International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Deep learning in the field of disease diagnosis

K.S. Harish Kumar [1] *, Dijo Micheal Jerald [1], A. Emmanuel [1]

[1] Department of Electronics and Communication Engineering, Loyola ICAM College of Engineering and Technology, Nungambakkam, Chennai-600034, India.
** Corresponding author E-mail: harishkumar.19ec@licet.ac.in*

Abstract

A good treatment is dependent on the accuracy of the diagnosis. The cure for the disease starts with the process of diagnosis. All these years, the grade and standard of the medical field has been increasing exponentially, yet there has been no significant downfall in the rate of unintentional medical errors. These errors can be avoided using Deep learning algorithm to predict the disease. The Deep Learning algorithm scans analyses and compares the patient's report with its dataset and predicts the nature and severity of the disease. The test results from the patient's report are extracted by using PDF processing. More the medical reports analyzed, more will be the intelligence gained by the algorithm. This will be of great assistance to the doctors as they can interpret their diagnosis with the results predicted by the algorithm.

Keywords: *Deep Learning; PDF Processing; Prediction.*

1. Introduction

Deep Learning in the field of Diagnosis is the need of the hour, as the unintentional errors are the major threat. Combining Deep Learning and the PDF processing techniques will be of assistance to both Doctors and patients as analyzing medical re-ports using supervised learning is a seamless process.

2. Supervised learning

Supervised learning is a Machine learning type for making the algorithm intelligent in its specified domain. We declare a few domains for the algorithm to learn and improve. The algo-rithm's knowledge is limited only to those specified domains and not beyond that. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. In our work, we use this tech-nique to improve the accuracy of the diseases' severity. This is also required to rectify the errors and repeated errors are elimi-nated. Convolutional Neural Networks (CNN) can also be implemented for improved accuracy, making the algorithm more complex.

3. Uniqueness

3.1. Our approach

Most of the programming for Machine Learning is done in Python, since it provides a wide range of libraries to choose from. Usually, a List (a python data type) is used to develop, train and test a model, which, is not a scalable approach. Also, a model should be able to work for both linear and nonlinear

So, we have changed the methods of giving inputs to train and test the model. Instead of lists, two CSV files are given as train-ing datasets. The feature and label matching are split manually, and the features are put in a separate CSV file and the Labels are but in a separate CSV file

These two CSV files are then parsed into an array of list separately, and then this huge array of lists is used to train the model. With this approach, the classification can be done using any method but here we use the Decision Tree Classifier. The testing data is given as input in the form of a list and using the training data the model predicts and prints the output for the given input.

This approach enables us to use this model for any type of dataset and use the required classification type and get the required output for most of the cases. Since the features and label are split manually before input, it reduces the time required to complete execution. The time of execution may vary according to the size of the dataset.

3.2. Dependencies used

1) CSV: This is used to get input the datasets in CSV format
2) Num Py: This is used to parse the data in the CSV file into corresponding arrays of list.
3) Sklearn: Scikit learn is one of the most used ML li-braries in python. It is used to predict the output by calling the classifier (here Decision Tree)

3.3. Dataset used for predicting heart disease

Heart Disease Dataset from UCI Machine Learning Repository
Features:

1) Age: age in years.
2) Sex: sex (one = male; zero = female).
3) CP: chest pain type.
1) Typical angina
2) Atypical angina
3) Non-anginal pain
4) Asymptomatic
4) Trestbps: resting blood pressure (in mm Hg on admission to the hospital)

5) CHOL: serum cholesterol in mg/dl.
6) FBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7) Restecg: resting electrocardiographic results
Value 0: normal
1) having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
2) showing probable or definite left ventricular hypertrophy by Estes' criteria
8) Thalach: maximum heart rate achieved
9) Exang: exercise induced angina (1 =yes; 0 = no)
10) Old peak = ST depression induced by Exer-cise relative to rest.
11) Slope: the slope of the peak exercise ST segment

Value 1: up sloping
Value 2: flat
Value 3: down sloping

12) Can number of major vessels (0-3) colored by Fluroscopy.
13) Thal: 3 = normal; 6 = fixed defect; 7 = re-versible Defect.

Label:

1) Num: diagnosis of heart disease (angiographic disease status)

Value 0: < 50percentage diameter narrowing
Value 1: > 50percentage diameter narrowing

Usage: The features are the inputs and the label are the predicted output. Refer the link of the dataset for more details.

4. PDF processing

In Python, two packages can be used to manipulate the PDF data.

4.1. PDF miner

This package is used to extract images and object coordinates. If the medical report consists of graphs such as ECG (Electro-Cardiograph), EEG (Electroencephalograph), EMG (Electromyograph), PDFMiner can be used to extract the same.

4.2. Py PDF2

PyPDF2 package is used to split, merge or crop the data in the PDF file. This library is helpful to obtain the text data such as Age, Fasting Blood Sugar, etc.

PDF file is analyzed using both PDFMiner a n d PyPDF2. The output from the PDFMiner consists of images of graphical reports of the patients. These graphs are then image processed and the necessary data is obtained. Similarly, output from the PyPDF2 contains the required text data for the algorithm to predict. The data mined is then inserted into an Excel sheet which acts as a CSV (Comma Separated Values). In this study, the CSV containing the extracted data is named as features. This CSV is passed as one of the two inputs to the Predictive Model. The other input CSV will be the previously predicted value, named label.

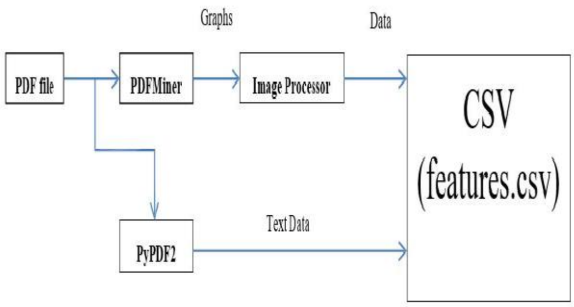

Fig. 1: Flowchart of PDF Processor.

5. Flowchart explanation

The Text Processing block processes the PDF of the patients' report and extracts 13 features which are required by the algo-rithm for prediction. These features are stored as a CSV (comma separated values). As shown in Figure 1, the label contains the predetermined severity ratings. Both the CSV files are fed as inputs to the Predictive Model. NumPy is a dataset to carry out the numerical operations in Python. It produces array of lists for both features and label. Scikit learn (sklearn) is one of the most used ML libraries in python. It is used to predict the output by calling the classifier (here Decision Tree). Decision Tree is a popular classifier which is simple and easy to implement.

The performance of decision trees can be enhanced with suitable attribute selection. The predicted output is from the classification block which gets input from the Decision Tree Classifier after comparing with the testing data.

The classification is done by the process of Clustering. Clustering is a basic machine learning method. The process in which the similar attributes form clusters for the decision to be precise.

Clustering also reduces the computation time.

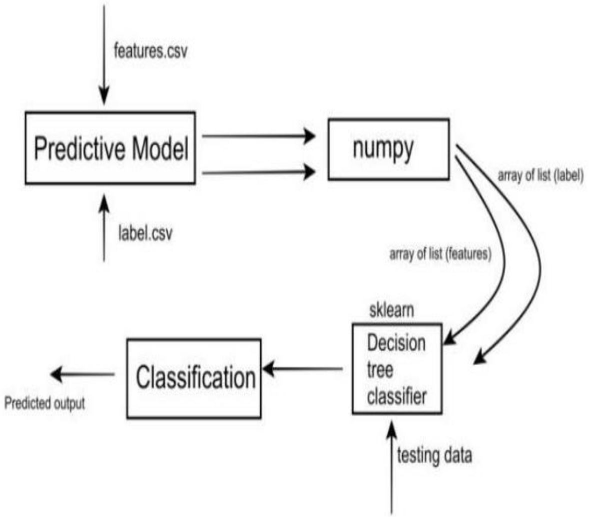

Fig. 2: Flowchart for Prediction Algorithm Using Two CSV'S.

6. learning and error control

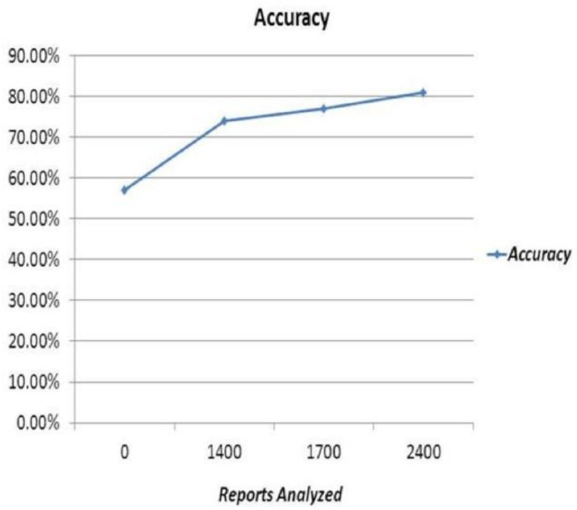

Fig. 3: Graph for Accuracy vs. Number of Reports.

Not all patients show the same symptoms for a disease. Due to this, the diagnosis becomes a complex process. Therefore, there is a considerable probability of error occurrence in the prediction. If the algorithm predicts a value that does not coincide with the exact value, then it is said to be an error.Analyzed
The error can be rectified by changing the hidden weights. For this purpose, there are certain formulae.

1) Derivative of Sigmoid function with respect to weight term:

A = B*(1-B) *x

Where
B is the function of x
A is the derivative of B
X is the weight to be adjusted

2) Derivative of Sigmoid function with respect to bias term

A = B*(1-B)

Where
B is the function of x
A is the derivative of B
X is the weight to be adjusted
From Figure 3, it is observed that the accuracy is increased, for more the number of reports analyzed.

7. Advantages

1) The label and the features CSVs are given separately. Due to this, the time taken to process the report is considerably redeuced.
2) The accuracy of diagnosis increases with respect to the number of reports analyzed.
3) The text processor fills the required fields for the algorithm to diagnose, results in eliminating the need for the physician to manually enter the details for prediction.
4) Assists the doctors in acquiring expertise in diagnosis.

8. Future enhancements

We have used only a single dataset related to heart ailments to prove that two CSVs can also be used for prediction. This method is not only limited to a single dataset and can also be used with multiple datasets. While using multiple datasets, the algorithm can be used to diagnose a wide range of disease oc-curring in various systems in the human body. Further medi-cines can also be prescribed based on the disease diagnosed. This will be of huge assistance to Doctors who lack experience in their service. Also, by enhancing the accuracy and database, with the present reports the patients can be warned whether they are susceptible to a disease and necessary precautions can be taken.

9. Conclusion

The objective of our work is to precisely diagnose the ailments from the medical reports of the patients. By changing the da-taset, we can diagnose diseases occurring in almost every sys-tem in our body. This algorithm is different from the other algo-rithms by using different CSVs for inputs and the predicted output. The PDF reader will get the fields from the PDF file of patients' report and fills the required fields all by itself. We have used De-cision Tree Classifier and Clustering for the prediction. The errors in the predicted output can be corrected by changing the weights for the symptoms manually so that the same error doesn’t occur the next time.

Acknowledgement

Our sincere thanks to Mr. A. Emmanuel, Assistant Professor, De-Apartment of Electronics and Communication Engineering, Loyola-ICAM College of Engineering and Technology for being Our motivation and a mentor.

References

[1] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay; 12(Oct):2825-2830, 2011.I. S. Jacobs and C. P. Bean, ―Fine particles, thin films and exchange anisotropy, ‖ in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[2] Yang, C.L., Chen, X., Nof, S.Y.: Design of a production conflict and error detection model with active protocols and agents. In: Proceedings of the 18th International Conference on Production Research, Italy, July 2005R. Nicole, ―Title of paper with only first word capitalized, ‖ J. Name Stand. Ab-brev., in press.

[3] Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, M. ANBARASI, E. ANUPRIYA, and N.CH.S.N. IYENGAR, School of Computing Science and Engineering, VIT University, Vellore – 632 014, India.

[4] link:http://archive.ics.uci.edu/ml/datasets/heart+Disease. Dataset for heart related ailments.