MACHINE LEARNING MODELS FOR THE DETECTION OF HUMAN EYE DISEASE

K. Arunkumar

Department of Computer Science, Annai Vailankanni Arts and Science College, India

Abstract

Glaucoma is a human eye illness that causes Irish-eye injury and ultimately can lead to full blindness in patients with diabetes. Glaucoma detection is vital at an early stage to prevent full blindness. Even if early diagnosis and persistent monitoring of the diabetes patient are required, effective Glaucoma treatment is available. Many physical tests can also be done to detect Glaucoma, but they are time-consuming, including visual acuity testing, pupil dilation and optical coherence tomography. The aim of the study was to decide on glaucoma by using machine learning ensemble classifying algorithms based on data taken from diverse Irish images. It gives us the precision of which algorithm is suitable and precise for disease prediction. The Random Forest, K-Nearest Neighbor, neural networks, and the support vector machine are responsible for the determination to glaucoma.

Keywords:

Random Forest, K-Nearest Neighbor, Neural Networks and Support Vector Machine

1. INTRODUCTION

Diabetes is a condition that happens chronically in an organ when insulin is not secreted or the body cannot effectively treat the pancreas. Glaucoma is a medical disorder in which the Irish have been injured by blood vessel fluid leaking into the Irish. It is one of the most prevalent disorders of the diabetic eye and a major cause of blindness. Close to 415 million people with diabetes are in danger of blindness. This develops in the light-sensitive tissue at the rear of the eye when diabetes affects the little blood vessels. This miniscale blood vessel leaks blood and fluid in the Irish forms of micro-aneurysms, blood bleeding, hard exudates, spots of cotton wool or vein loops.

The stages of DR can be defined depending on the presence of Irish characteristics. At NPDR, the condition may progress from mild, moderate to serious, with different levels of characteristics except less blood vessel formation. PDR is the advanced phase where new blood vessels are triggered by fluids supplied for food by the Irish. They develop along the ocular surface and over the transparent glass gel that fills the eye. A serious loss of vision and perhaps blindness may follow if they leak blood.

DI detection is now a time-consuming and tedious method which requires a professional clinician to analyse and evaluate pictures of the Irish digital fundus. When people offer their opinions, frequently a day or two later, the retarded results lead to lost monitoring, misunderstandings and delayed treatment.

The research focuses mostly on glaucoma prediction and analysis of several prediction algorithms. A training course can be trained using machine learning algorithms such as KNN, RF, SVM, NNET etc., and these can predict data by comparing data provided with training datasets. Then these algorithms may be predicted. The aim is to train the programme utilising training datasets to detect glaucoma using multiple classification algorithms.

2. BACKGROUND

Machine learning, an artificial intelligence discipline, involves system building and studying which can learn from data [4]. Calculation methods are used by the machine learning algorithm to "learn" data directly without relying upon a model equation. As the number of examples available for learning increases, the algorithms enhance their performance. With Mitchell, the definition is more commonly quoted and formal:

A computer programme has experience E of certain classes of T-tasks and measurements P if its performance at T-tasks is enhanced by the experience E, as measured by P [5].

The machine learning core is representation and widespread. All machine-learning systems are represented by the data instances and functions evaluated in these instances. Generalization is the ability of a machine learning system to operate accurately with new, invisible data incidents after learning data. The training examples originate from the generally unknown distribution of probabilities, and the pupil must create a general model for this space, so that adequate predictions may be produced in fresh circumstances. Generalization performance is usually 11 in relation to the capacity to reproduce known information from newer examples. Machine learning is varied, although the two primary types are: supervised learning and uncontrolled learning.

2.1 SUPERVISED LEARNING MODEL

Supervised learning is a master learning job that uses supervised training data to determine a function [6]. The supervised learning training data provides a number of instances with matched input topics and desires. A supervised learning algorithm analyses the training information and creates a function known as a classification or regression function. For any valid input item, the function should predict the proper output value. The learning algorithm demands that the training data be generated reasonably in unseen conditions.

The relationship between a student and a teacher is a straightforward parallel to guided learning. The teacher initially teaches a certain subject to the learner. Teach the pupils the concepts of the subject and answer numerous questions about the subject. The teacher then sends the student to take an examination paper in order to answer updated questions.

In Fig.1, the system learns from data supplied containing both the functionality and the output. After learning, more recent data is delivered without outputs, and the system produces the output with the expertise it has acquired from its data. This is how the supervised model of learning works.

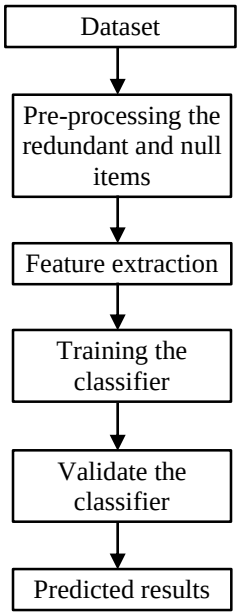

Fig.1. Workflow of supervised learning model

3. ALGORITHMS

Since there are so many machine learning algorithms, they cannot all be used for analytical purposes. For this research paper, the study will be using four of them neural networks (NNET), random forest (RF), K-Nearest Neighbor (KNN) and support vector machine (SVM).

3.1 NEURAL NETWORKS

In the machine-learning discipline, a subset of algorithms are embedded into three or more layers of artificial neurons [10]. There are several additional machine learning models that are remarkable for their adaptability. Each node of the neural network has its own area of knowledge about rules and features, which allows us to grow the neural networks through experience from earlier procedures. Neural networks are particularly adapted to identifying non-linear patterns, like in patterns where the link between input and output cannot be straight one-to-one [8]. It is an apprenticeship. Neural networks have adaptable weights along routes between neurons, which can be tailored to an educational algorithm that learns from data observed to enhance the model. A suitable cost function must be chosen. The cost function is the one utilised to learn how to solve the problem optimally [7]. In a word, it can adjust itself to the changing environment by learning from initial training and providing more knowledge about the world afterwards.

3.2 RANDOM FOREST

For classification and regression, a random forest approach can be used. Classification algorithms are supervised and a forest with a number of trees is created [9]. Generally speaking, the more trees, the healthier the forest seems. The greater the number of trees in the forest, the more accurate the results can be claimed to be. Random forest algorithms have several advantages. The classifier can process the values that are missing. The random forest classification can also be modelled for category values [2]. The overriding problem will never occur if the study of any classification problem uses a random forest technique. More crucially, functional engineering functionality may be used to determine the most important characteristics from the training dataset [4].

3.3 K-NEAREST NEIGHBORS

K-nearest Neighbors is a straightforward algorithm, which saves all available examples and classifies new cases on an analogy measure [11]. KNN was used to estimate statistics and to recognise patterns. KNN predicts a new instance (x) by searching for k most similar examples through the complete training set and resuming the output variable in those k cases. For a regression, this might be the medium output variable. For a classification, it might be the mode class to determine which of the k-sets in the training data set are the most similar to the new entry.

3.4 SUPPORT VECTOR MACHINE

A SVM generates a high-or infinite-dimension space hyperplane or a series of hyperplanes that can be employed in classification, regression, or other tasks. The hyperplane with the biggest distance to the closest data point of any class intuitively achieves a decent separation since the higher the margin, the smaller the general error of classification [1].

SVMs are in the generic kernel category. An algorithm which simply relies on the data via dot products is a kernel method. If so, a kernel function that calculates a dot product inside a possibly high-dimensional range can replace the dot product. There are two benefits: Firstly, the capacity to produce non-linear limits of prediction using linear classification methods. Second, by using kernel functions, a classier can be applied to data that does not have a clear representation of a fixed-dimensional vector space [4].

4. PROPOSED MODEL FOR PREDICTION

This section provides suggested models, datasets, descriptions, data visualisation and algorithms for analytical performance classification.

4.1 PROPOSED MODEL

Data collection is the first phase. The study collected data from the UCI Machine Learning website. The dataset includes the characteristics collected from the image of Messidor to forecast the existence of a glaucoma image or not. The data set will next be recognised with features and labels. Then the dataset is divided into two sets, one of which is used for workouts in which most of the data and the other tests. For the analytical performance of the model, the training set consisted of four distinct classification algorithms. The k-Nearest Neighbor, random forest, support vector machine and neural networks are the algorithms employed. After training datasets are learned by the system, new data is delivered without outputs. With the information gathered from the data she was trained with, the final model provides the output. In the last phase, the study will determine the accuracy of each algorithm and find out which specific algorithm will provide us with more accurate findings for glaucoma prediction.

5. IMPLEMENTATION

5.1 DATA COLLECTION

In the project, the study employed the UCI Machine Learning Repository data set. This dataset contains features taken from the image of Messidor to determine whether or not an image shows indications of glaucoma. Either a lesion that has been discovered, an anatomical portion or a picture-level descriptor are represented in all characteristics. In order to facilitate computer assisted diagnostics of Irish diabetes, the Messidor database has been built. The investigation saw many kinds of datasets used for various kinds of projects using Kaggle, Github, and other websites based on diabetes. Since the study sought to detect glaucoma, this set of data is suitable for the purpose because it has different types of characteristics.

5.2 DATA DESCRIPTION

The dataset includes many feature categories which are extracted from the image collection of Messidor. It used to predict if an image has or does not have symptoms of glaucoma. The value of diabetic patients in this case reflects a different Irish perspective. The first 19 columns of the dataset are separate variables, while the last column is the column or variables. Binary numbers represent outputs. "1" implies glaucoma patients, and "0" signifies the disease is not present. The study also calculated the number, mean, and max of the values standard in the dataset.

5.3 DATA VISUALIZATION

The skewness of each class is another key element in the data distribution. Visualization of data allows you to view how the data appears and how the data correlates in the study. A histogram represents accurately the distribution of numerical data. The probability distribution of a continuous variable is estimated. Histograms are an excellent way to learn the data. You can observe quickly where you can find a large and small amount of data.

5.4 SPLIT DATASET

Data is a key factor in the evaluation of data mining models when separated into training and test settings. In general, the majority of data is utilised for training when partitioning a data set into two parts, and a lesser proportion is used for testing. The study divided the dataset into two different sets. The first is for training and the second is for testing. The training set has a known output, and the model learns about this data so that it can be spread to other data. The study assessed the model by predicting the test set once the model was processed using a training set. Since the data in the test set already contains known values for the characteristics to be predicted in the study, it is simple to determine whether or not the assumptions in the model are right. Furthermore, 80% of the data was used for training and 20% for testing.

5.5 MACHINE LEARNING ALGORITHM

The study has undergone a trial and error procedure to settle on a small list of algorithms which yield better results, as the study works on glaucoma classification, and certain classification algorithms have been employed for the study. The study has a concept of which methods are adequate for the classification challenge in the data visualisation plots. The training data for modelling is used by the Machine Learning system, while test data is used to assess the prediction quality of the learned model. The machine learning system analyses forecast performance by comparing data sets with real values (called ground-reality) predictions using a range of measurement techniques.

Thus, four alternative machine-learning algorithms are evaluated in the study.

- Neural Networks (NNET)
- Random Forest
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)

5.6 K-FOLD CROSS VALIDATION

K-Fold Cross Validation is a popular type and is commonly used in computer training. Cross validation. The original sample is divided into sub-samples for k-fold cross-validation. Of the k sub-examples, the validation data for the model and the remaining k-1 sub-examples are maintained as training data. During the investigation, the validation was 10-fold. The advantage of this strategy is that it uses all observations both for training and validation and validates each observation exactly once.

6. EXPERIMENTAL RESULTS AND ANALYSIS

The study discussed the suggested system and the study in the previous chapter. The study showed how the data set, data set description, visualisation and algorithms were collected by the study. The paper now discusses the outcomes of the trials on the implementation of this system. The study split the data into two components: a dataset for training and testing. The results of the training and testing dataset are illustrated in this chapter. Four machine learning algorithms have been utilised, as described before. First, employing the four algorithms, the study trained the dataset and then a model was developed. The test data set of this model was then tested by the study. If the accuracy of the test set is close to the accuracy of the training, the study could conclude that a decent model was developed.

The study has a total of 1151 data points from different people. The dataset consists of 1151 rows and 20 columns. The study now consists of 920 rows for train data and 231 rows for test data after the division of the data into two sections. The train data was trained for the analysis of several algorithms during the investigation. This is the outcome of the study.

The study used for the training data set is a comparison between algorithms. Here, the high line shows the standard deviation and the median value is given by the rectangular box and the brown line in the box gives the mean value. From here, the study can understand the model's good algorithm.

The model is tested with the test dataset following the model training. The study has 20% of test data in the test set. The accuracy of the test, precision, recall and F1 results are shown in Table.1. The details of the test data assessment with the unigram model are as follows:

Table.1. Accuracy of test dataset

Models	Accuracy	Precision	Recall	F1 Score
SVM	57.07%	62%	57%	53%
KNN	64.50%	65%	65%	65%
RF	63.63%	64%	64%	64%
NNET	75.32%	78%	75%	75%

The objects of history are retrieved from calls to the model-training function. In the history part of an object returned, metrics are saved in a dictionary.

7. CONCLUSION

In an experimental investigation, the precision of both the training and the testing sets is relatively similar, and the NNET algorithm for both training and testing gives a greater accuracy rate of about 75%. The study can therefore state that this method gives us a better understanding of the condition. As the study is primarily aimed at building a model to as accurately as possible the glaucoma, the study expects that we will have the right results from that end model.

The study also measured the accuracy and loss of the train and test model. The study utilised keras to obtain this train for this visualisation model and test loss and accuracy. For this reason, the study also utilised a history callback. The History callback is one of the default callbacks recorded while training all profound learning models. For each epoch, it records training measures. This covers loss and precision in the test data set (for classification tasks), as well as loss and accuracy.

REFERENCES

[1] T. Zheng, W. Xie and L. Xu, L., "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records", *International Journal of Medical Informatics*, Vol. 97, pp. 120-127, 2017.

[2] K. Plis, R. Bunescu and C. Marling, "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-12, 2014.

[3] B. Patra, L. Jena and S. Bhutia, "Evolutionary Hybrid Feature Selection for Cancer Diagnosis", *Proceedings of International Conference on Intelligent and Cloud Computing*, pp. 253-263, 2021.

[4] S.A. Kaveeshwar and J. Cornwall, "The Current State of Diabetes Mellitus in India", *Australasian Medical Journal*, Vol. 7, No. 1, pp. 45-48, 2014.

[5] L. Jena, S. Nayak and R. Swain, "Chronic Disease Risk (CDR) Prediction in Biomedical Data using Machine Learning Approach", *Advances in Intelligent Computing and Communication*, Vol. 109, pp. 1-13, 2021.

[6] Maryjo M George and S. Kalaivani, "Intensity Inhomogeneity Correction and Tissue Segmentation of MR Images: A Parametric Approach", *International Journal of Pure and Applied Mathematics*, Vol. 115, No. 9, pp. 409-416, 2017.

[7] Nida M. Zaitoun and Musbah J. Aqel, "Survey on Image Segmentation Techniques", *Proceedings of International Conference on Communication, Management and Information Technology*, pp. 797-806, 2015.

[8] Hui Liua, Shanshan Liu, Dongmei Guo, Yuanjie Zheng, Pinpin Tanga and Guo Dan, "Original Intensity Preserved Inhomogeneity Correction and Segmentation for Liver Magnetic Resonance Imaging", *Biomedical Signal Processing and Control*, Vol. 47, No. 1, pp. 231-239, 2019.

[9] Xiao-Feng Wang, De-Shuang Huang and HuanXu, "An Efficient Local Chan-Vese Model for Image Segmentation", *Pattern Recognition*, Vol. 43, No. 3, pp. 603-618, 2010.

[10] A. Soudi, G. Neumann and A. Van den Bosch, "Arabic Computational Morphology: Knowledge-Based and Empirical Methods", *Proceedings of International Arab Conference on Arabic Computational Morphology*, pp. 3-14, 2007.

[11] J.H. Yousif, "Natural Language Processing based Soft Computing Techniques", *International Journal of Computer Applications*, Vol. 77, No. 8, pp. 1-7, 2013.