International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207132

A Survey on Diagnosis and Analysis of Diabetic Retinopathy using Feature Selection

Amalu Michael[1], Deepa S S[2]

[1]M. Tech Scholar, Department of Computer Science and Engineering, Government Engineering College Idukki, India

[2] Associate Professor, Department of Computer Science and Engineering, Government Engineering College Idukki, India

ABSTRACT

Diabetic retinopathy is one of the common forms of diabetic eye disease. DR occurs due to a high ratio of glucose in the blood, which causes alterations in the retinal vessels. Machine learning may be a broad multidisciplinary field that has its roots in statistics, algebra, data processing, and information analytics, etc. Machine learning is used to discover patterns from medical data and provide an efficient way to predict diseases.ML is an application of artificial intelligence it collects information from training data. There are several machine learning techniques are used for the diagnosis of diabetic retinopathy. This paper mainly focuses on the survey of such techniques and also various feature selection mechanisms. This study provides the basic categorization of feature selection techniques and discussing their use.

Keywords : Machine Learning, Diabetic Retinopathy, Feature Selection

I. INTRODUCTION

Nowadays, Diabetic retinopathy (DR) is one of the eye diseases that occurs due to diabetes mellitus and it has grown as the most common cause of blindness. Symptoms of DR disease include blurred vision, difficult to seeing colors, floaters, and even total loss of vision. The DR disease is mainly divided into Non-Proliferative Diabetic Retinopathy (NPDR) is a milder kind of diabetic retinopathy which is usually symptomless and Proliferative Diabetic Retinopathy (PDR) is the advanced stage of diabetic retinopathy which refers to the formation of abnormal blood vessels in the retina. Therefore regular screening of diabetic patient's medical details is a very essential and automated or computerized analysis of diabetic patient's medical data that can help eye care.

An Electronic Health Record (EHR) is a systematic collection of patient's medical information in digital format and it can be shared across different healthcare applications. Combining multiple types of clinical health record help the doctors to easily identify chronic diseases. The algorithms of Machine learning are useful in identifying complicated patterns within prosperous and huge data. This facility is especially well- suited to clinical applications, particularly those people who rely on advanced genomics and proteomics measurements. It is often used in numerous illness diagnosis and detection. In medical applications, ML algorithms will create higher decisions regarding treatment plans for patients by implementing a useful health-care system. Machine Learning (ML) has an effective medical diagnosis system. In medical diagnosis, a set of features represent all variations of diseases.

Feature Selection is the process which automatically or manually selects the most relevant features for the medical prediction or diagnosis.

The remaining part of the paper is organized as follows. In Section 2, the application of machine learning in healthcare is discussed; Section 3 discusses various feature selection mechanisms; Section 4 deals with summarizing the related work done under the scope of this paper; Section 5 concludes the paper.

II. MACHINE LEARNING IN HEALTHCARE

ML is an example of artificial intelligence which collects information from training data. It plays a vital role in many fields including finance, medical science, and security. Machine learning is used to discover patterns from medical data and these patterns are used to predict various diseases. Machine Learning is divided into the following categories:

Supervised Learning: This learning is taken under training and algorithm develops an exercise that matches inputs to related outputs. One common establishment of the supervised learning task is the classification issue.

Unsupervised Learning: In Unsupervised learning, the class labels are not known in the training phase. The technique used for unsupervised learning is clustering, fuzzy clustering, hierarchical clustering, K means clustering, association rule mining. Here a cluster is formed on the availability of trained data with unknown labels. These algorithms are used for developing a framework with the help of data samples. Semi-supervised Learning: It is the method of identifying the best classifier from each unlabeled and labeled information. By using unlabeled information it transfers high performance of classification. The success of this method depends on a few underlying assumptions.

Reinforcement Learning: In this learning, access is given by the computer program to the dynamic environment for performing a specific goal. Feedback in terms of rewards and punishments is provided to the program if it navigates its drawbacks.

III. MACHINE LEARNING TECHNIQUES

A. Decision Tree

Decision Tree is a supervised learning technique, which is used for solving classification problems. In the Decision tree, the given dataset iteratively divided into two or more sample data. The main goal of this method is to predict the class value of the target variable. The decision tree will help to isolate the data set and builds a decision model to predict the unknown class labels. A decision tree can be created by using both binary variables and continuous variables. The decision tree finds the root node based on the highest entropy values of variables. This gives the decision tree an advantage of selecting the most consistent hypothesis among the training dataset. The input to the decision tree is a dataset, which consists of several features and instance values and the output will be a decision model. The problems faced while building a decision model are selecting the splitting attribute, splits, stopping criteria, pruning, training sample, quality and quantity, the order of splits, etc.

The decision model is a graph structure, where the structure includes a collection of nodes. It includes decision nodes (with the condition) and leaf nodes. Among various variables in the dataset, choosing the right variable as the root node to start the break down is the difficult task. The decision node has 2 or more children. The model predicts the best variable as the root node or best predictor node from the set of nodes available. There are many ways to choose the best attribute as the root node, based on the degree of impurity of the child nodes. Entropy, Gini index, Mutual Information, etc, are used as the performance measures.

B. Random Forest:

Random forest is a supervised learning technique used for both classification and regression problems. The main logic of the random forest is bagging method to generate random samples of features. The main difference between the decision tree and the random forest is the process of finding the root node and splitting the feature node will run randomly.

C. Naive Bayesian:

A probabilistic classifier that is based on Baye's theorem with the independence assumption between the predictors. Naive Bayesian method takes the dataset as input, performs analysis and predicts the class label using Baye's Theorem. It calculates the probability of a class in input data and helps to predict the class of the unknown data sample. This technique is well suitable for large datasets. The Baye's Theorem formula computes the posterior probability of each class by using the formula below.

Where P(c|x) is the posterior probability of class (target) given predictor (attribute). P(c) is the prior probability of class. P(x|c) is the likelihood which is the probability of predictor given class. P(x) is the prior probability of predictor.

D. Support Vector Machine:

It is a supervised learning, discriminative classification method. This method can be used for both regression and classification problems. The logic behind the SVM is to find a hyper line between the dataset, which divides the dataset into two subclasses. SVM process includes 2 steps, identifies the optimal hyper line in data space and mapping the objects to the boundaries specified. The

SVM training algorithm builds a model that assigns new data samples to one of the two classes.

E. K-nearest neighbor (KNN):

KNN is a classification technique that classifies the new data samples based on similarity measure or distance measure between them. The KNN includes 3 distance measure which are Euclidean distance, Manhattan, Minkowski.

IV. FEATURE SELECTION

Machine Learning is effectively used as an effective medical diagnosis system. In the diagnosis system, a set of features are used to represent all variations of diseases. Feature selection is the process of selecting relevant features which contribute most to the model. Feature selection process approaches are[6] Filter approach, Wrapper approach, Embedded approach, and Hybrid approach.

Filter approach: Filter techniques assess the relevance of features by the intrinsic properties of the data. In most cases, a feature relevance score is calculated, and features with low-score are removed. After this, a subset of relevant features is presented as input to the classification algorithm. The filter techniques are useful for high- dimensional datasets, and also they are computationally simple and fast. Filter approaches are independent of the classification algorithm. The feature selection needs to be performed only once, and then different classifiers can be evaluated. A common disadvantage of filter methods is that they are independent of the classifier. Examples of filter methods are Euclidean distance, i-test, Information gain, Correlation-based feature selection (CFS), Fast correlation-based feature selection (FCFS), etc.

Wrapper approach: These methods are based on algorithms which evaluate all possible subset of features and select the subset which produces a better result for the specific machine learning problem. This method selects the useful features based on the classifier. The main advantage of the wrapper approach is, it provide better accuracy than a filter approach. But wrapper approach is more expensive than the filter approach due to the repeated evaluation of features. Examples of wrapper methods are Sequential forward selection (SFS), Genetic

Algorithms, Simulated annealing, Randomized hill climbing, etc.

Embedded approach: Embedded method is the seek for an optimal subset of features that are made into the classifier construction and might be seen as an exploration within the combined space of feature subsets and hypotheses. Rather than wrapper approaches, embedded approaches are thus specific to a given learning algorithm. The advantages of the Embedded approaches are: they depend on the classification model and less computationally intensive than wrapper methods. Examples are Feature selection using the weight vector of SVM, Decision trees, Weighted naive Bayes, etc.

V. RELATED WORKS

Literature Review

Raid Alzubi et, al. [2] proposed a Hypertension Disease Detection System using Machine Learning Techniques. Machine learning and data mining techniques are here used for the analysis of the human genome. One of the sources of human genome variation is Single Nucleotide Polymorphisms (SNPs), which have been associated with various chronic diseases such as heart disease, cancer, etc. Various machine learning techniques have been developed to differentiate between affected and healthy samples of SNP data. In this work, they use filter approach using Conditional Mutual Information Maximization (CMIM) method to find a subset of the most informative SNP samples to be used in various classification algorithms for the diagnosis of hypertension disease. Four classification algorithms, such as k-Nearest Neighbours (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) are evaluated here. Among that, SVM achieved the highest classification accuracy (93.21%)

Sakshi Gujral et, al [3] introduces an Early Diabetes Detection system using Machine Learning techniques. The data set used in this work is the Pima Indian Diabetic Data Set. Early diabetes detection is helpful to scale back the effects of diabetes. Various machine learning techniques such as artificial neural network, principal component, decision trees, genetic algorithms, Fuzzy logic are used for the detection of diabetic disease and their results are compared. They conclude that hybrid approaches provides better results than single classifiers.

Syed Sifat Rahman et, al [4]: proposed a machine learning Based Approach for early Diabetes Detection. This system can discover a person who has diabetes. This paper uses two different algorithms KNN and K-means on a supervised dataset to detect diabetes. It also shows nearby doctors by location tracking method. Through this system, a person can easily know if he has diabetes by giving their test reports and consult with nearest certified doctors by location tracking method. As a result, this method saves time and money to detect diabetes and to find preferable doctors as one does not have to go to doctors. As a result, KNN gives better accuracy than KMeans.

Tao Zheng et, al. [5]: introduces a Machine Learning-based framework to identify Type 2 Diabetes through Electronic Health Records is proposed. Their system is used to discover diverse genotype-phenotype associations affiliated with Type 2 Diabetes Mellitus (T2DM) via genome-wide association study (GWAS) and phenome- wide association study (PheWAS), more cases (T2DM subjects) and controls (subjects without T2DM) are required to be identified (e.g., via Electronic Health Records (EHR)). However, existing expert-based identification algorithms often suffer in a low recall rate and could miss a large number of valuable samples under conservative filtering standards. The goal of this work is to develop a semi-automated framework based on machine learning as a pilot study to liberalize filtering criteria to improve recall rate. Then, it uses DNA

complementary rules to encode an image and uses the key and DNA to compute each pixel value. It has a good encryption effect and security.

N. Sneha et, al [6]: Proposes a system for the analysis of diabetes mellitus for early prediction using optimal features selection. The constant hyperglycemia of diabetes is related to long-haul harm, brokenness, and failure of various organs, particularly the eyes, kidneys, nerves, heart, and veins. The objective of this research is to make use of significant features, design a prediction algorithm using machine learning and find the optimal classifier to give the closest result comparing to clinical outcomes. The proposed method aims to focus on selecting the attributes that ail in early detection of diabetes mellitus using predictive analysis. This approach includes the selection of the right attributes from the large database, based on the sensitivity of the dataset and the problem statement. The selection of optimal attributes for the problem, it requires an overall analysis of the attributes and ignoring the irrelevant attributes.

Peter B. Jensen et, al [8]: discusses the basic of Electronic Health Records and their research applications in medical care. Mining of Electronic Health Records (EHRs) has the potential for establishing new patient stratification principles and for revealing unknown disease correlations. Integrating EHR data with genetic data will also give a finer understanding of genotype- phenotype relationships. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining. Here, they consider the potential for furthering medical research and clinical care using EHR data and the challenges must be overcome. EHR is useful in Clinical Text Analysis, Knowledge Extraction, Prediction from data.

Karan Bhatia et, al [9] introduces a system for the diagnosis of diabetic retinopathy using classification algorithms. Detection of diabetic retinopathy in an early stage is essential to avoid complete blindness. Many physical tests like visual acuity test, pupil dilation, optical coherence tomography can be used to detect diabetic retinopathy but are time-consuming and affects patients as well. This paper focuses on the decision about the presence of disease by applying the ensemble of machine learning classification algorithms on features extracted from different retinal images. The features are the diameter of the optic disk, lesion-specific (microaneurysms, exudates), image-level (prescreening, AM/FM, quality assessment). Decision making for predicting the presence of diabetic retinopathy was performed using an alternating decision tree, AdaBoost, Naïve Bayes, Random Forest, and SVM. This paper highlights various technologies used for the diagnosis and detection of diabetic eye disease.

Dhiravidachelvi, E. et, al [10] proposes a Novel Approach for Diagnosing Diabetic Retinopathy in Fundus Images. This approach consists of 4 functional steps. The first step is to read an image I, where I is a Retinal image and register the image for correcting the alignment of the Image. Once the correct output is received, the input image is preprocessed, by removing the noise using the salt and pepper noise removal method and enhancing the

image by pixel level. The second step is to calculate the SIFT feature point by matching the input image and the template images and then the histogram value of input image and the template image and calculate the number of connected components of the input image and the template image by using morphological operations. This comparison output score decides whether the input image is a normal image or an abnormal image. Then, the image is checked to see whether it is affected by NPDR or PDR and to also determine the stage of the NPDR or PDR whether it is mild, moderate or severe. Once the stage of the NPDR or PDR is determined then the treatment is suggested.

Swati Gupta [11] introduces an automatic detection of diabetic retinopathy through computational techniques that would be a great remedy. There are many features present in retina like exudates and microaneurysm features. The presence of microaneurysms (MAs) is usually an early sign of diabetic retinopathy and their automatic detection from color retinal images somewhat tough job, so for that, we are using green Chanel images. The objective of this project is to detect retinal microaneurysms and exudates for automatic screening of DR using classifiers. To develop an automated DR screening system detection of dark lesions and bright lesions in digital funds photographs is needed. To detect retinal microaneurysms and exudates retinal fundus images are taken from the Messidor dataset. After pre-processing, morphological operations are performed to find the feature and then features are get extracted such as GLCM and Splat for classification.

R. Priya, et al. [12] introduces a system for the diagnosis of diabetic retinopathy using machine learning techniques. In this paper, to diagnose diabetic retinopathy, three models such as Probabilistic Neural network (PNN), Bayesian Classification and Support vector machine (SVM) are used and their performances are compared. The amount of the disease spread in the retina can be identified by extracting the features of the retinal images. The features like blood vessels, Hemorrhage of NPDR image and exudates of PDR image are extracted by using many image processing techniques and then the extracted features are fed to the classifier. Around 350 fundus images were used, out of which 100 images were used as a training set and remaining images were used for the testing phase. Experimental results show that PNN has an accuracy of 89.6 % Bayes Classifier has an accuracy of 94.4% and SVM has the highest accuracy as 97.6%.

Yunlei Sun et, al [13] introduces a system for diagnosis and analysis of diabetic retinopathy based on Electronic Health Records. In previous research, works uses various learning techniques to detect diabetic retinopathy in patients with medical images. The medical imaging achieves reasonable recognition accuracy, the application of lowcost easy-to-obtain free electronic health records (EHR) data in life can make an early diagnosis of the DR more convenient and quick. In this paper, they use five machine learning models such as SVM, LR, DT, RF, NB to detect the DR with the EHR data and formed a set of treatment methods. During data pre-processing, they consider that feature engineering may improve DR disease diagnosis accuracy for patients by removing difficult to handle text data.

VI. CONCLUSION

Machine learning is useful in the medical field to discover patterns from medical data and provide excellent proficiency to predict diseases. Detection of retinopathy in the early stage is an efficient way to avoid complete blindness. Many physical tests like visual acuity test, pupil dilation, optical coherence tomography are used to detect diabetic retinopathy but they take more time and also affect patients as well. Preceding works focus on the decision about the presence of disease by applying various machine learning classification algorithms on features extracted from different retinal images. Mining of Electronic Health Records (EHRs) has the potential for building up new persistent stratification standards and for uncovering obscure illness correlations. Because EHR data is also very easy to get, low-cost data, it only needs to be extracted from the daily physical examination report after simple processing. Different machine learning algorithms are used to predict the disease and the accuracy of each method is based on the feature of the dataset.

VII. REFERENCES

[1]. K. Shailaja et, al, "Machine Learning in Healthcare: Review" 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018)

[2]. Raid Alzubi et, al "Hypertension Disease Detection via Machine Learning Techniques "IEEE Transactions on Neural Networks and Learning Systems 2018

[3]. "Early Diabetes Detection using Machine Learning" International Journal for Innovative Research in Science & Technology 2017

[4]. Syed Sifat Rahman et, al "A Machine Learning-Based Approach for Diabetes Detection and Care" 2nd International Journal on Next Generation Computing Technologies (NGCT-2016)

[5]. Tao Zheng et, al "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records" International Journal of Medical Informatics 2016

[6]. N. Sneha et, al "Analysis of diabetes mellitus for early prediction" Journal of Big data 2019

[7]. Yvan Saeys et, al "A review of feature selection techniques in bioinformatics" International Conference on bioinformatics 2016

[8]. Peter B. Jensen et, al "Mining electronic health records: towards better research applications and clinical care" 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018)

[9]. Karan Bhatia et, al "Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm" International Conference on Next Generation Computing Technologies (NGCT-2016)

[10]. Dhiravidachelvi, E. et, al "Diagnosing Diabetic Retinopathy in Fundus Images" Journal of Computer Science 2018

[11]. Swati Gupta et, al "Diagnosis of Diabetic Retinopathy using Machine Learning" Journal of Research and Development 2015

[12]. R. Priya et, al "Diagnosis Of Diabetic Retinopathy Using Machine Learning Techniques" ICTACT Journal On Soft Computing, July 2013, Volume: 03, Issue: 04

[13]. Yunlei Sun et, al "Diagnosis And Analysis Of Diabetic Retinopathy Based On Electronic Health Records" Special Section On Healthcare Information Technology For The Extreme And Remote Environments 2019

Cite this article as :

Amalu Michael, Deepa S S, "A Survey on Diagnosis and Analysis of Diabetic Retinopathy using Feature Selection", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 1, pp. 170-176, January-February 2020. Available at doi : https://doi.org/10.32628/IJSRSET207132

Journal URL : http://ijsrset.com/IJSRSET207132