Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 3, June 2021:1002- 1010

Research Article

# An Ensemble Classification Techniques Based On 'Ml' Model ForAutomatic Diabetic Retinopathy Detection

J Manjula[1], Sajja Radharani[2], N.Hanumantha Rao[3], Y.Madhulika[4]

**ABSTRACT**

Diabetic Retinopathy causes the blindness, is commonly known as vision destruction which cannot cure with surgery, through the spectacles or medication completely. A retinal tissue of eye is damaged, in the condition of blindness. Through the automatic computational mechanism, diabetic retinopathy severity is detected in present days hence the early stage of retina damage can be recognized before attacking the blindness. Health records are having uncovering patterns which provides a number of data regarding various diseases diagnosis to medical practitioners by using the important role of machine learning. Sometimes the disease detection accuracy is reduces because of health records sensitivity. An ensemble based machine learning (ML) model is proposed in the paper and diabetic retinopathy dataset is uses the machine learning algorithms as Decision Tree Classifier, Random Forest, K-Nearest Neighbour, Adaboost Classifier, J48graft classifier and Logistic Regression for experiment. The ensemble based machine learning model gives the best performance in terms of sensitivity and accuracy than the individual machine learning algorithms. The proposed work of Diabetic Retinopathy (DR) uses the Image dataset collection which is predefined from Kaggle. The accuracy of proposed work is 96.34% and having the improvement in the performance when compared with previous works. Ophthalmologists society is very much helped by this proposed ensemble based retinopathy detection.

***KEYWORDS:*** Machine Learning (ML), K- Nearest Neighbour, Decision Tree, Diabetic Retinopathy, Machine Learning Repository, Random Forest, Adaboost, J48graft, Logistic Regression.

## INTRODUCTION

Large amount of data is generated continuously and it is remarkable and scalable in the digital world. Biomedical engineering research, social networks and security domains are collects and use the large amount of data frequently [1]. This significant amount of data processing is too difficult for human beings in a particular domain within a predetermined time. One of the sub domains of artificial intelligence is Machine learning and it provides possible solutions to maximum domains [2]. Artificial neural network architecture is used in deep learning, is a sub field of machine learning techniques [3]. This techniques are uses the datasets in extraction of automatic learning ability and features. Different patterns in many areas are identified with the exploitation of data miningtechniques and supported in the process of making decisions [4].

[1]Assistant Professor, Department of CSE, B V Raju Institute of Technology, Telangana, India
[2]Assistant Professor, Department of CSE, VFSTR Deemed to be University, A.P, India
[3]Assistant Professor, Department of CSE, RVR & JCCE, A.P, India
[4] Assistant Professor, Department of IT, RVR & JCCE, A.P, IndiaData mining techniques are used in

Biology and medicine fields for providing better services to the patients in different purposes. Retinal diseases are identified by the ophthalmologist after the investigation on retinal images. Diabetic Retinopathy (DR) detection is presented in this method and it is caused because of prolonged Diabetes [5]. Exudates, Haemorrhages and Microneurysms are the biomarks which characterizes the Diabetic Retinopathy. Early stage symptoms are not revealed in this DR even though if identified early then the treatments are successful. This DR detected in early stage through the patient's regular screening with the probability of affected patients. The bio-markers facilitating diagnosis with Manual investigation requires high proficiency and a time prone task in detecting the diseases. High accuracy results are obtained through the computational solutions. The screened fundus is examined with the help of Data Mining and Image processingtechniques [6]. In early stage of retinal fundus extraction features are uses the image processing then after Data Mining techniques are also adopted for learning model designing process by using if-then-else rules. In retinal fundus images, disease presence or absence is identified by the utilization of theserules [7]. In some cases diabetic retinopathy detection process is includes with two-level classification. In first level of classification, the learning models are does not fit for the instances so the quality of data is degraded with the effect of noise. These instances areremoved and the classification is done for resultant clean data in a second level. Retinal fundus images classification uses the results from second level classification.

## DIABETIC RETINOPATHY AND MACHINELEARNING ALGORITHMS

For the diagnosis of patients disease prediction and detection uses different machine learning methods [8]. Individual algorithmic implementations are not giving the satisfactory results. Therefore an approach of ensemble with the existing machine learning algorithms is proposed in this paper which deliversthe best results than the individual algorithms.Diabetic Retinopathy dataset detection is designed through the ensemble the different machine learning algorithms namely, Random Forest classifier (RF) [9], Decision Tree Classifier (DT) [10], Adaboost classifier [11], K-Nearest Neighbour classifier (KNN) [12], Logistic Regression (LR) [13] and J48graft classifier. The min-max normalization method is normalizes the diabetic retinopathy dataset, then the training is done by using ensemble model. In terms of accuracy and security, the performances of proposed ensemble model are calculated. Therefore the bestresults are observed at proposed ensemble model thanthe individual ML algorithms.

**Diabetic Retinopathy:** Diabetic retinopathy(DR) is causes the blindness and vision impairment inworking adults [14]. If the blood sugar levels are highthen it damages the eye retina eye retina which leads to cause the Diabetic retinopathy. With the help of experienced ophthalmologist who knows very well about DR gives the valuable information to the patient by observing fundus images. If DR is detectedin early stages through the regular screening and which reduces the effect of Diabetic Retinopathy. DRand NO DR fundus images classification are obtainedwith the applications of machine learning algorithmic techniques. Retina image datasets are taken from the Kaggle for experiment in this ensemble model.

**Machine Learning:** Predictions are done by using the machine learning algorithms in general. Huge amounts of data with training algorithms uses in this machines learning for learning ability. According to the similarity of functions or forms and learning style Machine learning is categorized. Humans are providing the training data in supervised learning every input element for learning the algorithm and establish a feedback relationship in between the input data and

output data. If accurate prediction of algorithm is occurred then the new setof data adopts the trained algorithm. Supervised learning is having the two main categories as regression and Classification. Some of the classification algorithms are Naïve Bayes, Support Vector Machine [15], Bayesian networks, Neural Network and Hidden Markov model.

**Deep Neutral Network (DNN):**

A multilayer neural network with the capability of deep learning is known as Deep Neural Network (DNN). The desired output is obtained from the proper designing of DNN which extracts complex and higher features progressively from the data at each layer. Automatic representation of learning data is mainly focused by the deep neural networks. When data extraction from the unstructured datasets, then DNN is came into action. Unstructured data sets are having the unstructured images. Restricted Boltzmann Machine, Convolution Neural Network and Sparse Auto encoders or Long Short Term Memory are implemented as deep neural networks.

## ENSEMBLE CLASSIFICATION BASED 'ML' MODEL FOR DIABETIC RETINOPATHY

## DETECTION

Diagrammatic representation of an ensemble classification techniques based ML model for automatic diabetic retinopathy detection is depicted in the Fig. 1.

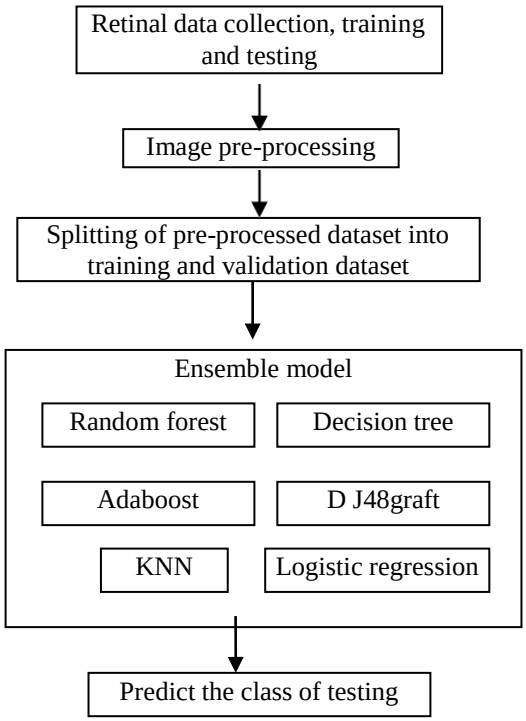


**Fig. 1: PROPOSED ENSEMBLE MACHINELEANING MODEL**

**Retinal data collection, training and testing:** In medical engineering field data collection is very important along with the privacy and security or protection. Unspecified datasets obtained at local or publicly available datasets are used in related fieldsby the researchers. Image dataset collection is used inthis ensemble model from the Kaggle.

**Image Pre-processing:** the image padding border pixels contained information which is preserved with zeros is done. RGB color image is obtained from the conversion of image. In MobileNetV2 model different sizes of images are present and the required input size of image is 224 x 224 therefore images are resized according to their requirements.

**Splitting of Pre-processed Images:** Preprocessed datasets are divided according to Random method into validation and training data. Designing model uses the training data and its effectiveness is examined through validation data. Two subsets are splitting according to 90/10, 80/20, 70/30 or 50/50. In preprocessed images 80% is used as training set and remaining 20% is used by test set.

**Ensemble Learning:** multiple machine learning models are trained at a time in the Ensemble learning process and generate an improvised performance if it compares to individual machine learning models. Ensemble word refers to trained predictors forpredictions computing. Ensemble learning implementation is employed in solving quantitative problems by using decision trees. Different decision trees are generates the ensemble results based final classification computed from the ensemble learning Instead of depending on the predictive result and analysis of one decision tree. Different machine learning models are used in the ensemble learning implementations and such as Random Forest Classifier, SVM classifier, Logistic regression and so on, finally presents these models aggregated results.

**Random Forest Algorithm:** An ensemble model is operated with many trees of Random forest. Class prediction is divided in the random forest individual trees and maximum gained votes class is becomes the models prediction. Different uncorrelated trees combinations are operating together for prediction process is the main concept of random forest. Each individual tree behavior is guaranteed by the two methods followed Randomforest and not correlated with other trees in the model. First method followed by the random method bagging, advantage of the fact are received by the random forest that is in training the data of decision trees are very delicate so a small change in the training set resulted to change the entire tree structure. Feature randomness is the second method, in which features random subsection is chosen by the each tree of random forest. This results the large variation in between the trees in the model and gives more diversification and lower correlation across trees.

**Decision Tree Algorithm:** Different branches, leaves and nodes are contained by hierarchical model of decision tree (DT). A test on a feature is depicted in this model by every node. Class label is denoted by the leaf and those class labels features are represented by branches. Classification policies are indicated by the emerging root from root to leaf. Based on different situations the datasets can undergo division through a logic and decision trees are constructed accordingly. In each process of division uses the features which are selected by information gain in constructing the tree.

**AdaBoost Algorithm:** Adaptive Boosting algorithm short form is AdaBoost algorithm. Strong classifiers are created from the weak classifiers using Adaptive Boosting algorithm with different techniques of machine learning. The ML algorithms performance is boosted or improved through the main aim of AdaBoost algorithm. A model is created for classifier boosting by using training data and first model errors are rectified by adding the second model. Models are added continuously until perfection is achieved by the predictions and more number of models is added.

**J48graft:** grafting is done on the Decision trees grafting as a post-process which means that instance space parts are reclassified greatly in which contains only misclassified data or no training data so error prediction is reduced. This method identifies leaf regions that should be reduced and afterward new leaves are generated with novel classifications through the branching out process, as a result produced a complex tree. Only branching is allowed by this process and classification errors introduction is avoided which are corrected previously. So in brief the errors are eliminated by the grafting technique rather than introducing them. To provide a more efficient means of evaluating the supporting evidence, the C4.5A algorithm was originally introduced by Webb. This algorithm is associated with grafting from leaf regions of All Test-But-One-Partition (ATBOP), those regions are defined and results as all surrounding decision surfaces are removed.

**K-Nearest Neighbour (KNN):** Regression and classification problems are solved by using the K- Nearest Neighbour (KNN) as a classical machine learning algorithm. According to the distance function similarity measurement datasets are classified by the KNN, it is a supervised learning and in several applications like pattern recognition data mining, image processing and intrusion detection are with non-parametric algorithm. In KNN, Data classification is depends on the respective nearest neighbor majority votes. KNN performance level is optimized by the minimum features counting. If the features counting is high, then faced an over fitting problem.

**Logistic Regression (LG):** Different class sets are allocated by the observations which use the classification method of Logistic regression (LR) and probability concept is also used. Multi-linear and binary are the two types in the Logistic regression. Sigmoid Function is the cost function used in the linear regression as Logistic regression. Between 0 and 1, any values are transformed by this function. Probability and predictions are correlated with the help of this function.

**Predicting the testing images class:** Testing dataset best predictions are obtained by considering all the predictions of individual classifiers.

## RESULTS

The dataset of diabetic retinopathy investigations are examined by using Each ML algorithms of hyper parameters set. These hyper parameters are tuned according to grid search in this work. In different stages experiment is carried out firstly, apply the machine learning methods to the dataset which follows the order as Random Forest classifier, Decision Tree, Adaboost, K-Nearest Neighbour, Logistic Regression, and J48graft. Final prediction is achieved in the ensemble model by applying all datasets to machine leartning algorithms. 3624 fundus images

were contained in the Kaggle and provide the training data with five class labels. These training datasets with DR five classes are provided in the Table 1.

Table 1: TRAINING DATASETS

| level | Type of DR | No. of. Records |
|---|---|---|
| 0 | No DR | 1789 |
| 1 | Mild | 410 |
| 2 | Moderate | 950 |
| 3 | Severe | 170 |
| 4 | Proliferative | 305 |

Various classifiers are combined in the ensemble model so these advantages are also combined which has its unique and own advantages. Therefore ensemble model gives the better results than the individual classifiers. The main aim of the ensemble model is best feature extraction from the classifiers, for this an ensemble model uses the voting mechanism. Each classifier predictions are treated as a vote in this voting methodology and majority of the classifiers are make the predictions or votes to the ensemble model so final prediction is computed.

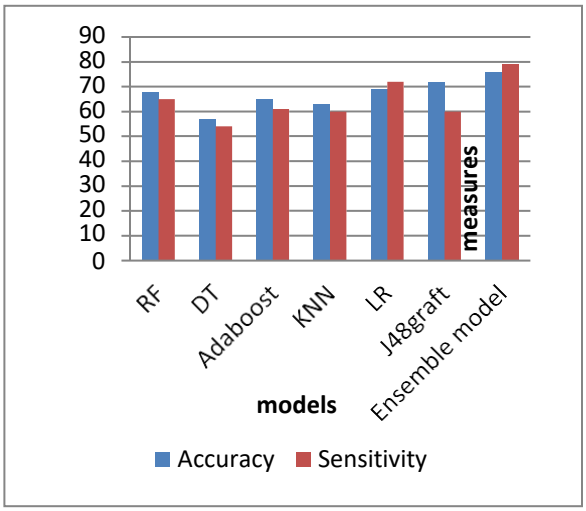


**Fig. 2: ACCURACY AND SENSITIVITY OFDIFFERENT MODELS**

Proposed ensemble model sensitivity and accuracy are calculated with the addition of all classifiers. From Fig. 2, it is clear that ensemble model gives the best accuracy than the other classification models.

The experiment is done in Weka, an open source data mining tool. Best rules are constructed for diabetic retinopathy detection from different experiments. Best rules are eveolved through the different classifiers namely Decision Tree (DT), Adaboost, KNN, J48Graft, Random Tree (RT), Logistic Regression (LR). Individual classifiers are classifying the 3624 instances at preliminary stage and then apply for the ensemble model. The bagging can attempt the ensemble with majority policy and Table 2 contains the results of ensemble model which are far better than individual classifiers. Results are not satisfied in the terms of validation and as well as accuracy.

Hence noise considered data prevents the effective building of learning model. Different ensemble models are used for removing the noise.

Table 2: CLASSIFICATION ACCURACY (%)WITHOUT ELIMINATION OF NOISE

| Classifier | Without ensemble | Bagging |
|---|---|---|
| RF | 64.90 | 67.51 |
| DT | 65.07 | 65.77 |
| KNN | 59.08 | 60.04 |
| LR | 74.28 | 74.20 |
| J48graft | 63.42 | 66.73 |

The highest possible result is achieved through Ensemble of J48Graft Trees demonstrating an accuracy of 96.34%. Through the achieved accuracy proposed system is justifiable in real time environment.

Table 3: CLASSIFICATION ACCURACY (%) OF INDIVIDUAL CLASSIFIERS ON NOISE ELIMINATED DATA

| **Classifiers** | **Without feature selection** | **CFS** | **PCA** |
|---|---|---|---|
| RF | 90.82 | 91.91 | 78.02 |
| DT | 94.57 | 91.91 | 77.05 |
| KNN | 81.40 | 85.63 | 73.31 |
| LR | 85.99 | 85.02 | 83.58 |
| J48graft | 94.93 | 92.03 | 76.69 |

BF Trees ensemble can effectively eliminate the noise. Without feature selection classification with correlation based feature selection (CFS) and Principal Component Analysis (PCA) are different experiments which are presented in this paper. Individual classification results are represented in the Table 3.

Higher accuracy is obtained through the ensemble which is represented in the below Table 4. Neither Feature selection nor Principal Component Analysis yielded higher results than the results obtained from the entire set of features. Classifier ensemble is conducted in the experiment so better accuracy prediction is obtained at classifiers ensemble. Improved results are

described in the Table 4.

Table 4: PERFORMANCE OF ENSEMBLED CLASSIFIERS (ACCURACY %) ON NOISE ELIMINATED DATA

| Classifiers | Without feature selection | CFS | PCA |
|---|---|---|---|
| RF | 90.46 | 92.76 | 81.76 |
| DT | 93.84 | 91.67 | 80.31 |
| KNN | 81.52 | 85.15 | 72.71 |
| LR | 85.99 | 85.27 | 83.45 |
| J48graft | 96.34 | 92.15 | 80.31 |

## CONCLUSION

Diabetic retinopathy dataset is evaluated in this work of ensemble model with machine learning algorithms such as decision tree, random forest, K-Nearest Neighbour, Adaboost, J48graft and Logistic Regression. The ensemble model trains the dataset. Designing model uses the training data and its effectiveness is examined through validation data. Machine learning algorithms individual performances are compared with the performance of proposed ensemble ML model. With limited size the proposed model is tested and it is the limitation of this work. The comparative analysis gives the best results at proposed ensemble model than the individual ML algorithms with improved performance reporting 96.34% accuracy on Diabetic Retinopathy dataset. This approach can be transferred to other biomedical engineering image classification problems using small training data.

## REFERENCES

1. Zeng Zeng, Cuntai Guan, Ziyuan Zhao, Cen Chen, Kaixin Xu, "DSAL: Deeply Supervised Active Learning from Strong and Weak Labelers for Biomedical Image Segmentation", IEEE Journal of Biomedical and Health Informatics, Year: 2021
2. Ajay Kumar Sharma, Mayank Patel, Divya Kothari, "Implementation of Grey Scale Normalization in ML & Artificial Intelligence (AI) for Bioinformatics using Convolutional Neural Networks (CNN)", 2021 6th International Conf. on Inventive Computation Technolo. (ICICT), Year: 2021
3. B.T. Jadhav, I.K. Mujawar, R.Y. Patil, V.B. Waghmare, "Development of Diabetes Diagnosis System with Artificial Neutral Network and Open Source Environment", 2021 International Conf. on Emerging Smart Com. and Informatics (ESCI), Year: 2021
4. Anu Sharma, R.K. Dwivedi, Shubhi Goel, "Analysis of Social Network using Data-Mining- Techniques", 2020 9th International Conf. System Modeling and Advancement in Research Trends (SMART), Year: 2020
5. Hui Zhou, Ying Zhu, Lifeng Qiao, "DR Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative DR Based on Deep Learning Algorithms", IEEE Access, Year: 2020

6. Teng Zhou, Shiqiang Tang, Changli Li, Jingwen Yan, Hon Keung Kwan, "Color Correction Based on CFA and Enhancement Based on Retinex With Dense Pixels for Underwater Images", IEEE Access,
7. Year: 2020
8. Nikita Demin, Nataly Ilyasova, Rustam Paringer, Alexandr Shirokanev, "Fundus Image Segmentation Using Decision Trees", 2020 International Conf. on Information Tech. and NT (ITNT), Year: 2020
9. Masahiro Ishikawa, Naoki Kobayashi, Satoki Homma, Hinako Okazaki, "Disease Detection Using Machine Learning in Vital Sign Data Telemonitoring", 2020 IEEE 2nd Global Conf. on Life Sciences and Technolo. (LifeTech), Year: 2020
10. Mariya Ivkina, Liliya Demidova, "Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier", 2019 1st International Conf. on CS, Mathematical Modelling, Automation and Energy Effici. (SUMMA), Year: 2019
11. Kudaravalli Deepika, Khadri Lalitha vanisri, Pradeep Kandimalla, Sajja Radharani," A meta- heuristic Bat inspired algorithm for Flow shop Scheduling"2020, International journal of future generation communication and networking. Year: 2020.
12. Delali Kwasi Dake, Esther Gyimah, "Using Decision Tree Classification Algorithm to Predict Learner Typologies for Project-Based Learning", 2019 International Conf. on Comp. Computational Modelling and Apps. (ICCMA), Year: 2019
13. Musa M. Ameen, Wisam A. Qader, "Diagnosis of Diseases from Medical Check-up Test Reports Using OCR Technology with BoW and Adaptive Boosting algorithms", 2019 International Engg.
14. Conf. (IEC), Year: 2019
15. Hong Shan, Xiaofeng Zhong, Liang Gao, Shize Guo, "A Transfer k-NN Classifier with the Bagging Method", 2018 Eighth International Conf. on Instrumentation & Measurement, Computer,
16. Communi. and Control (IMCCC), Year: 2018
17. Sajja Radharani, Amar Jukuntla, Madhubabu Chevuru, P Amarnatha Reddy,"A novel Enhanced ensemble clustering techniques in machine learning and data mining",2019 Journal of advance research in dynamic & control Systems, Year: 2019.
18. Linling Tang, Hengwei Lv, Qian Lei, Hong Sun, Haijian Zhang, "Logistic regression-based device-free localization in changeable environments", 2016 IEEE 13th International Conf. on Signal Proces. (ICSP), Year: 2016
19. Keshab K. Parhi, Dara D. Koozekanani, Sohini Roychowdhury, "Automated detection of neovascularization for proliferative diabetic retinopathy screening", 2016 38th Annual International Conf. of the IEEE Engg. in Medicine and Biology Society (EMBC), Year: 2016
20. Huseyin Seker, Charalambos Chrysostomou, "Structural classification of protein sequences based on signal processing and support vector machines", 2016 38th Annual International Conf. of the IEEE Engg. in Medicine and Biology Society (EMBC), Year: 2016