Learning And Predicting Diabetes Data Sets Using Semi-Supervised Learning

Radhika Tayal
Research Scholar
radhikatayal@gmail.com
Department Of Computer Science
Noida International University Uttar Pradesh

Achyut Shankar
Assistant Professor
achyutshankar@gmail.com
Department of computer Science and Engineering
ASET, Amity University Uttar Pradesh

***Abstract*—Now these days, many tools have been developed by the researchers to analyze the impact of diabetes disease on common people within a definite period. However, all these tools have predicted the results based on the labeled dataset or smaller dataset. But in a recent environment, we have collected a large amount of data using both online and offline media. Consequently, data are generated from heterogeneous sources, are in unstructured form and voluminous, etc. As a result, it is not possible to use huge data by using traditional prediction algorithms because they work only on the structured dataset. In this paper, we have used the semi-supervised learning approach that works on a partially labeled dataset for predicting diabetes disease. The partial dataset is the combination of a labeled and unlabelled dataset. For prediction, we have considered 80% unlabelled datasets and 20% labeled datasets. We developed a user based interface for the user to build their prediction model using labeled and unlabeled datasets and analyze the data according to their requirements and interest. Our main objective is to develop a diabetes prediction system that can be used by the researcher and the common people using with minimal labelled datasets.**

I. Introduction

In fast-growing life of information technology, data has become an indispensable part and covers every aspect of our lives. Big Data and machine learning techniques are now becoming the emerging topics for private as well as a government organization. As a result, both private and public company have invested an enormous amount of money in data mining field. Especially, health-care sector because it acts as a data goldmine and machine learning algorithm act like a blessing to give a very precise result. According to McKinsey [1] estimates that big data and machine learning techniques in the field of pharmacy and medicine could generate a value of up to 100 billion annually, for analyzing the problem very deeply so that common people, researchers, doctor, and pharmaceutical used it properly and improve the lifestyle. In this field, many researchers, pharmacy company, physicians, and clinics have focused at the intersection of machine learning and healthcare. Consequently, this has led to the successful deployment of a large number of applications. Healthcare sector is very wide so, here are focusing on the diabetes disease dataset for both modeling and prediction purpose.

In modeling phase, we have used 80% labeled and 20% unlabeled dataset to prepare the model. Our model was designed based on the extended self-training technique and the classification algorithm. In prediction phase, we predict the new unlabeled data (new patient record) using predicted model and shows the result in human understandable format. Here, we predict the health status of diabetic patients. Apart from prediction, we also evaluate the probability of having the diabetes disease in future. Our system will also capable of answering millions of queries. We interpreted the result in a human understandable form with the help of graph and decision boundary. We developed UI based disease prediction tool that will use by many user and physicians to analyze the health status of an individual.

The paper is organized as follows: Section 2 presents the motivation of the developing user interface tool using partially labeled datasets. All the related works relation to prediction and semi-supervised learning is introduced in section 3. The dataset, methodology and system architecture are discussed in section 4 5 and 6. Section 7 introduces our findings. Finally, section 8 is devoted to the conclusions of the proposed model.

II. MOTIVATION

The main motivation behind this learning and prediction the system is to create health awareness among the middle age people and prevent them from suffering diseases like diabetes. By research, it is found that billion of a population is suffering from diabetes irrespective of all age groups [2], [3]. Here, we developed a UI tool that will automatically detect people suffering from diabetes and also can predict the probability of having it in the future. Undoubtedly people are aware of these factors, but they are unable to follow this, due to their busy life and poor management. By keeping in mind, we develop a tool in which a user can enter few specific health-related parameters and analyze of having diabetes in the future with the help of confidence level. In our tool, it is not necessary to enter recent data to predict the results. Our main objective is to use recent data (a few weeks or months old) to predict and provide precise results. By keeping these requirements in mind, we developed a tool that is used for modeling the partially labeled dataset and predict the data based on the newly developed model.

III. RELATED WORK

As part of the literary work, we observed that number of research has conducted into diabetes disease diagnosis. Among all of them, 84% of researchers have used the supervised learning technique, 15% used the unsupervised learning technique and a very less percentage of approx 1% has used a semi-supervised learning approach in the field of medical diagnosis

978-1-7281-2791-0/20/$31.00 ©2020 IEEE

[4]. For implementing all these learning algorithm support vector machines and association rule mining is the most widely used classification algorithm. Based on several studies, we found that most of the researchers have used the Pima Indians Diabetes Dataset and UCI dataset for predicting the medical diagnosis[5]. In the field of semi-supervised learning, most of the work was done by listed researchers. Liang Y[6] used a semi-supervised learning method and developed a tool for survival analysis in clinical cancer research. Patil [7] used a weka toolkit to design a hybrid model using a k-mean and C4.5 algorithm and obtained an accuracy of 92.38%. Marcano-Cedeño [8] also used the Pima Indians Diabetes Dataset and Waikato Environment for Knowledge Analysis (WEKA) toolkit for developing a prediction model for diabetes by using multilayer perceptron model and obtained an accuracy of 89.3%. Han et al. [9] also used Pima Indians Diabetes Dataset and Rapid Miner tool to analyze and understand the links between plasma glucose and class attributes. Kavitha and Sarojamma [10] used the CART tool to analyze the hidden pattern from a complex dataset. S. Kumari and A. Singh [11] used Matlab software to predict diabetic patients at an early stage using a neural network algorithm.

However, no one developed a prediction tool based on the combination of labeled and unlabeled data. Almost all the tools are built based on the labeled dataset using different classification algorithms. Hence, there is a requirement of providing reliable, faster and cost-effective methods to provide information about the probability of a patient to have diabetes. present work, analyze the diabetes parameters and establish a combination of the multiple classifiers and improved the prediction system using labeled and unlabeled datasets.

IV. MATERIAL AND METHOD

A. Dataset

For our experiment, we have extracted diabetes dataset from UCI repository for evaluating our algorithm [17]. Here are details about these datasets.

1) Description of Diabetes Dataset: Pima India Dataset: Pima Indians Diabetes dataset is excerpted from the UCI Machine Learning Repository[5]. This dataset contains 8 categories and 768 instances of female patients at least 21 years old. This dataset is originally developed by the National Institute of Diabetes and Digestive and Kidney Diseases and Vincent Sigillito of the Applied Physics Laboratory of the Johns Hopkins University was the original donor of the dataset [5]. This Pima dataset has been used by the researcher to indicate the presence of diabetes in patients. PIMA dataset features are as follow:

1) Pregnancy: Number of times of pregnant.
2) Plasma-Glucose: Plasma glucose concentration measured using a two-hour oral glucose tolerance test of Blood sugar level.
3) Diastolic BP: Diastolic blood pressure (mmHg).
4) Triceps SFT: Triceps skinfold thickness (mm).
5) Serum-Insulin: 2-hour serum insulin (mu U/mt).
6) BMI: Body mass index (w in kg/h in m)
7) DPF: Diabetes pedigree function.
8) Age: Age of the patient (years).

The class variable consists of binary values 0 and 1. 0 means a patient has no diabetes indicating a healthy person and 1 means a patient has diabetes indicating the diabetic person. All patients in this dataset are Pima Indian women at least 21 years old and living near Phoenix, Arizona, USA. As shown in Table I there are 500 cases in class 0 and 268 cases in class1 as Outcome.

TABLE I. OUTCOME MATRIX

	Outcome
No. of Diabetic Patient (1)	268
No. of Non-Diabetic Patient (0)	500

V. METHODOLOGY

We used the self-training algorithm [12] with the SVM classifier to design our model. This algorithm uses single view dataset, single learning approach and single classifiers as SVM to build the model based on both labeled and unlabeled data. SVM is the classifier that used to classify the labeled and unlabeled data. This algorithm initially builds the model based on the available small set of labeled data (training data) using the SVM classifier. We have built a new classifier to classify the unlabeled data using the same classifier. Next, a set of unlabeled data points, of which class labels can be predicted with high confidence, is added to the training set. Then a model is re-trained based on a new pool of labeled dataset and again predicts the label of the unlabeled dataset. We have repeated this procedure eight times to convert all unlabelled datasets into a labeled dataset. When all the unlabeled samples are labeled then, make the final classifier using the SVM model.

VI. SYSTEM ARCHITECTURE

We designed a tool that is divided into phase modeling and prediction. In the modeling phase, we have designed the model based on a small set of labeled data and a large set of an unlabeled dataset. For designing and predicting the outcome, we used a self-training algorithm using the SVM classifier. The tool would take the partial dataset of a patient and multiscale simulation model for the diabetes patient, leading to more predict the future probability of having it in the future. The model could also serve as a screening tool for doctors and drug companies who are trying to develop new therapies. The detailed description of each phase is described below.

A. UI based tool: Diabetes Detection

We have designed a free UI based tool for surveillance of disease namely diabetes to detect spatial, temporal or spacetime disease boundaries and to examine whether these results are statistically significant or not missing. It is used to perform discrete statistics analysis. For discrete statistics analysis the number of diabetes patient cases and also observed the result of a dataset is not random, i.e. BMI, glucose level has a very high impact on diabetes disease. The main intuition behind this diabetes detection is to create awareness among the people and decrease the rate of having it among the people. It also provides a common interface to a user to build its model based on our newly designed algorithm. Our prediction tool used is designed in python using sckit-learn [15] and tkinter [16] library for developed GUI. For preparing the model, we

used was PIMA Indian diabetes dataset. Our tool contains the following functions.

1) Modeling: In the model phase, we are creating the model of the dataset as per the data are given by the user.
2) Anonymization: In the anonymization phase, the user converts the original dataset into an anonymized dataset, if it is required.
3) Visualization: In the visualization phase, the user analyzes the distribution of the model based on the decision boundary plot.

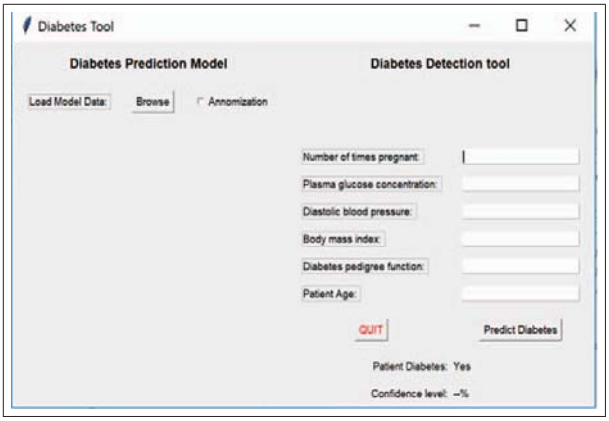

Fig. 1. UI Tool layout

1) Data Requirements: The input data needs to be stored in CSV files. For designing the model input files that are not mandatory, a user can use our default diabetes dataset. The input file should be in comma-delimited and the features are specifying in an array format. So for modeling the dataset in our tool. The user has to click only on the load data tab and browse the dataset from the system. If the data are not properly entered by the user, then it gives the error message,“ Data Loading Unsuccessfully” along with the file name. After successfully inserting the input file in the user interface tool. The algorithms are executed and generated a model. On the bottom left side, it shows the result in the human-understandable format along with the graph.

TABLE II. MODEL SAMPLE DATA

Pregnancies,Glucose,Blood Pres-sure,Skin Thickness, Insulin,Outcome
6,148,72,35,0,1
1,85,66,29,0,0
8,183,64,0,0,1
1,89,66,23,94,0
0,137,40,35,168,1

2) Outputs of detection tool: After successfully loading the data file, without any click, we build our model on the partial dataset using our semi-supervised algorithm. We showed the distribution of the data in the form of decision boundary graph. Decision boundary graph clearly gives us the idea of the positive and negative sample. So, that user can get a clear idea about the data and how the model performs. Along with the graph, it can also calculate the accuracy, sensitivity and ROC-AUC percentage of the model. Our tool works for all type of medical datasets. As here, in our UI based tool, we are using only diabetes disease.

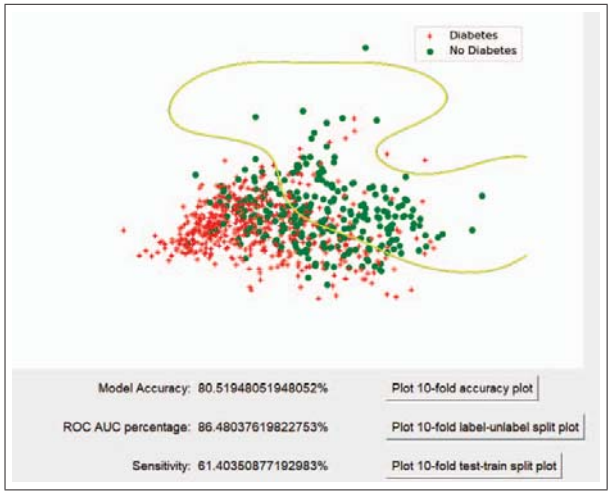

Fig. 2. Model Accuracy Using 10-Fold

B. UI based tool: Diabetes Prediction

A diabetes detection UI based tool is a good way to know the chance of having diabetes in advance. For detecting the diabetic patients, a user has to give some basic input parameter related to their health. The tool will automatically tell that whether you have a chance of having diabetes or not along with the confidence level.

1) Data Requirements:

- Plasma glucose concentration A blood sample or plasma glucose test can be taken not at any time. It has a very high impact on time when it has taken before a meal or after the meal. In our tool, we requested to use to enter the value after the meal. Our tool only required value. It has very less impact if the user entered the value months year old.
- Diastolic blood pressure: Diabetes and blood pressure have a direct relationship between them. About 25% of people with Type 1 diabetes and 80% of people with Type 2 diabetes have high blood pressure. Having diabetes raises your risk of heart disease, stroke, kidney disease, and other health problems. The input must be provided in mm Hg unit. So blood pressure is an important feature to evaluate diabetes.
- Body Mass Index: BMI is strongly and independently associated with the risk of being diagnosed with T2D. The incremental association of BMI category on the risk of T2D is stronger for people with a higher BMI relative to people with a lower BMI. The input must be provided in weight in $(kg/(height\_in\_meter)^2$.
- Diabetes pedigree function: The diabetes pedigree function provides a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject. It utilizes information from a person’s family history to predict how diabetes will affect that individual. In our tool, the user provides the value in the form of 0 and 1. The system will automatically calculate the function value.

- patient Age: Age has a high impact on diabetes. According to the Centers for Disease Control and Prevention (CDC) middle-aged and older adults are still at the highest risk for developing type 2 diabetes. Due to work pressure and tension middle-aged people not take care of their health. The input must be provided in the year.
- Number of times pregnant: User has to tell about the number of times pregnancy because Gestational diabetes, generally happens during pregnancy. It does not mean the person is having diabetes.

For predicting the result, we required these basic input parameters is needed. For predicting the result, we required these basic input parameters is needed.

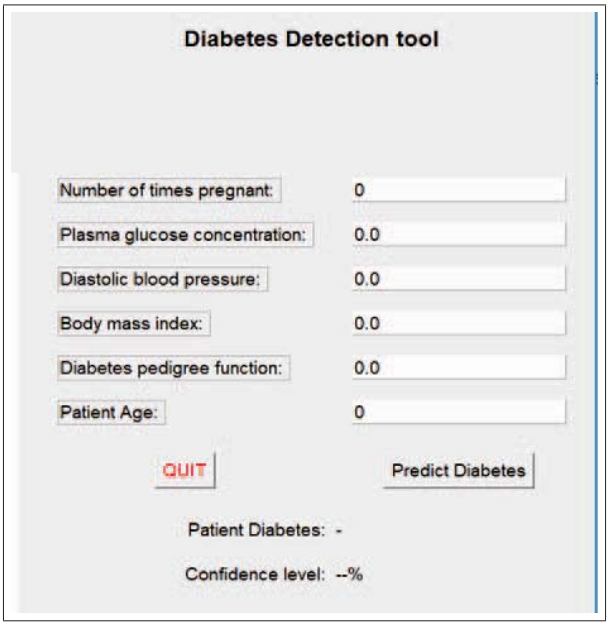

Fig. 3. Input for predicting diabetes

2) Outputs of prediction tools: After providing the input parameter in proper units, our system predicts the result using the pre-built model. After giving the user input to the model. Our model gives the result in binary form ("Yes/No"). Apart from the result, it also predicts the confidence level. The confidence level shows the probability of having diabetes disease in the future. With the help of confidence level, we can understand the impact of blood sugar level in our body.

VII. RESULTS AND ANALYSIS

We evaluated the performance of our model on different datasets with the help of the prediction tool and analyze the performance of our semi-supervised model. We have used 10-fold cross-validation for our accuracy calculation to get robust metrics. Our tool provides the functionality to get the 10- fold cross-validation accuracy plot with different label sample percentages. Fig. 2 shows the screenshot of a tool providing the model accuracy 10-fold plot. As we can see the graph compare our model accuracy with SVM [13] and TSVM [14] and our model shows great improvement in the diabetes dataset. Fig. 2 also shows the tool providing a visual image of decision boundary for a given dataset along with accuracy and other metrics in the left pane of a tool which gives handsome confidence to a user using the tool about the model. The tool is flexible enough to handle different datasets and able to plot decision boundaries and other parameters accordingly. Fig. 3 shows the visualization of the right side of the pane which is focused on prediction system, the tool provides the GUI option to input the different parameter values to calculate prediction output, the tool is simple enough to hide the complexity from the end-user and provide the output in a human-readable format. It not only shows the output but along with confidence which helps end-user in accessing the output and interoperate it for better understanding the result.

VIII. CONCLUSION

As per new emerging technologies, a tremendous amount of data is generated by many organizations. The main intuition here is to generate a model that used both the labeled and unlabeled dataset to make the model and use it for prediction. In this paper, diabetes datasets have been used for modeling and prediction. We designed a system, in such a way that can be useful to the medical student, doctors, and normal end-user to analyses the different dataset in the form of graph and statistical measures. As we have developed diabetes prediction system. Our system is categories into two parts namely learning and prediction. For prediction, we used new patient data as an input parameter for predicting the result. We generalize our leaning system by modifying the dataset. In the future, we can use also generalize the prediction system. So this work can be extended to make this entire process automatic where a user just needs to provide the dataset and its metadata along with few pre-defined tuning values and system will prepare a highly accurate and efficient model by performing all pipeline steps including preprocessing, feature extraction, data split and predicting unknowns. GUI tool can also be improved to include other useful features like providing dataset information and other meta-information required to build an effective model and then populate input form to allow a user to provide those features and able to predict the output. This will become the E2E solution and can be considered as a generic solution for classifications.

REFERENCES

[1] M. Levy, Notes from the AI frontier: Applications and value of deep learning, June 6, 2018. Accessed on: June 23, 2019. [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

[2] A. Fernández, "Non-Adherence to Newly Prescribed Diabetes Medications among Insured Latino and White Patients with Diabete", JAMA internal medicine 177. vol. ED-3, pp. 371–379, June 2017

[3] N. Nirala, R. Periyasamy, B.K. Singh, Awanish Kumar, "Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine", Biocybernetics and Biomedical Engineering, 2018.

[4] I. Kavakiotis, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal- 15, pp. 104–116, June 2017.

[5] "UCI Machine Learning Repository: Data Sets", June 6, 2007. Accessed on: Feb 12 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.php

[6] Y. Liang, H. Chai, XY.Liu, ZB. Xu, H. Zhang, KS. Leung, “Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with L1/2 regularization”, BMC Med Genomics, pp.9-11,Mar 2016.

[7] B.M. Patil, “ Hybrid prediction model for Type-2 diabetic patients”, International Journal, vol. ED-37, pp. 8102-8108, Dec. 2010.

[8] A.M. Cedeno, J. Torres, D. Andina, “A prediction model to diabetes using artificial metaplasticity”, IWINAC , Part II, LNCS, vol. 6687, pp. 418-425, May 2011.

[9] J. Han, J. C. Rodriguez, and M. Beheshti, “Diabetes data analysis and prediction model discovery using rapidminer ”, Future Generation Communication and Networking (2008), pp. 96-99, Dec 2008.

[10] K. Kavitha and R. M. Sarojamma, “Monitoring of diabetes with data mining via cart method”, International Journal of Emerging Technology and Advanced Engineering, vol. 2(11), pp. 157 - 162, Nov 2012.

[11] S. Kumari and A. Singh, “A data mining approach for the diagnosis of diabetes mellitus”, 7th International Conference on Intelligent Systems and Control (ISCO), pp. 373-375, Jan 2013.

[12] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods”, 95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pp. 189-196, June 1995 .

[13] C. Cortes, V. Vapnik , “Support-vector networks”, Machine Learning, vol. 20(3), pp.273–297, Sep 1995.

[14] Y. Shi, K. Yao, H. Chen, Y. C. Pan and M. Y. Hwang, ”Semi-supervised slot tagging in spoken language understanding using recurrent transductive support vector machines”, 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 353-360, Dec 2015.

[15] “scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation”, 2017. Accessed on: June 28, 2019. [Online]. Available: https://scikit-learn.org/stable

[16] “24.1. Tkinter — Python interface to Tcl/Tk — Python 2.7.17rc1 documentation”, 2017. Accessed on: June 18, 2019. [Online]. Available: https://docs.python.org/2/library/tkinter.html

[17] K. Kamer, AU Yildirim Tlay, “Medical diagnosis on Pima Indian diabetes using general regression neural networks”, 2003.