An Improved Approach for Detection of Diabetic Retinopathy Using Feature Importance and Machine Learning Algorithms

S M Asiful Huda†, Ishrat Jahan Ila‡, Shahrier Sarder┼, Md. Shamsujjoha*, Md. Nawab Yousuf Ali▲

†‡┼▲Department of Computer Science and Engineering, East West University, Dhaka-1212, Bangladesh
* Faculty of Information Technology, Monash University, VIC 3800, Australia

Email: †hudaasiful@gmail.com, ‡ishratjahanila@gmail.com, ┼shahriersarder@gmail.com,
*dishacse@yahoo.com, ▲nawab@ewubd.edu

***Abstract*—Diabetic Retinopathy is a human eye disease that causes damage to the eye's retina and may ultimately result in complete blindness. Early detection of diabetic retinopathy is needed to avoid complete blindness. Physical tests, such as visual acuity test, dilation of pupils, optical consistency tomography, is used to detect diabetic retinopathy. However, it is costly in terms of time and might affect the patients. In these consequences, this paper detects the presence of Diabetic Retinopathy in the human eye using a machine learning algorithm. The proposed method applies classification algorithms on several features (e.g., optical disk diameter, lesion-specific (microaneurysms, exudates) or presence of hemorrhages) of an existing Diabetic Retinopathy dataset. Then the features were extracted and used for the final decision making to predict the presence of diabetic retinopathy. The proposed system used Decision Tree, Logistic Regression. Support Vector Machine for the prediction. The proposed method achieved 88% accurate results which is much better than the existing works. Moreover, the proposed method achieves a better score in precision and recall which are 97% and 92%, respectively compared to the existing result 72% and 63%, i.e., more the 25% in each category on average which proves the enormousness of the proposed method.**

Keywords—Diabetic Retinopathy, Machine Learning, Logistic Regression, Feature Importance, SVM.

I. Introduction

Diabetic retinopathy or diabetic eye disease is caused by diabetes mellitus commonly referred to as diabetes. This is a group of metabolic disorders where there is a high sugar level in the blood for a long period. The mellitus manifests itself in the retina. Diabetic eye disease is one of the most frequent causes of complete blindness in many developed countries. The detection of retinal pathologies became much easier using automated retinal image analysis whereas other methods like dilation of eye pupil are time-consuming and the patient has to suffer for some time. Nowadays it is one of the most severe and widely spread eye diseases. It is the most regular cause of permanent blindness in the group of the working-age population of developed countries [1]. Diabetic retinopathy occurs when diabetes damages the blood vessels inside the retina, leaking blood and fluids into the surrounding tissue. This fluid leakage produces microaneurysms, hemorrhage, hard exudates, and cotton wool spots (a.k.a., soft exudates) [2], [3]. This is a silent disease and may only be recognized by patients when changes in the retina have progressed to a level where treatment becomes difficult or even impossible.

Thus, this work makes the following contribution. This paper proposes an automated diagnosis approach of Diabetic Retinopathy based on an immense and crucial feature extracted from the DIARET-DB dataset to help people detecting diabetic retinopathy at the primary level. The dataset contains 15945 samples along with 66 features associated with each sample referring to various symptoms of a primary, mild and advanced level of Diabetic Retinopathy. Some of the features are Area, bounding box, convex area, Regional intensity coefficients corresponding to the max, min, mean variance of the green plane, red plane etc. By applying a feature selection technique, the most important and crucial features were extracted first. Then those features were used to train some of the superior machine learning models to figure out how accurate our model works in determining the presence of the symptoms on a patient's eye. The main goal here is to automatically classify the proliferative as well as non-proliferative diabetic retinopathy grade of any retinal image.

The contribution of this proposed method is a fully automated, fast and almost accurate DR detection As a result, we obtained a maximum accuracy of 88% in both Logistic Regression Classifier and Support Vector Machine and an underperforming 72% in Decision Tree Classifier which means our model is nearly very much successful in predicting the symptoms of Diabetic Retinopathy. A K-Nearest neighbor classifier was also implemented to get an overall idea of the dataset which achieved an accuracy of 63%, which is comparatively falling short than SVM and Regression Classifier. This whole result was achieved using only 30 important features from the 66 features out of the 15945 samples.

II. Background Study and Dataset

The huge number of Diabetic Retinopathy cases worldwide requires assisting the whole diagnosis process in detecting Diabetic Retinopathy. A great amount of time and effort is saved by implying automatic detection of Diabetic Retinopathy. Thus, a support system on decision making was examined by

978-1-7281-1557-3/19/$31.00 ©2019 IEEE *2019 7th International Conference on Smart Computing &Communications (ICSCC)*

978-1-7281-1557-3/19/$31.00 ©2019 IEEE

Maher et al [4]. A Support Vector Machine classifier was implemented in this case. A great number of image processing techniques were implemented to detect Diabetic Retinopathy [5] – [8]. A neural network classifier based on the area surrounded by vessels and exudates was evaluated and examined by Nayek et al [9]. Detection accuracy of 93% was achieved by them. A support vector machine classifier was implemented by Acharya et al where bi-spectral invariant features were fed into the classifier to classify the fundus image into prolific and non-prolific [10]. An automated approach in the diagnosis system to classify several 3 types of early symptoms like microaneurysms, red lesions, hemorrhages and, exudates and cotton wool spots of DR was proposed to classifyNPDR [11]. An accuracy of nearly 83%, and 88.3% using 430 images and another in the range of 85-87%, and 93% using 360 images were achieved by them, respectively, for microaneurysms, hard exudates, hemorrhages and, and cotton wool spots.

A. The DIARET-DB Dataset

As there has been huge research and studies on theautomated diagnosis of early detection of Diabetic Retinopathy there has been numerous datasets available online on several sites like Kaggle, GitHub and lot more.

We used a dataset called DIARETDB1 which has widely been used for evaluating and implementing computer-aided diagnosis models. This local dataset has been formed using those 89 fundus images. It includes manually marked images which show various levels of Diabetic Retinopathy.

This dataset includes Region based lesion information with 6 individual classes having 15,945 samples each with corresponding 66 features. [12]

The class labels have the following meaning:

- Class 0: Bright non-lesion (Healthy Eye)
- Class 1: Hard exudates (Early DR)
- Class 2: Cotton wool spots (Early DR)
- Class 3: Red non-lesion (Early DR)
- Class 4: Microaneurysms (Early DR)
- Class 5: Hemorrhages (Severe DR)

Some of the crucial features among those 66 features are an eccentricity, Area, equivalent diameter, perimeter, Orientation, 16 Region based Intensity Coefficients like min, max, mean, and variance of pixels in the red, green, intensity and hue plan and so on.

III. METHODOLOGY

From the above discussion, we already came to know that there are six classes in total in our dataset, each expressing a label that is seen in the early diabetic retinopathy stage. As the effectiveness of detecting the symptoms manually depends on years of experience, scientists and researchers are trying hard to automate this whole procedure. In our work that is our main objective.First, we look to preprocess our dataset through some feature scaling and feature selection method, then we build our model using several Machine Learning Algorithms and lastly, we will evaluate our findings

A. Algorithms

In our model, we used 4 different types of superior machine learning algorithms and for the implementation work, we used Python 3.6 as our programmable language and Anaconda Platform for the implementation. The algorithms we used are Logistic Regression, Support Vector Machine, K- Nearest Neighbor and Decision Tree[13] [14] [15]. These algorithms are good for different classifications and they got their own properties and performance based on different datasets. As we said earlier, Logistic Regression and SVM are more commonly used algorithms for classification problems.

B. Preparing the Dataset

1. Normalization: As our dataset contains many features, and some of the features have some extreme values, we do not know how our overall data is distributed. In this context, we used Normalization on all the attributes except the label of our whole dataset. It is the process of rescaling one or more attributes to the range of 0 to 1. It helps to get an overall idea over the distribution of our dataset.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \ldots\ldots\ldots (1)$$

2. Standardization: Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. As some of the attributes of this DIARETDB1 dataset have higher ranges, this overall process minimizes that difference between the maximum and minimum value and gives a better visualization of the data points than before.

$$x_{new} = \frac{x - \mu}{\sigma} \quad \ldots\ldots\ldots (2)$$

Feature Selection: As our dataset comes with a great number of features, it would take a lot of computation time if we want to train our model with all the attributes we have. In cases like these features, selections are one of the crucial concepts which have a direct impact on the performance of the model. It can reduce training time, improve accuracy measure and avoids the problems related to overfitting. There are several feature selection techniques like Univariate Selection, Correlation matrix with heatmap and so on. Here in our study, we have used feature importance to get the most important features that we need to consider while building our model.

Feature Importance: Feature importance of each feature of the dataset can be obtained by using the feature importance property of the model. It is an inbuilt class that comes with a tree-based classifier. It gives a score for each feature or attributes of the data, the higher the score the more important. We have used the extra tree classifier for extracting the top 30 features for the dataset. As this is a tree-based classifier, every node in a decision tree is a condition on a single feature. This is

designed to split the dataset into two in such a way so that values having similar responses end up going onto the same set. The parameter which defines this criterion is called as Impurity. Based on this measure, the optimal condition is chosen. For classification problem, it is typically Gini Impurity. Thus, while training a decision tree it can be easily calculated how much the weighted impurity is decreased by each individual feature or attribute. The impurity decrease calculated from each feature can then be averaged and features are ranked according to this feature. The most important features will decrease the weighted impurity less than the less important features in the whole dataset.

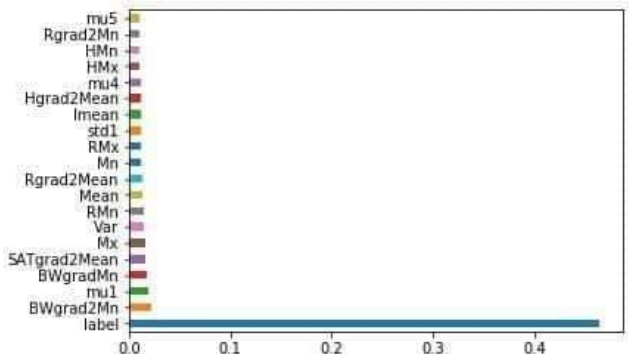

Figure 1. Top features using feature importance technique.

For each decision tree,scikit learn calculates the importance of a node using Gini Importance, assuming only two child nodes. The final feature importance of a node or a feature is its average over all the trees. The value is normalized finally to 0 to 1 by diving by the sum of all feature importance values [16-20].

We selected the top 30 features for training our model. There are several reasons behind selecting 30 others than any number. Firstly, feature selection based on Gini impurity is biased towards preferring variables with more categories. Secondly, when any dataset has two or more than correlated features, any of those can be used as a predictor. But once any of them is used, the importance of others significantly reduces. As the first features already removed the impurity which had to be effectively removed.As in the first place we thought about reducing the features by half, 30 features were taken as the number of final features as any other number of features would give less effective result then.

After having all the important features in our hand, It is time to build our classifier. In our model, we used 4 different types of superior machine learning algorithms and for the implementation work, we used Python 3.6 as our programmable language. The algorithms we used are Logistic Regression, Support Vector Machine, K- Nearest Neighbor and Decision Tree. These algorithms are good for different classifications and they got their own properties and performance based on different datasets. As we said earlier, Logistic Regression and SVM are more commonly used algorithms for classification problems.

Logistic Regression: As out dataset contains multiple labels, this makes our problem a Multi-Class Classification problem. Logistic Regression is a Regression model based on some Statistical parameters. Here our output variables are categorical, unlike continuous values. For multiclass classification one vs all method has been used in our study.So, the values are passed to the argument called “Multiclass” in the constructor of the algorithm. A separate model is trained for each class based on the features to determine whether a prediction is under that class or not. Thus, that makes it a Binary Classification Problem. It is assumed that every class is independent. The 30 features extracted earlier is fed into the Logistic Regression hypothesis. The output obtained here is the probability expressing if the sample belongs to class 0 or 1.

Sigmoid activation: The Sigmoid curve looks like an S-shaped function. The reason for the use of sigmoid is that it exists between 0 and 1. It is therefore especially used for models in which we have to predict the probability as an output. Since there is only a probability of anything between 0 and 1, Sigmoid 's the right decision [14].

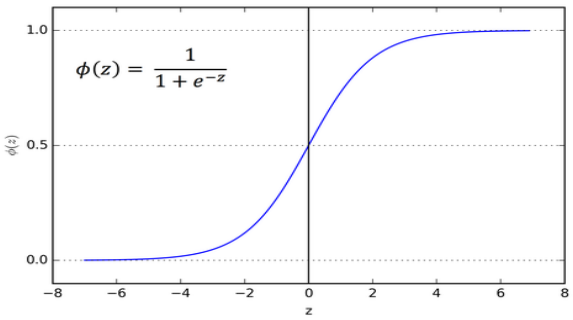

Figure 2. Sigmoid Function.

Support Vector Machine: Support Vector Machine is another superior classification algorithm which mainly draws a boundary between two classes to differentiate two different classes. When the number of features is a big number, SVM in cases like this sometimes outperforms Regression model and Neural Networks [13]. In our problem one vs rest scheme to determine the classes with the help of Linear SVC which has an alternative multiclass strategy.

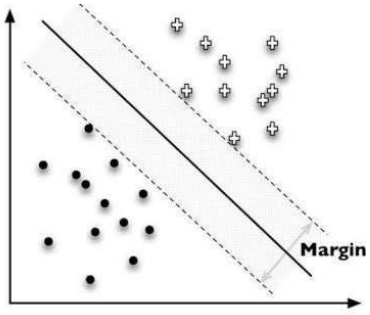

Figure 3. Support Vector Machine.

IV. IMPLEMENTATION

We split our dataset into 80% training data and 20% test data. Then A K-Nearest Neighbor Classifier was tested first on each of the features extracted earlier through the feature importance technique. It gives an underperforming 63% accuracy, but It gives an overwhelming score using all the features. Then A Decision Tree Classifier was performed on the train data and then evaluated based on the test data. Here in this case as our dataset contains multiple class levels, one vs all method was followed. This gave an accuracy of 72% which is slightly better than what we achieved using KNN model. Thirdly Logistic Regression Model was implemented to observe if there's any further scope of improvement over the results so far. A One vs all method was followed as the problem is multiclass classification. The logistic regression model achieved an accuracy of 87.8% or nearly 88%. As a result, it can be declared that the Logistic Regression Classifier is able to predict the DR level 88% accurately which is by far the best result so far.Lastly and finally Support Vector Machine model was trained with the trained data, to observe if SVM classifier is also able to perform at the same level as Logistic Regression. After applying some kernel tricks and regularization value of C equals to 5, an Accuracy of 88.3% was achieved which is undoubtedly the best-achieved result of our study. A sigmoid kernel was used to get classify the support vectors at this point.

So, it can be concluded that SVM and Logistic Regression performs better compared with the other classifiers that were used in our dataset.

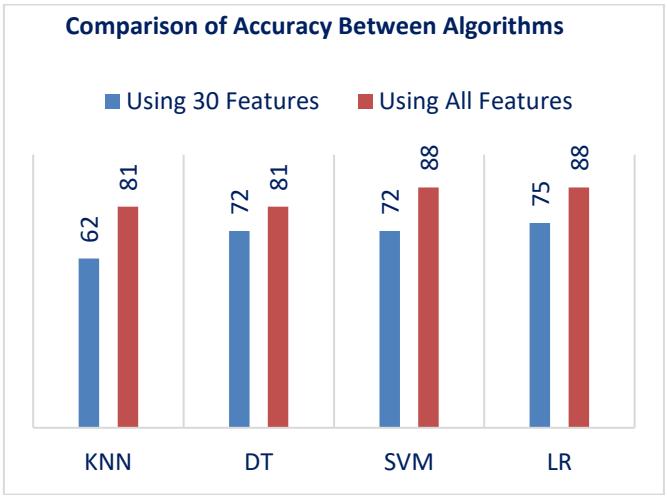

Figure 3. Accuracy comparison.

V. RESULTS AND DISCUSSION

The aim of our study was to build a classification model which can achieve an improved metric than the previous studies in this field. There has been an extreme number of researches that are still going on in the development of automated approaches of classifying early symptoms of Diabetic Retinopathy. To evaluate our work, we used 3 Popular Machine Learning Metrics which are Accuracy, Precision, and Recall.

Accuracy: Accuracy is the percentage of correct predictions out of the total number of predictions.

Precision:Precision is the ratio of correctly predicted positive results to the number of total predicted positive observation.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall: Recall is the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

The chart below shows the comparison between past studies and our proposed method.

TABLE I. PERFORMANCE OF THE PROPOSED SYSTEM WITH EXISTING WORK

Reference	No of classes	Method	Accuracy of Classification
Acharya et al. 2008 [10]	5	Higher order spectra	82%
Acharya et al. 2009 [13]	5	Blood Vessel, Exudates, Microaneurysms, cotton wool spots	86%
Proposed Method	6	Bright non-lesion, Hard exudates, Cotton wool spots, Red Non-lesion, Microaneurysms and Hemorrhages	88%

Also, in particular, our model was able to obtain a significantly higher precision and recall in hierarchical level 3 of our dataset using only 30 features compared to there all features.Using only 30 features our model has obtained a 92% precision and 97% recall. This means out of 100 positive identity's 92 were actually correct and not any other samples.

Furthermore, the recall explains if we take 100 samples of retinopathy having the symptoms of red nonlesions, 97 were actually predicted as red non lesion. This is notable as our model only needed 30 features to predict with such numbers of the result. Using all features can definitely raise the numbers a bit higher but would need more computation time.

TABLE II. PERFORMANCE OF THE PROPOSED SYSTEM IN TERMS OF PRECISION AND RECALL COMPARISON

Reference	Precision	Recall
Matthew et al. [12] using Decision Tree	0.72	0.626
Proposed Method using Logistic Regression	0.92	0.97

VI. CONCLUSION

The main objective of this research is to build an automated system model that can successfully detect the early non-

proliferative Diabetic Retinopathy (DR) symptoms among diabetes patients. DR is a disease that cannot be cured. To forestall permanent vision loss optical laser analysis is typically effective but it hampers the retina. Since decisive symptoms do not accumulate until the disease turns into inexorable, initial discovery via screening is mandatory. Therefore this research propsed an automated method to diagnosing and detect early stages of diabetic retinopathy e.g., microaneurysms, exudates, cotton wool spots, etc. based on supervised machine learning alagorithms. The proposed method also obtains a better result than existing approaches by introducing a tree-based feature selection method rather than irrelevant features-building of existing approaches [10,12,13]. The proposed work also achieves significant improvement in Precision and Recall in a hierarchy level over the existing works. All these tighter proves the supremacy of the proposed work to develop an automated tool for Diabetic Retinopathy.

VII. REFERENCES

[1] B. Wu, W. Zhu, F. Shi, S. Zhu, and X. Chen, "Automatic detection of microaneurysms in retinal fundus images," Computerized Medical Imaging and Graphics, vol. 55, pp. 106–112, 2017.

[2] R. Maher, S. Kayte, and D. M. Dhopeshwarkar, "Review of automated detection for diabetes retinopathy using fundus images," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 3, pp. 1129–1136, 2015.

[3] D. J. Browning, Diabetic retinopathy: evidence-based management. Springer Science & Business Media, 2010.

[4] R. Maher, S. Kayte, D. Panchal, P. Sathe, and S. Meldhe, "A decision support system for automatic screening of non-proliferative diabetic retinopathy," International Journal of Emerging Research in Management and Technology, vol. 4, no. 10, pp. 18–24, 2015.

[5] B. Singh and K. Jayasree, "Implementation of diabetic retinopathy detection system for enhance digital fundus images," International Journal of advanced technology and innovation research, vol. 7, no. 6, pp. 874–876,

[6] N. Thomas and T. Mahesh, "Detecting clinical features of diabetic retinopathy using image processing," International Journal of Engineer-ing Research & Technology (IJERT), vol. 3, no. 8, pp. 558–561, 2014.

[7] M. Gandhi and R. Dhanasekaran, "Diagnosis of diabetic retinopathy using morphological process and SVM classifier," in Communications and Signal Processing (ICCSP), 2013 International Conference on. IEEE, 2013, pp. 873–877.

[8] E. M. Shahin, T. E. Taha, W. Al-Nuaimy, S. El Rabaie, O. F. Zahran, and F. E. A. El-Samie, "Automated detection of diabetic retinopathy in blurred digital fundus images," in Computer Engineering Conference (ICENCO), 2012 8th International. IEEE, 2012, pp. 20–25.

[9] Nayak, J., Bhat, P. S., Acharya, U. R., Lim, C. M., and Kagathi, M., Automated identification of different stages

[10] Acharya, U.R., Tan, P. H., Subramaniam, T., Tamura, T., Chua, K. C., Goh, S. C., Lim, C. M., Goh, S. Y., Chung, K. R., and Law, C. Automatic Identification of diabetic type 2 subjects with and without neuropathy using wavelet transform on pedobarograph. J. Med. Syst. 32(1):21–29, 2008.

[11] Lee, S. C., Lee, E. T., Wang, Y., Klein, R., Kingsley, R. M., and Warn, A., Computer classification of nonproliferative diabetic retinopathy. Arch. Ophthalmol. 123(6):759–764, 2005.

[12] Bihis, Matthew; Roychowdhury, Sohini, "A generalizedflow for multi-class and binary classification tasks: An Azure ML approach," in Big Data (Big Data), 2015 IEEE International Conference on , vol., no., pp.1728

[13] L. Chen and L. Chen, "Support Vector Machine - Simply Explained," Towards Data Science, 07-Jan-2019. [Online]. Available: https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496. [Accessed: 12-May-2019].

[14] Logistic Regression for Machine Learning," Machine Learning Mastery, 06-Apr-2019. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/. [Accessed: 12-May-2019]

[15] S. M. Mazinani and K. Fathi, "Combining KNN and Decision Tree Algorithms to Improve Intrusion Detection System Performance," International Journal of Machine Learning and Computing, vol. 5, no. 6, pp. 476–479, 2015.

[16] Diving into data. [Online]. Available: https://blog.datadive.net/selecting-good-features-part-iii-random-forests/. [Accessed: 12-May-2019].

[17] H. Abedy, F. Ahmed, M. N. Q. Bhuiyan, M. Islam, M. N. Y Ali, and M. Shamsujjoha "Leukemia Prediction from Microscopic Images of Human Blood Cell Using HOG Feature Descriptor and Logistic Regression", 16th Int'l Conf. ICT & Knowledge Engineering, pp: 1-6, year: 2018

[18] N. Ahmed, M. Shamsujjoha, M. N. Y. Ali, and W. Marsh, "An Efficient REDCap Bassed Data Collection Platform for the Primary Immune Thrombocytopenia and Its Analysis Over the Conventional Approaches", 18th Int'l Conf. on Computer and Information Technology, pp:353-357, year: 2015.

[19] M. Shamsujjoha, M. S. Ahmed, F. Hossain, and T. Jabid, "Semantic Modelling of Unshaped Object: An Efficient Approach in Content Based Image Retrieval", 17th Int'l Conf. on Com. and Inf. Technology, Dhaka, Bangladesh, pp:30-34, year: 2014

[20] M. Shamsujjoha, and T. Bhuiyan, "A Content-based Image Retrieval Semantic Model for Shaped and Unshaped Objects", Journal of Computer Engineering, vol: 18, no: 1, pp:43-60, year: 2016