# Bagging of Convolutional Neural Networks for Diagnostic of Eye Diseases

Mahmoud Smaida[1][0000-0002-5552-2768], Serhii Yaroshchak [2][0000-0001-9576-2929]

The national university of water and Environmental Engineering, Revine, Ukraine
Smaida20012001@gmail.com
The national university of water and Environmental Engineering, Revine, Ukraine
s.v.yaroshchak@nuwm.edu.ua

**Abstract**- Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the structure of the human brain itself, learn from large amounts of data. In this paper, we will introduce the part of the techniques of deep learning to perform multi-class classification, in order to classify eye diseases. One of the biggest issues in image recognition is the classification of medical images, and it aims to classify medical images into different categories to help doctors diagnose the disease. But the most important idea will be addressed in our paper is the evaluation performance model using a bagging ensemble. In this study, we will compare three models of the convolutional neural network, CNN, Vgg16 and InceptionV3 in order to evaluate the performance of the models using bagging ensemble.
In our work, a deep learning convolutional network based on Keras and Tensor Flow is deployed using python for image classification. A number of different medical images have been used as a data set to diagnose eye diseases, which contain four types of diseases such as, Diabetic retinopathy, Glaucoma, Myopia and Normal. CNN, VGG16 and InceptionV3 neural network structures are compared singly and together using bagging ensemble, in order to diagnose eye diseases. All experiments were applied and the result was obtained. It has been shown that using a bagging ensemble yields better predictive efficiency than can be obtained using learning algorithms alone. Moreover, the use of the confusion matrix in our experiments shows us where our classifiers are confused when it makes predictions.

**Keywords.** InceptionV3, Vgg16, eye diseases, ensemble learning, Deep Learning, Diabetic retinopathy, Glaucoma, Myopia, bagging.

## 1 Introduction.

Diabetic retinopathy, glaucoma and myopia are some of the most common eye diseases and one of the most common causes of blindness in the world if they are not detected at an early stage.

In recent years, the diagnosis of diseases of the human visual system has advanced greatly to technological innovations and developments in the field of artificial intelligence. Taking into account the diversity and complexity of eye functions, a large

number of diagnostic equipment, tools, methods and algorithms have been developed. Sometimes a doctor can discover a specific disease after a visual analysis of the image. However, in a large number of cases, the diagnosis is not made due to many factors, such as bad experience, fatigue, a variety of shapes, similarities, poor image quality, etc. In these cases, the second opinion is very important and useful, which comes from another expert who uses advanced information technology and algorithms to accurately analyze the image to diagnose eye diseases [1] bagging ensemble is a kind of ensemble learning, it is a set of machine learning models combined together to obtain better results. In this study, we focus on bagging ensemble to improve the model's prediction and make it better. We are not talking about creating a new algorithm, but instead assembling together several different algorithms or several different models to create an ensemble learner, called bagging, in order to increase the accuracy of the model. In general, predicting the target variable using any deep learning method leads to a difference between actual and expected values, due to noise, variance and bias. The Bagging ensemble helps reduce variance. In summary, as shown in Figure 1, different and same algorithms are used in ensemble learning to achieve a better prediction efficiency that can be achieved from any of the constituent learning algorithms alone [2].

## 2 Formal problem statement

Eye diseases have a wide range of shapes, sometimes the textures are difficult to identify and recognize by an ophthalmologist. Therefore, information technology must be used to provide maximum comfort to the patient and ophthalmologist, and improve health care system.

In this paper, we will use bagging ensemble to evaluate three different CNN structures to identify eye diseases, Diabetic retinopathy, Glaucoma, Myopia, and Normal.

## 3 Ensemble learning

Ensemble learning models are a technique that combines several base models to create a perfect predictive model, and it is divided into two groups: Simple ensemble Techniques and advanced ensemble techniques [3].

### 3.1 Simple ensemble Techniques [3]:

a. Max Voting: each model in max voting makes a prediction and votes for each sample. The category with the most votes will be the last predictive category.

b. Averaging: It is the process of creating many models and combining them to get the desired result. The result will be better average performance than single model.

$$\tilde{y}(\mathbf{x};\alpha) = \sum_{j=1}^{p} \alpha_j y_j(\mathbf{x}) \tag{1}$$

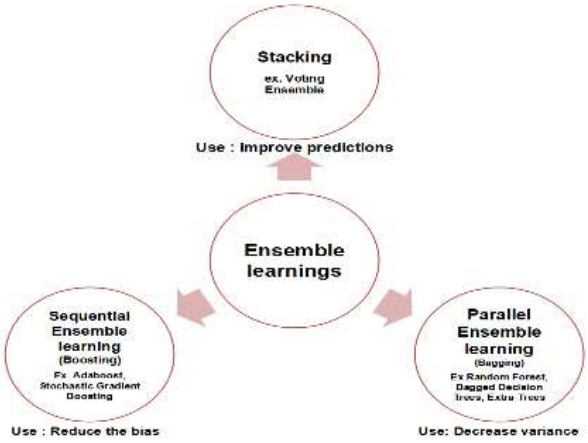


**Fig. 1.** Types of ensemble learning [5]

c. Weighted Averaging: is an extension of a model averaging ensemble where the contribution of each member to the final evaluated is weighted by the performance of the model. A weighted average mean value takes the form of a sum on quantum energy states, rather than continuous integration.

$$\bar{A} = \frac{\sum_i A_i e^{-\beta E_i}}{\sum_i e^{-\beta E_i}} \tag{2}$$

### 3.2 Advanced ensemble techniques [4], [15]:

a. Bagging: type of ensemble learning that relies on creating a number of sub-datasets called bagging.

b. Boosting: Is a fairly simple variation on bagging that strives to improve the learners by focusing on areas where the system is not performing well.

c. Stacking: in stacking all models are trained based on a complete data set, and the output used as input features to train ensemble function.

## 4 Bagging:

In this paper we will focus on bagging ensemble which is a type of ensemble learning that relies on creating a number of sub-datasets called bagging; each bag is a subset of the original dataset, which contains a number of different instances. Inside each bag a set of instances of random data with replacement. We use each of these collections of data (bag) to train a different model. Finally, collect all of the outputs (predicts) and calculate the average or voting value as shown in fig.2.

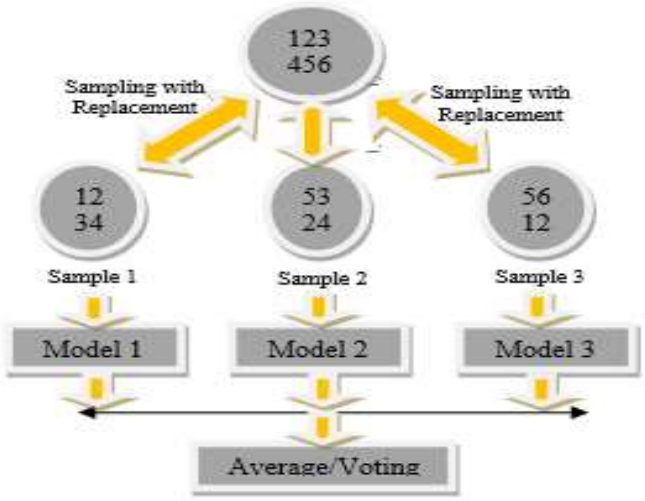


**Fig. 2.** Bagging example has three steps

Bagging ensemble learning model is category bases on its use; the details are addressed in Table.1

**Table 1.** bagging ensemble learning basis of it's use [5]:

| | Partitioning of the data into subsets | Goal to achieve | Methods where this is used | Function to combine single models |
|---|---|---|---|---|
| **Bagging** | Random | Minimize variance | Random subspace | (weighted) average |

## 5 Related work

Many researchers have used bagging ensemble techniques using neural networks in their research, and most of these studies have been done recently, focusing on recent research. A few reviews are as follows:

Ju, Cheng, Aurélien Bibaut and Mark van der Laan. [6] In this work, authors in this work are used neural network, VGG, GoogleNet and ResNet to applied some of ensemble learning technical, including: Max voting, unweighted average, bayes optimal and super learner. the authors trained their models based on same and different Networks. Ensemble of the same and different networks has been trained multiple times. the results obtained and listed based on the best performance on the testing set. all learners used CIFAR 10 as a dataset, and the unweighted average provided the best result when the performance of the base learners is comparable.

Huang, Jonathan, et al. [7] four deep neural networks has been applied in this work in order to improve the accuracy. These networks are, Vgg12, ResNet50, AclNet and AclSincNet, all these models were pre-trained with audio dataset. Ensemble learning was achieved in all these models and the result obtained over the validation set. The best accuracies were achieved when all the networks combined together based on ensemble average by score 83.01%.

Mo, Weilong, et al. [8] the authors suggest an image recognition algorithm based on the ensemble learning algorithm and the structure of the ELA-CNN to solve a problem that single model can not correctly predict. They used the bagging ensemble to train their models. the networks structure was used are combines of ResNet, DenseNet, DenseNet-BC and Inception-Resnet-v2 architecture. in their experiments they used cifar-10 as images dataset, it consists of 60,000 color images. These images were divided into 50,000 in the training set and 10,000 in the test set. The final result was the average probability of the prediction vector.

Kumar, Ashnil, et al. [9] For classification of medical images based on diagnosis, training, and biomedical research, a set of convolutional neuronal networks of fine-tuned were used to classify medical images. They used 6,776 training images and 4166 test images. The authors used two different CNN designs, AlexNet and GoogleNet, to images classification. The experiments were performed using individual models and ensemble models. By the end result, the ensemble method reached an accuracy corresponding to the best accuracy among other methods of the overall method of 96.59%.

Beluch, William H., et al. [10] the authors in this paper explore some of the recently proposed active learning methods that contain big data and CNN classifiers. They compare ensemble-based methods against Monte-Carlo Dropout and geometric approaches.. They have found that the ensemble learning better and leads to a more predictable uncertainty, which is the basis of many active training algorithms of convolution Neural Networks, such as S-CNN, K-CNN, DenseNet, InceptionV3 and ResNet -50 to classify Diabetic retinopathy. The dataset was used with MNIST, CIFAR, and ImageNet. They found that ensembles which based on several active learning algorithms were better predicted and achieved a set test accuracy of 90% of the approximately 12,200 images presented.

Minetto, Rodrigo, Maurício Pamplona Segundo, and Sudeep Sarkar. [11] Hydra: An Ensemble of Convolutional Neural Networks for Geospatial Land Classification in satellite image. Hydra is an initial CNN that is coarsely optimized, which will serve as the Hydra's body. in this article, authors created ensembles for their experiments using two state-of-the-art CNN architectures, ResNet and DenseNet. they demonstrated

their application of Hydra framework in two datasets, FMOW and NWPU-RESISC45. The final result ensemble was achieved accuracy around 94.51%.

## 6 Data Description

Kaggle is a data science website that contains a variety of interesting data sets. In its main menu, you can find all kinds of specialized data sets, from the Ramen classifications to basketball and animal licenses data in Seattle [11].

We used our data from competition in kaggle Diabetic Retinopathy Detection [17] and iChallenge-GON Comprehension which is a large collection of 1,200 retinal fundus images for both subjects without glaucoma (90%) and glaucoma patients (10%).

The data set includes more than 35 types of eye diseases. To simplify, we will reduce the data set with 4 main breeds. The dataset includes images of glaucoma, myopia, diabetic retinopathy, and Normal eye provided as a subset of photos from a large dataset of 2781 Retinal Image as it shown in table.2. All the images were collected in total from Kaggle dataset and iChallenge-GON Comprehension, in high resolution images.

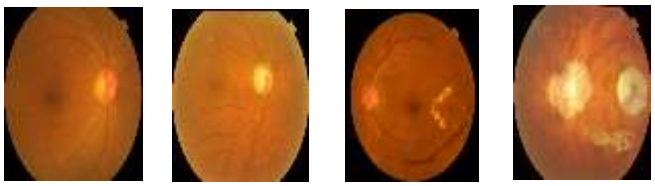

**Fig. 3.** Normal Fundus, Glaucoma, Diabetic retinopathy and Myopia

Images will be the entrance to CNN architecture. our images divided into training dataset, validation and test dataset; each type of image has a separate folder and each image has a file name, which is its unique identifier. Python will be used to achieve our goal using Google Colab.

**Table 2.** Number of images according to eye diseases

| Diabetic retinopathy | Glaucoma | Myopia | Normal eye |
|---|---|---|---|
| 975 | 721 | 486 | 599 |

## 7 Research Methodology

The three proposed methodologies will be used in our experiment are CNN, VGG16 and InceptionV3 in order to evaluate singly and using bagging ensembles to identify eye diseases. First, a set of image data is prepared step-by-step; there are 4 folders in the data set, which contain 2781 images of diabetic retinopathy, glaucoma, myopia and normal, where 1951 images were used for training, 415 images were used for tests, and 415 images were used for validation. In the next steps, fitting our CNN model, then, obtain the accuracy of the data set for different CNN structures and final-

ly, compare these accuracies separately and using bagging ensemble to measure performance.

This article covers three ways to evaluate the performance of our learners:

- CNN based on three hidden layers, pooling layers and fully connected layers.
- Pre-trained CNN based on VGG 16 algorithms using the last block layer training (Block 5).
- Pre-trained CNN based Inception v3 algorithms using the last block layer training ('mixed6).

### 7.1 Convolution Neural Network

The size of the input image is 150 * 150 pixels with 3 channels (RGB). To extract the image features, we used 32 filters 3 * 3 pixels. And 2 * 2 pixel window, used to minimize the size of image(Pooling layer). Next, we applied another convolution layer used 32 filters with a size of 3 * 3 and a max pooling size of 2 * 2. In the last convolution layers, 64 filters of 3 * 3 are used with a max pooling of 2 * 2., then we use the Fully connected layer (64 dense units) and softmax layer (4 units) to predict eye diseases. CNN networks adjust the weight of the filters during the back propagation, which means that after forwarding, the network can look at the loss function and carry out the backward transfer process to update the weight.

| Function | Formula | Derivative |
|---|---|---|
| Weighted input | $Z = XW$ | $Z'(X) = W$<br>$Z'(W) = X$ |
| ReLU activation | $R = max(0, Z)$ | $R'(Z) = \begin{Bmatrix} 0 & Z < 0 \\ 1 & Z > 0 \end{Bmatrix}$ |
| Cost function | $C = \frac{1}{2}(\hat{y} - y)^2$ | $C'(\hat{y}) = (\hat{y} - y)$ |

(3)

$$^{*}W_x = W_x - a\left(\frac{\partial Error}{\partial W_x}\right) \tag{4}$$

New weight — Old weight — Learning rate — Derivative of Error with respect to weight

### 7.2 VGG 16.

This is a convolutional neural network structure developed by the University of Oxford's Visual Engineering group in 2014. This model loads a set of pre-trained weights into ImageNet using a 16-layer network.

The size of the images entered on the VGG16 network is 224x224 RGB, the images are passed through 5 blocks of convolutional layers, with each block consisting of an increasing number of 3x3 filters, stride is fixed to 1 while the convolutional layer inputs are padded. The blocks are separated by the max pooling layers. The max pooling is made on 2 * 2 windows with stride 2. Five blocks of convolutional layers are followed by three fully connected (FC) layers. The last layer is a soft max layer representing the output layer [12].

### 7.3 Inception V3.

Is a convolutional neural network consisting of 48 deep layers trained in over a million images in an ImageNet database. It can categorize images into 1000 categories of objects [13], [14].

Inception-v3 is one of the most popular models that can be used for transfer learning. This allows us to retrain the last layers of existing models, which leads to a significant reduction in training time. inception-v3 has been trained in over a million images from the ImageNet database, which means that the model had learned during its original training and could be applied to smaller dataset with highly accurate classifications without the need of training all the model.

The Inception Layers is a mixture of a set of layers (i.e. 1 × 1 convolutional layer, 3 × 3 convolutional layer, 5 × 5 convolutional layer) with combinations of output filters combined into one output vector, forming the inputs for the next step.

### 7.4 Selected Measures.

In this section, we officially describe the most common measures used to compare our works. The various measures are based on the marginal rate of the confusion matrix. In this article, comparisons will be made using confusion matrix to measure model accuracy. Accuracy: This is a measure of how much the classifier predicted the class correctly.

$$\text{Accuracy} = \quad TP / TP+FP+FN+TN \tag{5}$$

## 8 Experiments and Results.

All the models above are applied using Python; the dataset is a set of fundus images representing eye diseases, such as diabetic retinopathy, glaucoma, myopia and normal. In our experiments, we compare the empirical performance for bagging ensemble method which we mentioned before to obtain different models to get better accuracy for eye diseases detection.

### 8.1 Results on CNN, VGG16 and InceptionV3 individually:

CNN, VGG16 with fine-tune the final layers and InceptionV3 with pre-training the final layers uses eye diseases dataset including bagging have been applied. Table.3 shows the result on the test dataset.

**Table 3.** Prediction accuracy in individual models

| Model | Number of Epoch | Prediction Accuracy |
|---|---|---|
| **CNN** | 50 | 71.57% |
| **VGG16** | 50 | 83.86 % |
| **InceptionV3** | 50 | 87.71 |

From the above models, there are three classification accuracies obtained as shown in Table.3. These accuracies are graphically represented in the graphs below, where each model structure is shown with epochs and accuracies.

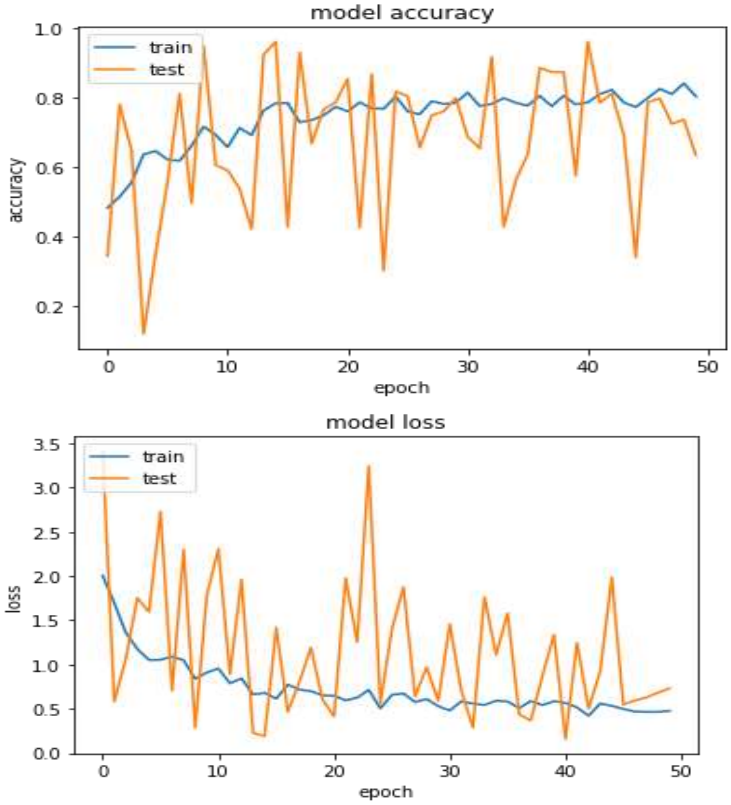


**Fig. 4.** Accuracy on train and test set in CNN

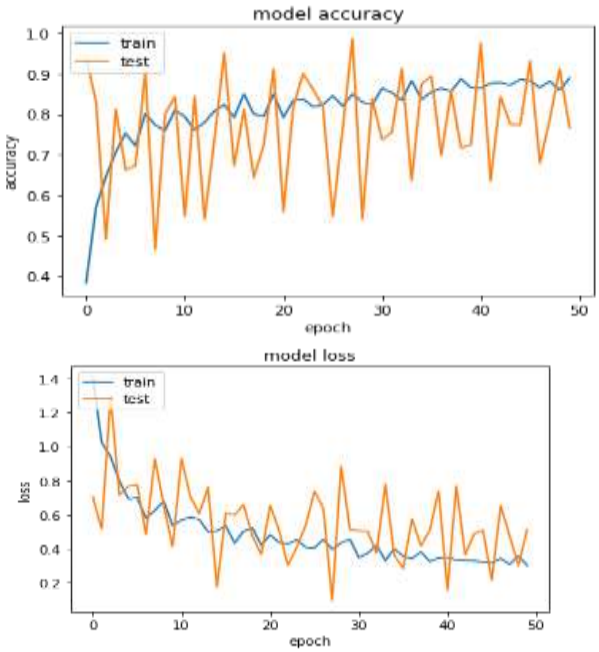


**Fig. 5.** Accuracy on train and test set in VGG architecture

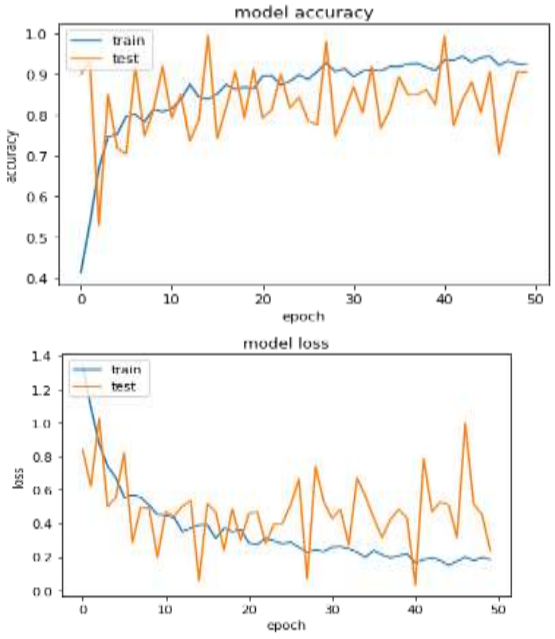


**Fig. 6.** Accuracy on train and test set in InceptionV3 architecture

### 8.2 Results on CNN, VGG16 and InceptionV3 using bagging ensemble learning:.

Bagging ensembles with different and the same structures have been applied. Three models were trained by CNN, VGG16 and InceptionV3 to implement a bagging ensemble. Therefore, compare the performance for all the bagging ensemble methods, and the results presented in Table.4 of each net on the test set.

**Table 4.** Prediction accuracy on the testing set for same and different models using ensemble learning

| Model | Type of ensemble | Prediction Accuracy |
|---|---|---|
| Three model of CNN | Bagging | 77.1% |
| Three model of VGG16 | Bagging | 79.8% |
| **Three model of InceptionV3** | **Bagging** | **87.2%** |
| CNN, VGG16 and InceptionV3 | Bagging | 86.5% |

Table 4, shows the accuracy of bagging ensemble with same and different architectures, these accuracies are graphically represented in below graphs, where each model structure is shown with epochs and accuracies.

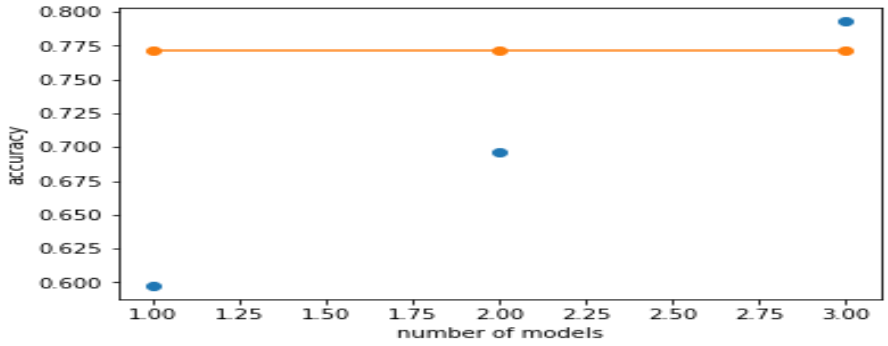


**Fig. 7.** Accuracy on train and test set in bagging ensemble models using CNN

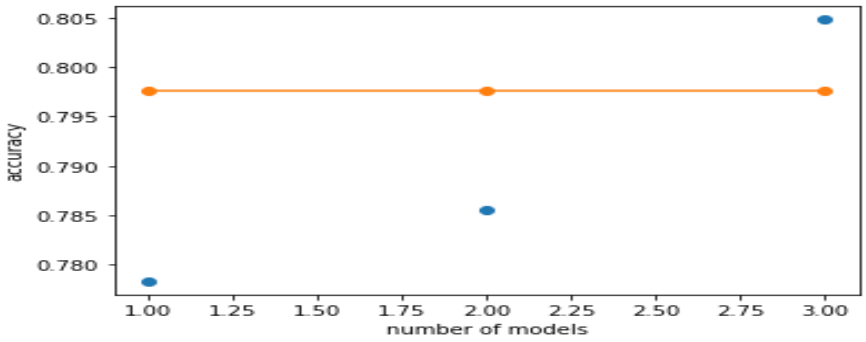


**Fig. 8.** Accuracy on train and test set in bagging ensemble models using VGG16

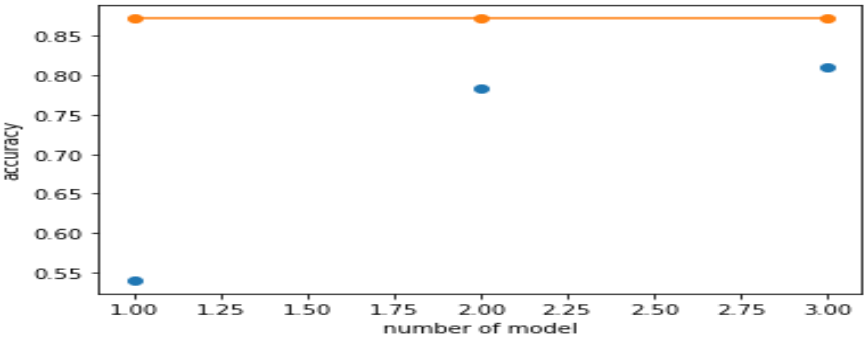


**Fig. 9.** Accuracy on train and test set in bagging ensemble models using InceptionV3

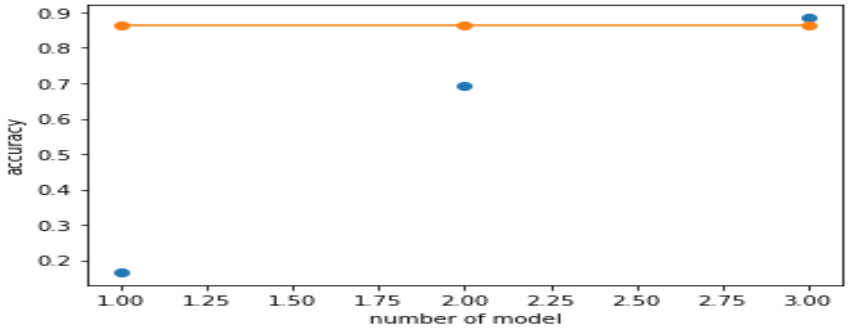


**Fig. 10.** Accuracy on train and test set in bagging ensemble models using CNN, VGG and InceptionV3

We compare accuracies of graphs above, and find out the following:

- Combining Inception v3 and All the models as a bagging ensemble which shown in Fig.9 and Fig.10 gives the best accuracy 87.20 % and 86.50%, which is far better than accuracy of graph in Fig.7 and graph in Fig.8.
- Due to the varied results between the models alone. We used the CNN which has poor accuracy compared to the other models. Therefore, we recommend using deep learning networks such as Alex Net or ResNet with Inception V3 to obtain the best accuracy.
- The confusion matrix in Fig.11 shows that all classification models are confused with Glaucoma and Normal eye when it makes prediction. Therefore, this problem must be addressed to optimize the classification.

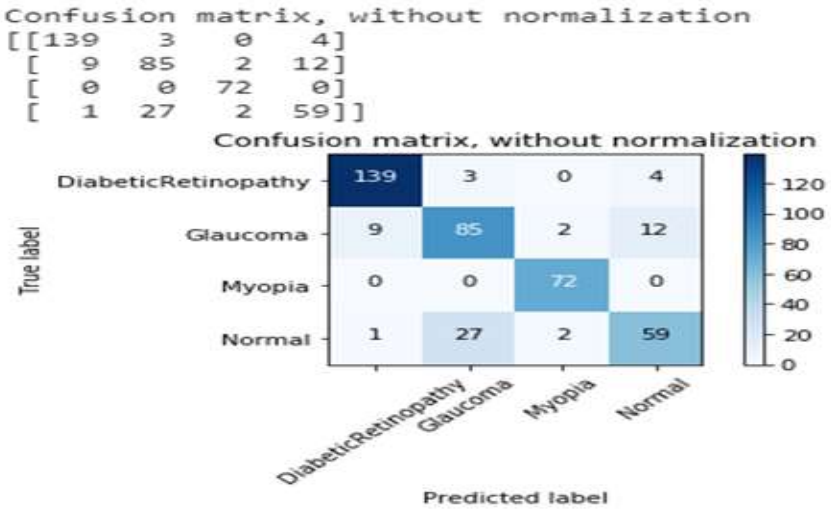


**Fig. 11.** Confusion matrix shows where our model confuse

## 9 Conclusion.

We studied the relative performance used bagging ensemble methods with deep convolutional neural networks as base learners on eye diseases data set, for image classification. In this work we have applied three systems for multi-class classification using a bagging ensemble, and we found that assembly of deep neural network models can outperform traditional methods that rely on learning algorithms alone.

Three models of multiclass classification CNN, VGG16 and Inception V3 have been compared in order to measure the accuracy and to know the effects of models assembly compared with learning algorithms alone. Due to the small number of the training datasets, we implemented the Fine-tuning and data augmentation to increase the accuracy of experiments in the test set. All the models mentioned above are deployed using python for multiclass image classification. We compared these three different structures of CNN on GPU systems using google Colab. With experiments,

as shown in table.4 we obtained results for each combination and observed that bagging ensemble based on Inception V3 combination gives better classification accuracy (87.20 %) than any other models.

We recommend using deep learning networks such as AlexNet or ResNet with Inception V3 to obtain better accuracy. Confusion matrix has been applied in our experiment to know in which class our models were confused. The results show that all classification models in varying proportions are confused with Glaucoma when it makes prediction as it shown in Fig.11. Therefore, this problem must be addressed to optimize the classification.

## References


1. American Macular Degeneration Foundation www.macular.org, 2019.
2. Ensembling ConvNets using Keras https://towardsdatascience.com/ensembling-convnets-using-keras-237d429157eb, 01/2020.
3. Ensemble averaging (machine learning), https://en.wikipedia.org/wiki/Ensemble_averaging_(machine_learning), 01/2020.
4. A Comprehensive Guide to Ensemble Learning, https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/, 01/2020.
5. Ensemble Learning- The heart of Machine learning, https://medium.com/ml-research-lab/ensemble-learning-the-heart-of-machine-learning-b4f59a5f9777, 01/2020.
6. Ju, Cheng, Aurélien Bibaut, and Mark van der Laan. "The relative performance of ensemble methods with deep convolutional neural networks for image classification." Journal of Applied Statistics 45.15 (2018): 2800-2818.
7. Huang, Jonathan, et al. "Acoustic scene classification using deep learning-based ensemble averaging." (2019).
8. Mo, Weilong, et al. "Image recognition using convolutional neural network combined with ensemble learning algorithm." Journal of Physics: Conference Series. Vol. 1237. No. 2. IOP Publishing, 2019.
9. Kumar, Ashnil, et al. "An ensemble of fine-tuned convolutional neural networks for medical image classification." IEEE journal of biomedical and health informatics 21.1 (2016): 31-40.
10. Beluch, William H., et al. "The power of ensembles for active learning in image classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
11. Minetto, Rodrigo, Maurício Pamplona Segundo, and Sudeep Sarkar. "Hydra: an ensemble of convolutional neural networks for geospatial land classification." IEEE Transactions on Geoscience and Remote Sensing (2019).
12. Tindall, Lucas, Cuong Luong, and Andrew Saad. "Plankton classification using vgg16 network." (2015).
13. Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
14. ImageNet. http://www.image-net.org/ ,01/2020
15. The interests of truth require a diversity of opinions, https://www.zest.ai/blog/many-heads-are-better-than-one-making-the-case-for-ensemble-learning, 12/2019

16. https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789136609/2/ch02lvl1sec16/max-voting, 01/2020
17. Diabetic Retinopathy Detection, https://www.kaggle.com/c/diabetic-retinopathy-detection/data, 01/2020
18. Visa, Sofia, et al. "Confusion Matrix-based Feature Selection." MAICS 710 (2011): 120-127.
19. Nezami, Omid Mohamad, et al. "Automatic Recognition of Student Engagement using Deep Learning and Facial Expression." *arXiv preprint arXiv:1808.02324* (2018).
20. Loussaief, Sehla, and Afef Abdelkrim. "Machine learning framework for image classification." 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT). IEEE, 2016.
21. Bizios, Dimitrios, et al. "Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by Stratus OCT." Acta ophthalmologica 88.1 (2010): 44-52.
22. LeCun, Y., Bengio, Y., et al. (1995). Convolutional networksforimages,speech,andtimeseries. Thehandbook of brain theory and neural networks, 3361(10):1995.
23. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105.
24. Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pages 6645– 6649. IEEE.
25. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
26. Open Dataset Finders, https://lionbridge.ai/datasets/the-50-best-free-datasets-for-machine-learning/,2019
27. Tindall, Lucas, Cuong Luong, and Andrew Saad. "Plankton classification using vgg16 network." (2015).
28. Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
29. Smolyakov, V. "Ensemble learning to improve machine learning results." (2017).