International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN: 2456-172X | Vol. 1, No. 4, January, 2017 | Pages 89-94

DISCOVERING THE TYPE2 DIABETES USING ELECTRONIC HEALTH RECORD IN MACHINE LEARNING

[1] **Jerina Begum S** [2] **Mageshwari M**
[1] Assistant Professor, [2] UG-Scholar
Department of Information Technology, Agni College Of Technology, Chennai, India

Abstract- Diabetic retinopathy is the most frequent cause of new cases of blindness among adults aged 20–74 years. During the first two decades of disease, nearly all patients with type 2 diabetes have retinopathy. In the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), 4.6% of older-onset patients (type 2 diabetes) were legally blind. In the younger-onset group, 86% of blindness was attributable to diabetic retinopathy. In the older-onset group, in which other eye diseases were common, one-third of the cases of legal blindness were due to diabetic retinopathy. With traditional machine learning methods and leave-10-patients-out cross- validation, our method outperformed a deep learning based MA detection method, with AUC performance improved from 0.962 to 0.985 and F-measure improved from 0.913 to 0.926, using the same DIARETDB1 database. Furthermore, we validated our method on a different dataset – Retinopathy Online Challenge (ROC) dataset. The performance of three classifiers and the pattern with different percentage of principal components are consistent on the two datasets. Especially, we trained the random forest on DIARETDB1 and applied it to ROC; the performance is very similar to that of the random forest trained and tested using cross validation on ROC dataset. This result indicates that our method has the potential to generalize to different datasets.

Keywords: Diabetic retinopathy, Retinopathy Online Challenge, type 2 diabetes

.

1.0 Introduction

Machine Learning is the most popular technique of predicting the future or classifying information to help people in making necessary decisions. Machine Learning algorithms are trained over instances or examples through which they learn from past experiences and also analyze the historical data. Therefore, as it trains over the examples, again and again, it is able to identify patterns in order to make predictions about the future.

Data is the core backbone of machine learning algorithms. With the help of the historical data, we are able to create more data by training these machine learning algorithms. For example, Generative Adversarial Networks are an advanced concept of Machine Learning that learns from the historical images through which they are capable of generating more images. This is also applied towards speech and text synthesis. Therefore, Machine Learning has opened up a vast potential for data science applications.

Machine Learning combines computer science, mathematics, and statistics. Statistics is essential for drawing inferences from the data. Mathematics is useful for developing machine learning models and finally, computer science is used for implementing algorithms. However, simply building models is not enough. We must also optimize and tune the model appropriately so that it provides you with accurate results. Optimization techniques involve tuning the hyper parameters to reach an optimum result. The

world today is evolving and so are the needs and requirements of people. Furthermore, we are witnessing a fourth industrial revolution of data. In order to derive meaningful insights from this data and learn from the way in which people and the system interface with the data, we need computational algorithms that can churn the data and provide us with results that would benefit us in various ways. Machine Learning has revolutionized industries like medicine, healthcare, manufacturing, banking, and several other industries. Therefore, Machine Learning has become an essential part of modern industry.

Data is expanding exponentially and in order to harness the power of this data, added by the massive increase in computation power, Machine Learning has added another dimension to the way we perceive information. Machine Learning is being utilized everywhere. The electronic devices you use, the applications that are part of your everyday life are powered by powerful machine learning algorithms. With an exponential increase in data, there is a need for having a system that can handle this massive load of data. Machine Learning models like Deep Learning allow the vast majority of data to be handled with an accurate generation of predictions.

Machine Learning has revolutionized the way we perceive information and the various insights we can gain out of it. These machine learning algorithms use the patterns contained in the training data to perform classification and future predictions. Whenever any new input is introduced to the ML model, it applies its learned patterns over the new data to make future predictions. Based on the final accuracy, one can optimize their models using various standardized approaches. In this way, Machine Learning model learns to adapt to new examples and produce better results.

2.0 PROPOSED SYSTEM

We analyzed MA detect ability using small 25 by 25 pixel patches extracted from fundus images in the Diabetic Retinopathy Database - Calibration Level 1 (DIARETDB1). Raw pixel intensities of extracted patches served directly as inputs into the following classifiers: a random forest (RF), a neural network (NN), and a support vector machine (SVM). We also explored the use of two techniques (principal component analysis and random forest feature importance) for reducing input dimensionality.

With traditional machine learning methods and leave-10- patients-out cross-validation, our method outperformed a deep learning based MA detection method, with AUC performance improved from 0.962 to 0.985 and F- measure improved from 0.913 to 0.926, using the same DIARETDB1 database. Furthermore, we validated our method on a different dataset – Retinopathy Online Challenge (ROC) dataset. The performance of three classifiers and the pattern with different percentage of principal components are consistent on the two datasets. Especially, we trained the random forest on DIARETDB1 and applied it to ROC; the performance is very similar to that of the random forest trained and tested using cross validation on ROC dataset. This result indicates that our method has the potential to generalize to different datasets.

3.0 ARCHITECTURE

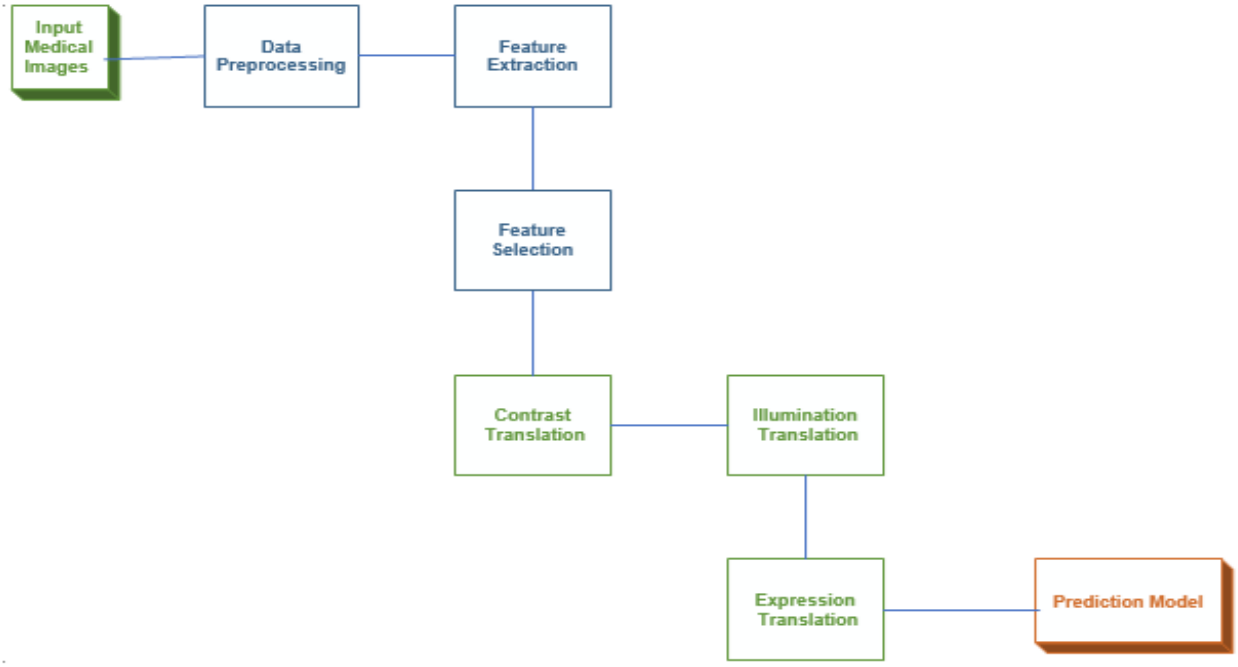

Figure:4.0 :System Architecture

4.0 MODULE DESCRIPTION

4.1. Module 1: Data Import and Preprocessing

Pre-processing is a common name for operations with images at the lowest level of abstraction both input and output are intensity images. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Convert color images to grayscale to reduce computation complexity: in certain problems we will find it useful to lose unnecessary information from your images to reduce space or computational complexity. For example, converting your colored images to grayscale images. This is because in many objects, color isn't necessary to recognize and interpret an image. Grayscale can be good enough for recognizing certain objects. Because color images contain more information than black and white images, they can add unnecessary complexity and take up more space in memory (Remember how color images are represented in three channels, which means that converting it to grayscale reduces the number of pixels that need to be processed).

One important constraint that exists in some machine learning algorithms, such as CNN, is the need to resize the images in your dataset to a unified dimension. This implies that our images must be preprocessed and scaled to have identical widths and heights before fed to the learning algorithm.

4.2. Module 2 : Data Augmentation

We use effective methods that you can use to build a powerful image classifier, using only very few training examples --just a few hundred or thousand pictures from each class you want to be able to recognize. In order to make the most of our few training. Examples, we will "augment" them via a

number of random transformations, so that our model would never see twice the exact same picture. This helps prevent overfitting and helps the model generalize better. The right tool for an image classification job is a covnet, so let's try to train one on our data, as an initial baseline. Since we only have few examples, our number one concern should be overfitting. Overfitting happens when a model exposed to too few examples learns patterns that do not generalize to new data, i.e. when the model starts using irrelevant features for making predictions. For instance, if you, as a human, only see three images of people who are lumberjacks, and three, images of people who are sailors, and among them only one lumberjack wears a cap, you might start thinking that wearing a cap is a sign of being a lumberjack as opposed to a sailor. We would then make a pretty lousy lumberjack/sailor classifier.

Data augmentation is one way to fight over-fitting, but it isn't enough since our augmented samples are still highly correlated. Your main focus for fighting over- fitting should be the entropic capacity of your model --how much information your model is allowed to store. A model that can store a lot of information has the potential to be more accurate by leveraging more features, but it is also more at risk to start storing irrelevant features. Meanwhile, a model that can only store a few features will have to focus on the most significant features found in the data, and these are more likely to be truly relevant and to generalize better. There are different ways to modulate entropic capacity. The main one is the choice of the number of parameters in your model, i.e. the number of layers and the size of each layer. Another way is the use of weight regularization, such as L1 or L2 regularization, which consists in forcing model weights to taker smaller values.

4.3. Module 3 : Model Building

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition”. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous model submitted to ILSVRC-2014. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel- sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU’s.

The input to cov1 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations, it also utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2.

Model Checkpoint helps us to save the model by monitoring a specific parameter of the model. In this case I am monitoring validation accuracy by passing val_acc to Model Checkpoint. The model will only be saved to disk if the validation accuracy of the model in current epoch is greater than what it was in the last epoch.

4.4. Module 4 : Model Performance

As we train your classification predictive model, we want to assess how good it is. Interestingly, there are many different ways of evaluating the performance. Most data scientists that use Python for predictive modeling use the Python package called scikit-learn. Scikit-learn contains many built-in functions for analyzing the performance of models.

4.5. Confusion matrix

Given an actual label and a predicted label, the first thing we can do is divide our samples in 4 buckets:
True positive — actual = 1, predicted = 1
False positive — actual = 0, predicted = 1
False negative — actual = 1, predicted = 0
True negative — actual = 0, predicted = 0

5.0 CONCLUSION

Diabetes is the most common dangerous disease that can lead to additional problems such as heart attack, stroke, blindness, nerve damage, kidney failure, disease of the blood vessels. The predictive analytics in health care is primarily used to determine patients having initial stages of diabetes, asthma, heart disease and other critical lifetime disease. The proposed method using Neural network so it provides better accuracy for predicting type 2 diabetes.

References:

a. F.S. GHAREHCHOPOGH,"Approach and Review of User Oriented Interactive Data Mining", 4th International Conference on Application of Information and Communication Technologies (AICT2010), Digital Object Identifier: 10.1109/ICAICT.2010.5611792, IEEE, Tashkent, Uzbekistan, pp.1-4, 12-14 October 2010.
b. C. Olaru, L. Wehenkel, "Data Mining" , in IEEE Computer Applications in Power, Vol. 12, no. 3, pp. 19-25, July 1999.
c. M. Chen, J. Han, P. S. Yu, "Data mining: an overview from a database perspective", IEEE Transactions on Knowledge and data Engineering, pp. 866-883, 1996.
d. Q. Luo, "Advancing knowledge discovery and data mining", 1st international Workshop on Knowledge discovery and data mining (WKDD'08), Adelaide, South Australia, pp. 3-5, 2008.
e. Diagnosis and Classification of Diabetes Mellitus - NCBI – NIH, by American Diabetes Association - 2010, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797383/, DIABETES CARE, VOLUME 33, SUPPLEMENT 1, JANUARY 2010.
f. Seema Abhijeet Kaveeshwar, Jon Cornwall , The current state of diabetes mellitus in India, Australas Med J. 2014; 7(1): 45–48, doi: 10.4066/AMJ.2013.1979.
g. Prevalence of diabetes and prediabetes in 15 states of India: http://www.thelancet.com/journals/landia/article/PIIS2213-8587(17)30174-2/fulltext?elsca1=tlprDiabetes Prevalence in India: https://www.indiaspend.com/
h. Sadri Sa'di, AmanjMaleki, RaminHashemi, Zahra Panbechi, Kamal Chalabi, " COMPARISON OF DATA MINING ALGORITHMS IN THE DIAGNOSIS OF TYPE II DIABETES", International Journal on Computational Science & Applications (IJCSA) Vol.5,No.5,October 2015
i. DaniahAlmadn, AbdolrezaAbhari, "Comparative analysis of classification models in diagnosis of type 2 diabetes", 2016 Society for Modeling & Simulation International (SCS) SpringSim-MSM, 2016 April 3-6,
j. Pasadena, CA, USA Mahmoud Heydari, Mehdi Teimouri, ZainabolhodaHeshmati, Seyed Mohammad Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran", International Journal of Diabetes in Developing Countries, June 2016, Volume 36, Issue 2, pp 167-173

k. Angus G. Jones, Timothy J. McDonald, Beverley M. Shields, Anita V. Hill, Christopher J. Hyde, Bridget A. Knight, and Andrew T. Hattersley, "Markers of b-Cell Failure Predict Poor Glycemic Response to GLP-1 Receptor Agonist Therapy in Type 2 Diabetes" Diabetes Care Publish Ahead of Print, published online August 4, 2015, DOI: 10.2337/dc15-0258
l. UCI Machine Learning Repository: Data Sets, http://mlr.cs.umass.edu/ml/datasets/