Using Ensemble Learning for Diagnostics 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
m.e.smaida@nuwm.edu.ua
The national university of water and Environmental Engineering, Revine, Ukraine
s.v.yaroshchak@nuwm.edu.ua

DOI: 10.29322/IJSRP.10.10.2020.p10639
http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639

Abstract- Deep learning is the most widely used image recognition technology today, and it provides more insight into how computers can understand and learn from data. In deep learning, artificial neural networks analyze a large set of data to automatically detect patterns. In this article, we'll introduce some of these techniques to see how we can use deep learning to create our own model for diagnosing eye disease. When we talk about determining a diagnosis by medical imaging, we mean that we are going to explain some of the main areas of computer vision and how we can classify images data using convolutional neural networks, one of most important idea is the evaluation of model performance using ensemble learning. In this study, CNN, Vgg16 and Inceptionv3 networks will be compared in detail using ensemble learning. In our work, a convolutional neural network based on Keras and TensorFlow is implemented using python. 1692 images were used for classification, which contained four types of eye diseases (diabetic retinopathy, glaucoma, myopia and normal), including 1,200 for training, 246 for validation and 246 for testing. CNN, VGG16 and InceptionV3 networks are trained in Colab GPU processors with three different combinations of classifiers using bagging, boosting and stacking. All types of ensemble learning have been implemented and the result is obtained. We found that through ensemble learning we could achieve higher performance than we could achieve from any of learning algorithms alone. On the other hand, we used the confusion matrix in our experiments to show us where our classifier is confused when it makes predictions.

Keywords. Vgg16, eye diseases, ensemble learning, Diabetic retinopathy, Glaucoma, Myopia, bagging, boosting, stacking.

III. Ensemble learning

Ensemble learning is a combining of a set of models in order to improve the model accuracy. Ensemble learning divided into simple ensemble techniques and advanced ensemble techniques.

I. Introduction

In this part, we will review the concepts relating to eye diseases such as glaucoma, myopia and diabetic retinopathy and its detection. These diseases are the most common eye diseases, and leads to blindness if it is not detected early. In recent years, due to the development of information technology and artificial intelligence, the diagnosis of diseases of the human visual system has made significant progress. Considering the complexity and variety of functions, where the number of diagnostic equipment, tools, algorithms and methods are developed. Sometimes a doctor can detect a specific condition after the analysis of visual images. However, in some number of cases, a diagnosis unsuccessful, due to many reasons, for example, low experience, fatigue, various forms of similarity, poor image quality, etc. In these cases, We must use intelligent image analysis systems in order to diagnose eye diseases [1], [18].

Ensemble learning is a collection of machine learning models combined to produce better results. In our study, we will use ensemble learning to improve model performance and make better learners. It is not a new algorithm, but assembling together several different algorithms or several different models to create an ensemble learning to improve the model accuracy. over all, ensemble learning helps to reduce noise, bias and variance. As it shown in fig.1 ensemble learning used a set of learner algorithms to improve predictive performance. [2], [19].

II. Problem statement

Eye diseases come in a variety of forms and are sometimes difficult for an optometrist to recognize and recognize. Therefore, information technology and artificial intelligent must be used to improve the existing system.

In our work, we will use ensemble learning to evaluate accuracy of three different CNN structures to determine eye diseases.

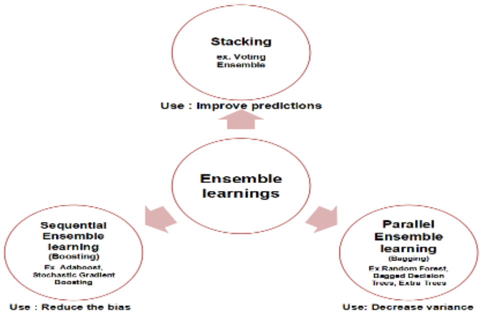

Stacking
ex. Voting Ensemble
Use : Improve predictions
Ensemble learnings
Sequential Ensemble learning (Boosting)
Ex. Adaboost, Stochastic Gradient Boosting
Use : Reduce the bias
Parallel Ensemble learning (Bagging)
Ex Random Forest, Bagged Decision Trees, Extra Trees
Use: Decrease variance

Fig. 1. Ensemble learning technique [5], [18]

A. *Simple ensemble* *Techniques [3], [18], [19]*

In this technique we use three different ways to predict:

a. Max Voting: the idea of this techniques is all models makes prediction and the voting will be for each sample. The final predictive class will be highest votes.

b. Averaging: It is creating and combining a set of models to produce a desired output. The output will be better performance than a given single model alone.

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

c. Weighted Averaging: In this way, each member contribution to the final evaluated is weighted by the performance of the model. the formula of weighted average takes the form:

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

B. Advanced ensemble techniques [4], [15], [18], [19]:

a. Bagging: This technique relies on a number of sub-datasets that we create, called bagging. Each bag containing a subset of the original dataset and a set of instances of random data with replacement. We use all bagging of our data to train all the models, obtain all results and implement the average or voting value as it shown in Fig. 2.

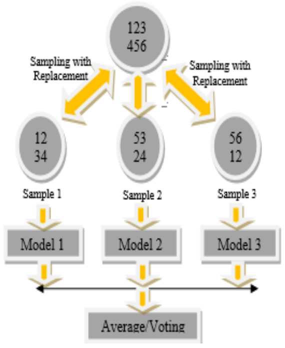


Fig. 2. Bagging example

b. Boosting: in this technique attempts to improve the learners by focusing on areas where the system is not performing correctly. We train a sequence of models where more weight is given to examples that were misclassified by earlier iterations. Classification tasks are solved through a weighted majority vote and regression tasks are solved with a weighted sum to produce the final prediction as it shown in Fig.3. in boosting, each subsequent model try to correct the errors of the previous model.

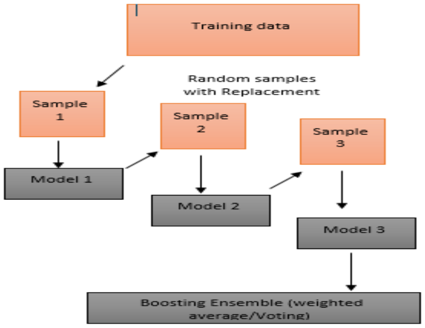

Training data
Random samples with Replacement
Sample 1
Sample 2
Sample 3
Model 1
Model 2
Model 3
Boosting Ensemble (weighted average/Voting)

Fig. 3. Boosting example

c. Stacking: As Fig.4. shows all models are trained in stacking based on all data set, and the output used as input features to train ensemble function. In our experiment we used logistic regression as a classifier.

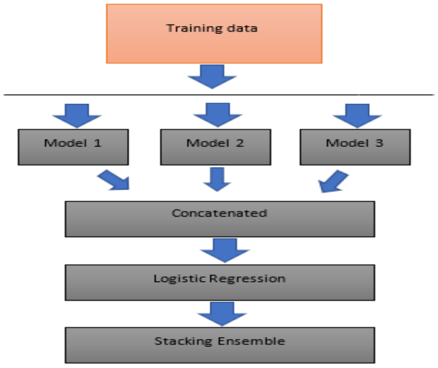


Fig. 4. Stacking example

Ensemble learning models are divided into three main category bases on their use, Table.1 shows the details.

Table 1. Types of ensemble learning based on the use [5]:

	Bagging	Boosting	Stacking

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639

Divide the data into subsets	Random	Prefer misclassified samples	Various
The goal to be achieved	Minimize variance	Increased predictive power	Both
The methods in which it is used	Random subspace	Gradient descent	Blending
The function of combining individual models	(weighted) average	Weighted majority vote	Logistic regression

IV. Related work

Ensemble learning has been used in many works using neural networks.

For example, Ju, Cheng, Aurélien Bibaut and Mark van der Laan. in their paper [6], they compared the empirical performance for some of the ensemble methods, such as, Neural network, VGG, GoogleNet and ResNet, including: Unweighted Averaging, Majority Voting, Bayes Optimal Classifier, Super Learner. The authors trained their models based on Same Network with Different Training Checkpoints, Ensemble of Same Network Trained Multiple Times, Ensemble of Networks with Different architectures, Learning from Weak Learner and Prediction with All Candidates. The results are obtained and achieved the best performance on testing set. All learners based on the CIFAR 10 data set. The unweighted averaging proved the best result when the performance of the base learners is comparable.

Huang, Jonathan, et al. [7] the authors used four different deep neural networks: Vgg12, ResNet50, AclNet, and AclSincNet in order to improve the accuracy of the models. All these models were pre-trained with Audio set data. The ensemble learning has been achieved in all these models and the experiment results were obtained over the validation set. The best result was achieved when all models combined together 83.01% through ensemble averaging.

Mo, Weilong, et al. [8] proposed image recognition algorithm based on ensemble learning algorithm and CNN structure (ELA-CNN) to solve the problem that single model cannot correctly predict. Bagging model has been used to train their different learners. The network structure used are combinations of ResNet, DenseNet, DenseNet-BC and Inception-Resnet-v2architecture. In their work they used cifar-10 as a dataset. Among them, there are 60,000 color images. These images were divided into 50000 as a training set and 10000 as testing set. They used the average probability in their final result.

Kumar, Ashnil, et al. [9] An Ensemble of Fine-Tuned CNN for Medical Image Classification has been used to classification-based diagnosis, teaching, and biomedical research. The authors used 6776 as a training images and 4166 as test images. They used two different architectures of CNN, AlexNet and GoogleNet to images classification. The experiments have been obtained using individual models and ensemble methods. Finally, the method achieved an accuracy that was consistent with the best accuracy among the other methods in ensemble method 96.59%.

Beluch, William H., et al. [10] investigate some recently proposed methods for active learning with high-dimensional data and CNN classifiers. They were compared ensemble methods against Monte-Carlo Dropout and geometric approaches. They found that the best result was obtained from ensembles, which are the basis for many learning algorithms of CNN such as, S-CNN, K-CNN, DenseNet, InceptionV3 and ResNet-50 for Diabetic Retinopathy classification. The dataset was used from MNIST, CIFAR and ImageNet. And they found that ensembles which are based on many active learning algorithms were the best result, and achieve a test set accuracy around 90% with 12,200 labeled images.

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

V. Data Description

Kaggle is a data science site that contains a different of interesting datasets. You can find all kinds of datasets in its list [11].

I used my data from competition in kaggle Diabetic Retinopathy Detection and from iChallenge-GON Comprehension which is a very big dataset of large number of annotated retinal fundus images from both non-glaucoma and glaucoma patients.

Inside the dataset more than 35 types of eye diseases. We will reduce the dataset with the four types. The dataset is consisting of images of Diabetic retinopathy, Glaucoma, Myopia, and Normal eyes which provided as a subset of images from a dataset of 1692 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 [17], [18], [19].

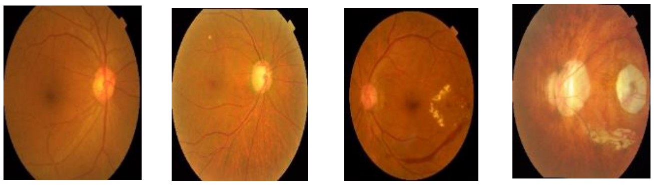


Fig. 5. Normal Fundus, Glaucoma, Diabetic retinopathy and Myopia

All our images have been used as the input of CNNs structure. We split our dataset into training, validation and testing data; each type of images has an individual folder and each image has a filename that is its unique id. Python language were used to achieve the goal with Google colab.

Table 2. images of eye diseases

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639

VI. Research Methodology

The block diagram of our proposed methodologies is shown in Fig.6 and Fig.7. Each block is labeled and represents the processing steps. In our work we compared three different structures of CNN, VGG16 and InceptionV3 in order to evaluate individually and using ensemble learning in order to determine eye diseases.

Firstly, we need to analyze and preprocess data we have. Here are 4 folders which contain 1692 images of Diabetic retinopathy, Glaucoma, Myopia and Normal where 1200 images used for training, 246 images used for testing and 246 images used for validation. Next steps include choosing the right model architectures, training chosen models and then measure model performance with accuracy metric. With obtained results we can compare performance of different approaches: individual training and ensemble learning. In this paper we have used three methods to study to evaluate performance of our classifier: CNN model which consists of three hidden layers and pooling layers as it shown in Fig 6A, Pre-trained CNN based on VGG 16 algorithms using the last block layer training as it shown in Fig 6B, and Pre-trained CNN based Inception v3 algorithms using the last block layer training as it shown in Fig 7.

Diabetic retinopathy	Glaucoma	Myopia	Normal eye
432	396	432	432

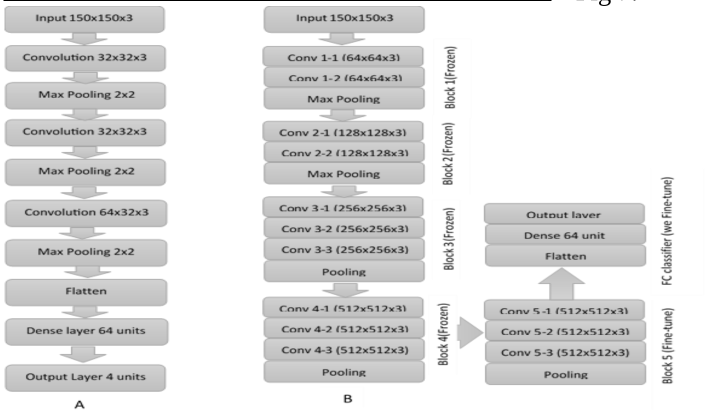


Fig. 6. CNN and VGG16 block diagram [18], [19]

A. Convolution Neural Network

Fig.6A shows convolutional layers, 150*150 pixels has been chosen as the size of input image (RGB). We used 32 filter of size 3*3 pixels to extract the features. The next layer is pooling layer, the size of windows we have used is 2*2 pixels to minimize the size of images. Another convolution layer we have used is 32 filters with size 3*3 and then max pooling size 2*2 pixels. We used 64 filter of size 3*3 pixels as the last convolution layers with pooling layer 2*2 pixels. Finally, fully connected layer 64 units and output layer 4 unit to prediction. After the forward pass and adjust its filter weights through backpropagation, the network is able to calculate the loss function and update the weights [18].

Function	Formula	Derivative
Weighted input	$Z = XW$	$Z'(X) = W$ $Z'(W) = X$
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)

$$\underset{\text{New weight}}{{}^{*}W_x} = \overset{\text{Old weight}}{W_x} - \underset{\text{Learning rate}}{a} \overset{\text{Derivative of Error with respect to weight}}{\left(\frac{\partial Error}{\partial W_x}\right)} \quad (4)$$

B. VGG 16.

Is CNN architecture developed by Visual Geometry Group from oxford university in 2014. This model using a 16-layers, 224*224 with three channels were used as the input images. The input images are passed through five blocks of convolutional layers. The five blocks layers were followed by three fully connected layers. The last layer is the output layer used for prediction. Fig.4B shows the block diagram of vgg16 [12], [18], [19].

C. Inception V3 [13],[14].

It is famous models consisting of 48 deep layers, one of the models that can be used for transfer learning and the ability of retrain the last layers of the model. As the fig.5 shown the inception layer is a comprised of set of layers [18], [19]:

- 1*1 Convolutional layer.
- 3*3 Convolutional layer.
- 5×5 Convolutional layer.
- 1×1 Convolutional layer before applying another layer used for dimensionality reduction.
- Max-Pooling layer.

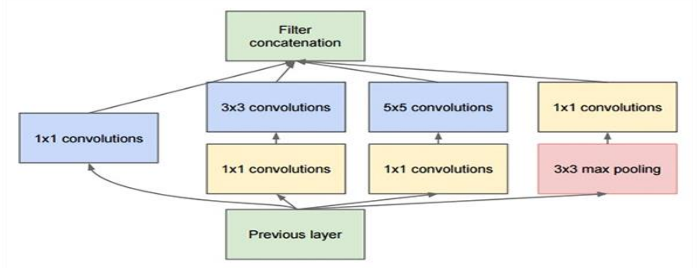


Fig. 7. Inception V3 block diagram

D. Confusion matrix.

It contains information about actual and predicted classifications which obtained from classification system. The performance of the systems is usually evaluated by the data in confusion matrix. It is the most common measures used to compare

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639

the classifiers. In our work, we will use this matrix to compare the model's accuracy.
Accuracy: A measure of how much we predicted the classes correctly [18], [19].

Accuracy = TP / TP+FP+FN+TN (5)

VII. Experiments and Results.

In our experiments, all the ensemble methods mentioned above have been deployed and the result obtained. The performance of all the ensemble methods were addressed including: bagging, boosting and stacking.

A. Results on models individually:

All the models mentioned before, namely CNN, VGG16 and InceptionV3 have been applied to eye diseases dataset as ensemble using bagging, boosting and stacking approaches. The result addressed in table.

Table 3. Individual models accuracy

Model	Epochs	Accuracy
CNN model	Early stopping	64.5%
VGG-16 model	30	68.39%
Inception-V3 model	30	74.17

Table.3. shows the results we have obtained from the three individual models. These results were graphically represented in fig.8.

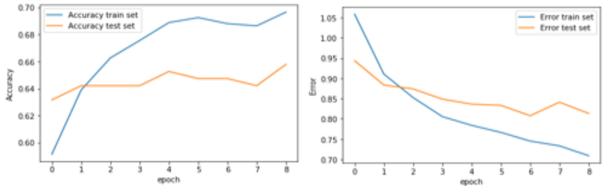


Fig. 8. Accuracy in CNN architecture

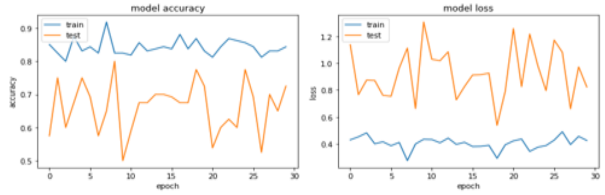


Fig. 9. Accuracy in VGG architecture

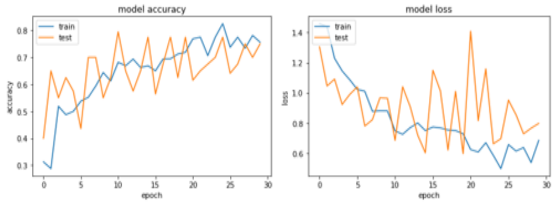


Fig. 10. Accuracy in InceptionV3 architecture

B. Results on three models using ensemble learning:

Ensemble learning of all the three models with different and the same structures have been applied, and the results obtained used bagging, boosting and then stacking. The performance was compared, and the results presented in Table.4 of each model.

Table 4. Models accuracy using ensemble learning

Model	Ensemble Types	Accuracy
Three CNN models	Bagging	66.73%
	Boosting	60.45
Three VGG16 models	Bagging	69.19%
	Boosting	62.78
Three InceptionV3 models	Bagging	83.76%
	Boosting	67.45%
CNN, VGG16 and InceptionV3 models	Bagging	73.22%
	Stacking	76.60%

Table 4, shows the accuracy of ensemble learning networks according to the types of ensemble, bagging, boosting and stacking. These accuracies are graphically represented in the graphs below, where each structure of the model is represented with epochs and accuracies.

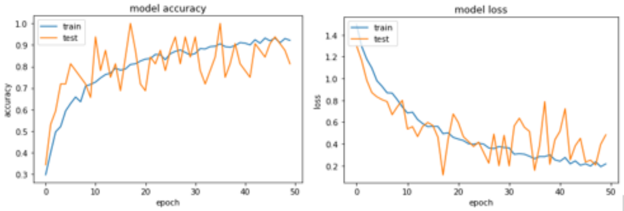


Fig. 11. Bagging ensemble accuracy using InceptionV3

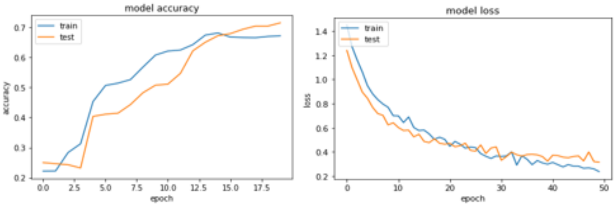


Fig. 12. Bagging ensemble accuracy using three models

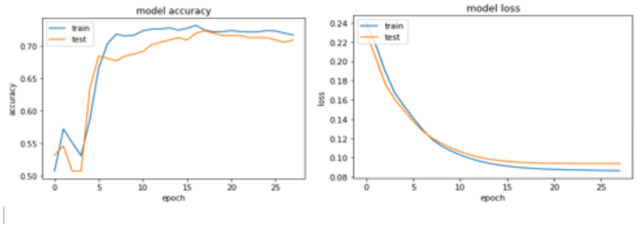
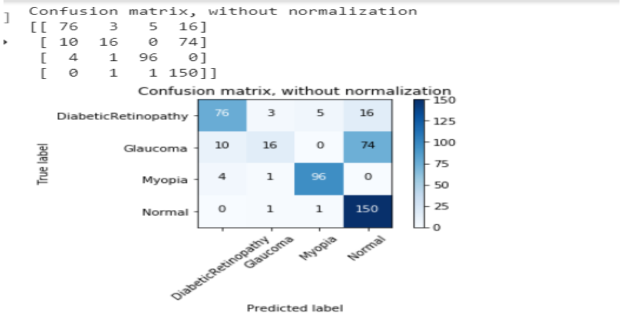


Fig. 13. Stacking ensemble models accuracy using three models

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639

The graphs above have been compared according on accuracy, and we conclude that combining all different structure as a bagging ensemble gives the best accuracy 86.43 % as shown in Fig.12. in addition, using of confusion matrix in Fig.14. shows that all models are confused with Glaucoma and Normal when it makes predictions. Therefore, in order to optimize the classification, this problem needs to be solved.



```
Confusion matrix, without normalization
[[ 76   3   5  16]
 [ 10  16   0  74]
 [  4   1  96   0]
 [  0   1   1 150]]
```

Fig. 14. Confusion matrix results

VIII. CONCLUSION

Deep learning is becoming very popular these days for solving image classification problems. In this work, we applied three models for classifying multiple classes using ensemble learning. we found that ensemble learning with convolutional neural network models can outperform traditional methods based on learning algorithms alone.

The comparative has been applied between three models, CNN, VGG16 and InceptionV3 in order to measure the model's accuracy using two ways, individual and ensemble techniques, and to know the effects of model's assembly on classification compared with learning algorithms alone. We implemented the fine-tuning and data augmentation to increase the accuracy of experiments in the test set. Python language with google colab have been used to compare these models. The results were obtained for each combination and observed that ensemble learning gives better results relying on the problem in your own learner. Bagging is used if your learner suffers from variance, boosting is used to reduces bias and stacking minimizes both variance and bias. Fig.12 shows the best results we obtained in bagging ensemble when we combine three different architectures. That means ensemble learning with multiple learner algorithms obtains better predictive performance than could be obtained from constituent learning algorithms alone.

We used CNN consisting of 3 hidden layers which has poor accuracy compared to the other models. Therefore, the authors recommend to use deep learning networks such as AlexNet or ResNet with Inception V3 to obtain better accuracy. Due to the lack of precision in our experiments, we used the confusion matrix to know in which class our models were confused. The confusion matrix shows us that all classification models in varying proportions are confused with Glaucoma, as it shown in Fig.14. Therefore, this problem needs to be corrected in order 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] Smaida, Mahmoud, and Serhii Yaroshchak. "Bagging of Convolutional Neural Networks for Diagnostic of Eye Diseases." *COLINS*. 2020.

[19] Smaida, Mahmoud, and Yaroshchak Serhii. "Comparative Study of Image Classification Algorithms for Eyes Diseases Diagnostic."

[20] Nezami, Omid Mohamad, et al. "Automatic Recognition of Student Engagement using Deep Learning and Facial Expression." *arXiv preprint arXiv:1808.02324* (2018).

[21] 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.

[22] Diabetic Retinopathy Using Fundus Image." Molecules 22.12 (2017): 2054.

[23] Adrian Rosebrock, "Keras Conv2D and Convolutional Layers", https://www.pyimagesearch.com/2018/12/31/keras-conv2d-and-convolutional-layers/, on December 31, 2018.

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639

[24] Maher, Raju, et al. "Automated detection of diabetic retinopathy in fundus images." International Journal of Emerging Research in Management andTechnology 4.11 (2015): 137-145.

[25] Attia, Mohammed, et al. "Multilingual multi-class sentiment classification using convolutional neural networks." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[26] Pavithra, G., and T. C. Manjunath. "Design of algorithms for diagnosis of primary glaucoma through estimation of CDR in different types of fundus images using IP techniques." Int. J. Innov. Res. Inf. Secur.(IJIRIS) 4.5 (2017): 12-19.

[27] Kumar, P. S. J., and Sukanya Banerjee. "A survey on image processing techniques for glaucoma detection." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 3.12 (2014).

[28] González, Javier, Michael Osborne, and Neil D. Lawrence. GLASSES: Relieving the myopia of Bayesian optimisation." (2016).

[29] Fandango, Armando. Mastering TensorFlow 1. x: Advanced machine learning and deep learning concepts using TensorFlow 1. x and Keras. Packt Publishing Ltd, 2018.

[30] Ganguly, Kuntal. Learning Generative Adversarial Networks: Next-generation Deep Learning Simplified. Packt Publishing, 2017.

[31] Pujari, Pradeep, Md Rezaul Karim, and Mohit Sewak. "Practical Convolutional Neural Networks." (2018).

[32] Ketkar, Nikhil. Deep Learning with Python. Apress, 2017.

[33] Smolyakov, V. "Ensemble learning to improve machine learning results." (2017).

AUTHORS

First Author
Mahmoud Smaida
Ph.D. student, The national university of water and Environmental Engineering, Revine, Ukraine
m.e.smaida@nuwm.edu.ua

Second Author
Serhii Yaroshchak
The national university of water and Environmental Engineering, Revine, Ukraine
s.v.yaroshchak@nuwm.edu.ua

Correspondence Author
Mahmoud Smaida
Ph.D. student, The national university of water and Environmental Engineering, Revine, Ukraine
m.e.smaida@nuwm.edu.ua

http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10639