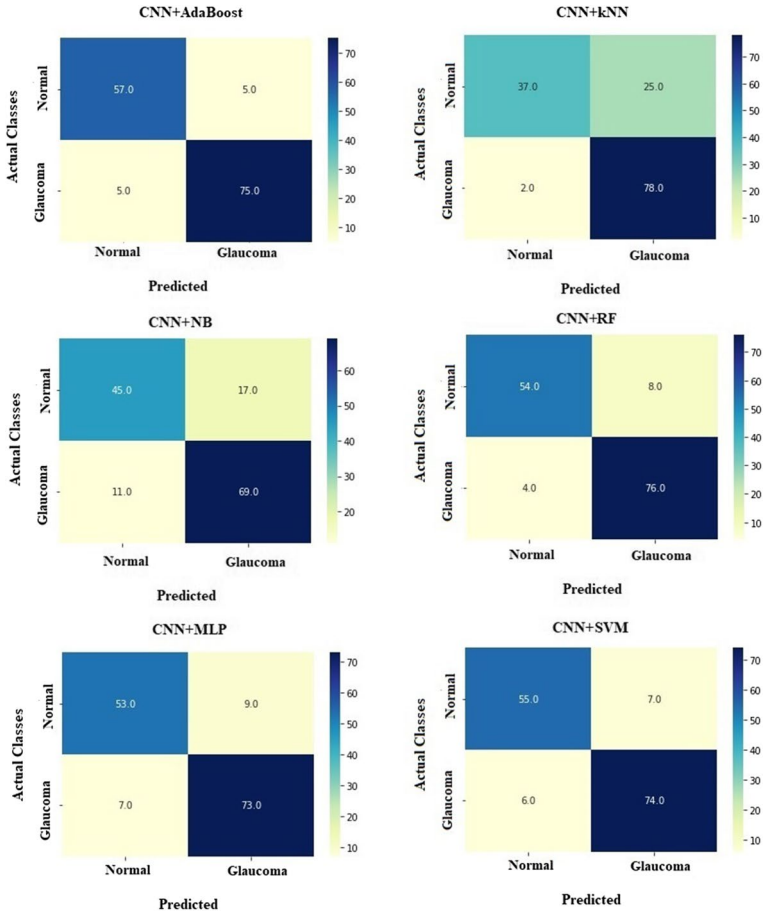

Fig. 4 Confusion matrixs obtained with hybrid models

The plots of accuracy and losses of Basic CNN, VGG16, VGG19, MobileNetV2, DenseNet121 and ResNet50 are shown in Figs. 6, 7, 8, 9, 10, and 11 respectively. The plots show the decrease in the loss rate and the increase in the accuracy rate. They show that the training process and learning of the model have a good learning rate. While the loss rate of the models decreases in each epoch of the training process, the accuracy increases and learning takes place with the given training set. The training was stopped when the performance increased in very small units in each model.

A comparison of the proposed model with 5 common CNN models can be found in Table 6. As can be seen in the table, Vgg16, Vgg19, DenseNet121, and ResNet50 performed better than the recommended model in the sensitivity value. However, for the other success criteria, the proposed CNN-based hybrid model scored higher values than the well-known CNN models.

Table 5 Success Criteria of the Hybrid Models

Model	Accuracy(%)	Specificity(%)	Sensitivity(%)	Precision(%)	F1-score(%)	AUC
CNN + SVM	90.8	88.71	92.5	91.36	91.93	0.906
CNN + kNN	80.99	59.68	97.5	75.73	85.25	0.786
CNN + RF	91.55	87.1	95	90.48	92.68	0.91
CNN + AdaBoost	92.96	91.94	93.75	93.75	93.75	0.928
CNN + MLP	88.73	85.48	91.25	89.02	90.12	0.884
CNN + NB	80.28	72.58	86.25	80.23	83.13	0.794
CNN	86.62	82.26	90	86.75	88.35	0.869

5 Discussion

In this study, several hybrid classification models integrating Deep Learning (CNN) and traditional Machine Learning classifiers (SVM, kNN, RF, Adaboost, MLP, NB) were used to automatically classify glaucoma and healthy cases from CFIs.To compare the success of the proposed model with the classification methods used in previous studies, the dataset used, the number of images, and the success rate are shown in Table 7. As can be seen in the table, the proposed model performed better than the studies [9, 25] that used the same data set (ACRIMA). In their study, Diaz Pinto et al. [9] applied five DL models to five different datasets. In the ACRIMA dataset, the Xception model achieved the highest performance with an accuracy of 70.21%. Serte and Serener [25] evaluated 3 different DL models on 5 different datasets. On the ACRIMA dataset, the GoogLeNet model achieved the highest performance with 65% accuracy. The hybrid model of CNN and Adaboost proposed in this study achieved 92.96% accuracy.

Patil et al. [21] achieved 98% accuracy in their study. However, considering the sensitivity (83%), it could not show the same success in classifying glaucoma images. The reason for this can be seen in the imbalance of the data set. 280 images with glaucoma and 1058 normal images might have resulted in the model learning better from normal images.

The proposed model obtained better results than the methods presented in other studies [1, 2, 5, 7, 21] with different datasets in the literature. It is clear that studies [11, 17, 22, 29] with a high number of images in the dataset achieve high performance. Collecting lots of data is often helpful, but this can be a very costly activity [24, 30]. Roccetti et al. [24]

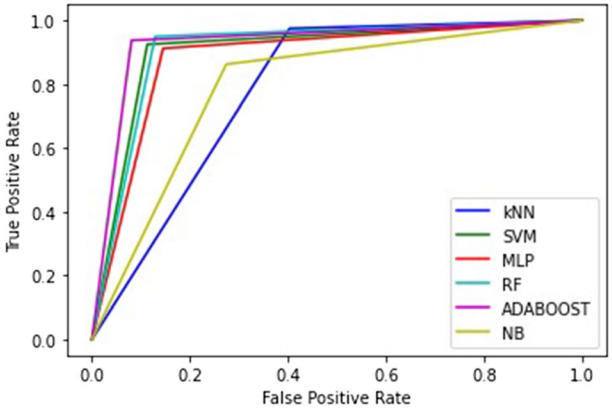

Fig. 5 Roc curve of hybrid models

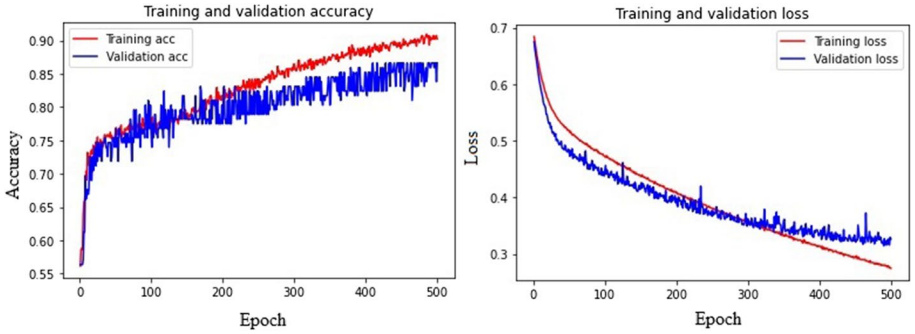

Fig. 6 Accuracy and loss graph of CNN

showed that using good data instead of having a large number of data yields better results. In many industries, such as the healthcare industry, where datasets with a large amount of data do not exist, the focus needs to shift from big data to quality data. It is thought that better results will be obtained by paying special attention to the quality of the data used. In our future work, we intend to produce a model with higher performance and less cost by improving the data quality in the data preprocessing step (such as noise removal).

6 Conclusion

Glaucoma is a disease that is difficult to detect. If glaucoma is diagnosed early, vision loss can be prevented with regular exams and treatment. If diagnosed too late, the disease can cause severe damage to the optic nerve that cannot be repaired, leading to loss of central vision and blindness. Therefore, early diagnosis of the disease is critical.

Most automated computer-based systems developed for glaucoma detection have proposed Deep Learning-based architectures based on manually created segmentation features in fundus images. In this study, a hybrid classification technique is proposed to assist experts in glaucoma disease detection with a model that integrates Deep Learning and traditional ML classifications using raw fundus images.

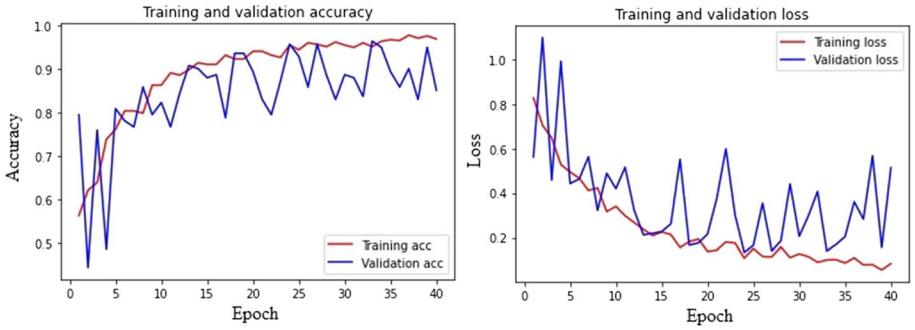

Fig. 7 Accuracy and loss graph of Vgg16

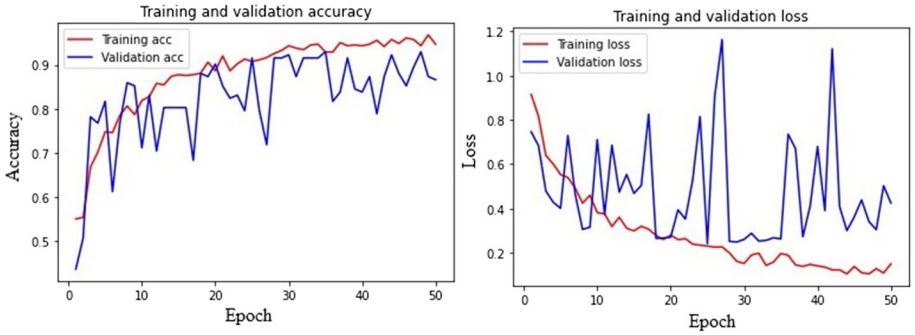

Fig. 8 Accuracy and loss graph of Vgg19

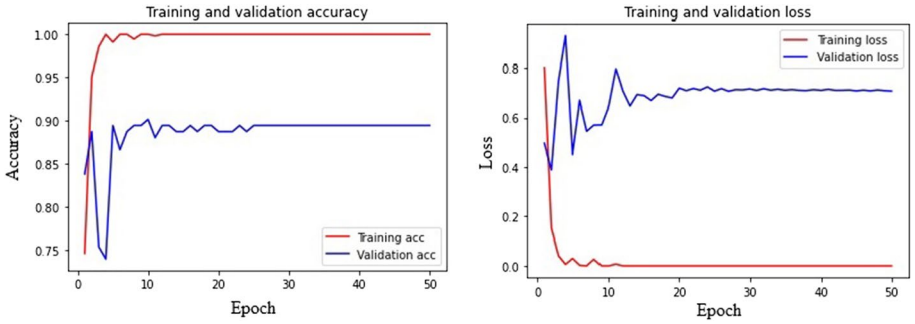

Fig. 9 Accuracy and loss graph of MobileNetV2

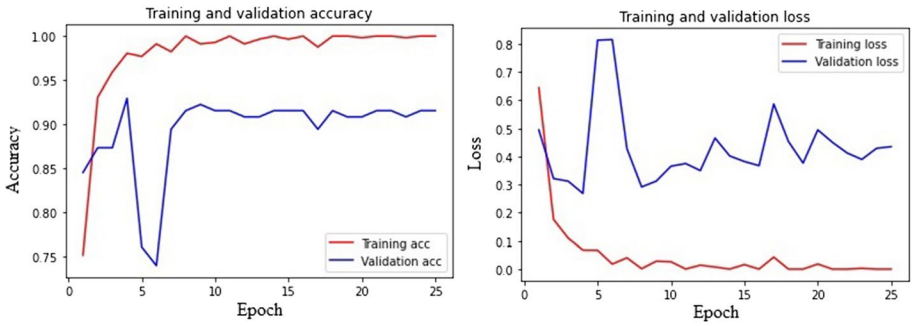

Fig. 10 Accuracy and loss graph of DenseNet121

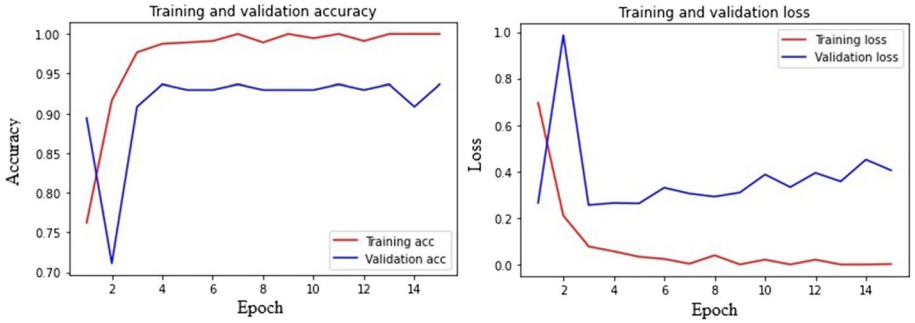

Fig. 11 Accuracy and loss graph of ResNet50

The proposed hybrid model achieved higher performance than studies in the literature using the same data set. A new hybrid model was obtained by combining the traditional ML structure and the DL structure. While feature extraction was made from unstructured data with DL structure, the classification process was performed using these attributes with traditional ML structure. In this way, cost and training time were reduced. The proposed hybrid model avoids high computational resources and long training times by using fewer parameters than end-to-end CNN, Transfer Learning and RNN-based methods. The proposed Hybrid model achieved better results than the end-to-end CNN model and state-of-the-art transfer learning models, with less cost, shorter training time, and high accuracy. The proposed model may give better results with a large number of data. However, this situation creates a problem in many sectors such as the health sector, where data sets containing large amounts of data are not available. Collecting large amounts of data is also very costly. That's why the focus needs to shift from big data to quality data. It is thought that better results will be obtained by paying special attention to the quality of the data used. In our future work, we aim to produce a higher performance and less costly model by

Table 6 Comparative result with popular CNN models

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	Precision (%)	F1-score (%)
Vgg16	85.21	66.13	100	79.21	88.4
Vgg19	82.39	59.68	100	76.19	86.49
MobileNetV2	89.44	83.87	93.75	88.24	90.91
DenseNet121	91.55	85.48	96.25	89.53	92.77
ResNet50	92.25	83.87	98.75	88.76	93.49
Proposed Model	92.96	91.94	93.75	93.75	93.75

Table 7 Comparison of Glaucoma Detection Studies

Works	Method	Dataset	Number of images	Performance
[7]	6 layer-CNN	ORIGA SCES	650 1676	0.831 0.887 AUC
[1]	FNN	University of Tokyo Hospital	195	92.5% accuracy
[22]	18 layer-CNN	Kasturba Medical College	1426	98.13% accuracy
[5]	MB-NN		2554	91.51% accuracy
[9]	Xception	ACRIMA HRF Drishti-GS1 RIM-ONE sjchoi86-HRF	705 45 101 455 401	70.21% accuracy 80% accuracy 75.25% accuracy 71.21% accuracy 70.82% accuracy
[25]	GoogLeNet	ACRIMA HRF Drishti-GS1 RIM-ONE sjchoi86-HRF	705 45 101 455 401	65% accuracy 76% accuracy 55% accuracy 70% accuracy 72% accuracy
[2]	Two-step DL	ORIGA	650	0.874 AUC
[21]	CNN	PRV-Glaucoma sjchoi86-HRF	1338	98% accuracy 83% sensitivity
[11]	CNN and RNN	University of New South Wales	1810CFIs 295fundus videos	96.2% F1-score
[29]	Community-based CNN	LAG RIM-ONE	11,760 455	99.51% accuracy 93% accuracy
[17]	CoG-NET	RIM-One Drishti REFUGE ACRIMA	2172	93.5% accuracy
This study	CNN + AdaBoost	ACRIMA	705	93% accuracy 93.75% sensitivity

improving data quality in the data preprocessing step (such as noise removal). This can be counted as the limits of the proposed model.

Future studies will aim to achieve more successful glaucoma disease detection results by developing hybrid features and optimizing the features with a model such as Deep Belief Networks. In addition, our future study will investigate the fusion of OCT and fundus images into the dataset.

Data availability The full version of dataset is available at https://figshare.com/articles/dataset/CNNs_for_Automatic_Glaucoma_Assessment_using_Fundus_Images_An_Extensive_Validation/7613135. The database consists of images from the ACRIMA project to develop automatic classification algorithms for retinal images.

Declarations

Conflict of interest The author has declared that no conflict of interests or competing interests exist.

References

1. Asaoka R, Murata H, Iwase A, Araie M (2016) Detecting Preperimetric Glaucoma with standard automated Perimetry using a deep learning classifier. Ophthalmol 123(9):1974–1980
2. Bajwa MN, Malik MI, Siddiqui SA, Dengel A, Shafait F, Neumeier W, Ahmed S (2019) Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. MC Med Inf Decis Making 19(1):1–16
3. Breiman L (2001) Random Forests Mach Learn 45(1):5–32
4. Burgansky-Eliash Z, Wollstein G, Chu T, Ramsey JD, Glymour C, Noecker RJ, Ishikawa H, Schuman JS (2005) Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. Invest Ophthalmol Vis Sci 46(11):4147–4152
5. Chai Y, Liu H, Jie X (2018) Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. Knowl-Based Syst 161:147–156
6. Chaudhuri BB, Bhattacharya U (2000) Efficient training and improved performance of multilayer perceptron in pattern classification. Neurocomputing 34:11–27
7. Chen X, Xu Y, Wong DWK, Wong TY, Liu J (2015) Glaucoma detection based on deep convolutional neural network. In 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp. 715–718
8. Devalla SK, Chin KS, Mari J-M, Tun TA, Strouthidis NG, Aung T, Thiéry AH, Girard MJA (2018) A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. Invest Ophthalmol Vis Sci 59(1):63–74
9. Diaz-Pinto A, Morales S, Naranjo V, Köhler T, Mossi JM, Navea A (2019) CNNs for automatic glaucoma assessment using fundus images: an extensive validation. Biomed Eng Online 18(1):1–19
10. Freund Y, Schapire RE (1999) A short introduction to boosting. J-Japan Soc Artif Intell 14:1612
11. Gheisari S, Shariflou S, Phu J, Kennedy PJ, Agar A, Kalloniatis M, Mojtaba Golzan S (2021) A combined convolutional and recurrent neural network for enhanced glaucoma detection. Sci Rep 11(1):1–11
12. Hacisoftaoglu RE, Karakaya M, Sallam AB (2020) Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems. Pattern Recogn Lett 135:409–417
13. Hand DJ, Keming Y (2001) Idiot's Bayes-not so stupid after all? Int Stat Rev 69(3):385–398
14. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intell Syst Appl 13(4):18–28
15. Jiang Y, Zhou Z-H (2004) Editing training data for kNN classifiers with neural network ensemble. International symposium on neural networks, 356–361
16. Juneja M, Minhas JS, Singla N, Thakur S, Thakur N, Jindal P (2022) Fused framework for glaucoma diagnosis using optical coherence tomography (OCT) images. Expert Syst Appl 201:117202
17. Juneja M, Thakur S, Uniyal A, Wani A, Thakur N, Jinda P (2022) Deep learning-based classification network for glaucoma in retinal images. Comput Electr Eng 101:108009
18. Koh V, Tham Y-C, Cheung CY, Mani B, Wong TY, Aung T, Cheng C-Y (2018) Diagnostic accuracy of macular ganglion cell-inner plexiform layer thickness for glaucoma detection in a population-based study: comparison with optic nerve head imaging parameters. PLoS One 13(6):e0199134
19. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst, 25
20. Padmanayana, Anoop BK (2022) Binary classification of DR-diabetic retinopathy using CNN with fundus colour images. Mater Today: Proceed 58:212–216
21. Patil N, Patil PN, Rao PV (2021) Convolution neural network and deep-belief network (DBN) based automatic detection and diagnosis of Glaucoma. Multimed Tools Appl 80(19):29481–29495
22. Raghavendra U, Fujita H, Bhandary SV, Gudigar A, Hong TJ, Rajendra Acharya U (2018) Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. Inf Sci 441:41–49
23. Raja H , Akram MU , Hassan T , Ramzan A , Aziz A, Raja H (2022) Glaucoma detection using optical coherence tomography images: a systematic review of clinical and automated studies. IETE J Res, 1–21
24. Roccetti M, Delnevo G, Casini L, Cappiello G (2019) Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. J Big Data 6(1):1–23
25. Serte S, Serener A (2015) A generalized deep learning model for glaucoma detection. 2019 3rd International symposium on multidisciplinary studies and innovative technologies (ISMSIT), pp. 1–5

26. Shinoj VK, Hong XJJ, Murukeshan VM, Baskaran M, Tin A (2016) Progress in anterior chamber angle imaging for glaucoma risk prediction–a review on clinical equipment, practice and research edical. Eng Phys 38(12):1383–1391
27. Shoukat A, Akbar S (2021) Artificial intelligence techniques for glaucoma detection through retinal images: state of the art. Artif Intell Int Things, 209–240
28. Singh LK, Garg H, Khanna M (2021) An artificial intelligence-based smart system for early glaucoma recognition using OCT images. Int J E-Health Med Commun (IJEHMC) 12(4):32–59
29. Stalin David D (2021) Enhanced glaucoma detection using ensemble based CNN and spatially based ellipse fitting curve model. J Ambient Intell Human Comput, 1–12
30. Strickland E (2022) Andrew Ng: Unbiggen AI-IEEE Spectrum. https://spectrum.ieee.org/andrew-ng-data-centric-ai. Accessed 22 May 2022
31. Sunija AP, Varun P, Gopi, Palanisamy P (2022) Redundancy reduced depthwise separable convolution for glaucoma classification using OCT images. Biomed Signal Process Control 71:103192
32. Zilly J, Buhmann JM, Mahapatra D (2017) Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation, 55, 28–41