JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Detection of Type2 Diabetes Using FIMMG Dataset based on Machine Learning Algorithms

[1] Birada Taruni Lakshmi Ganasai
[1]M.tech, Department of Computer Science
Andhra University, Visakhapatnam, AP, India

ABSTRACT: The field of biosciences have progressive to a higher extent and have generated large amounts of information from Electronic Health Records. This have given rise to the acute need of knowledge generation from this enormous amount of data. Data mining methods and machine learning play a major role in this aspect of biosciences. An adequate Type 2 Diabetes unified administration system and regular timely checkup has key role in treatment of Type-2 Diabetes at initial stages. In Recent years there is rapid increase of evolution of Machine learning technique and FIMMG Dataset which is category of Electronic Health Record. Over fitting, Model interpretability and computational cost are the challenges while managing these much of information. Based on these challenges, we proposed a Machine Learning technique called Sparse Balanced Support Vector Machine (SBSVM) Based Type 2 Diabetes discovering by using Electronic Health record dataset named FIMMG dataset. We have collected data for Type 2 diagnosis from uniform age group that related to Electronic Health records such as exemptions, examination and drug prescription. Machine Learning and Deep neural networks are mainly used in solving task. Results proved that Sparse Based SVM provide better predictive performance and computation time when compared to techniques that are present in existing system. To increase model interpretability, we introduced induced sparsity which manages data which have high dimension. In proposed system we used Random Forest tree, Linear Regression, Decision Tree, Voting Classifier algorithms. Among all algorithms we found Random Forest tree accuracy, sensitivity is high.

KEYWORDS: FIMMG Dataset, Random Forest tree, Linear Regression, Decision Tree, Voting Classifier algorithms,T2d detection

I. INTRODUCTION

Diabetes is a disorder traditionally subdivided into two types. Type 2 diabetes (T2D) is a chronic condition that affects how our body metabolizes sugar or glucose, inducing either resistance to the effects of insulin, or lack of its production in a way sufficient to maintain normal glucose levels. No cure exists for such disorder affecting populations that include adults as well as children. Control of body weight, diet, and exercise can help T2D management, complementing (or as an alternative to) medications or insulin therapy [1].

T2D is of interest to this work. The classification of diabetes depends primarily on age at onset and the presence or absence of conditions such as obesity, metabolic syndrome, insulin deficiency, and others. Several mechanisms can lead to diabetes, and these can be modified by genetic, lifestyle, and environmental factors. Clearly, all such factors make T2D a very heterogeneous disease, one for which many types of data should be analyzed for achieving superior precision of diagnoses and therapies[2].

Precision Medicine (PM) explores the distinct characteristics in individuals that make their disease signatures or risk profiles possibly unique. This complexity involves the acts of first collecting widespread information specific to the individual and then streamlining data-driven processes subject to treatment by automated rules. A domain of support to PM data analytics is electronic health records (EHR). EHR represent complex heterogeneous information systems that call for algorithmic approaches in order to quantify their saliency and the related uncertainty [3]. Once these two properties are suitably assessed, one of the most valuable roles that EHR may play is to generate new disease phenotypes, leading to novel classifications, and taxonomies (Pendergrass and Crawford, 2019). PM approaches beyond one-size-fits-all models respond to the challenge of deciphering the interactions between EHR components. These have value especially in predictive terms, knowing that prediction represents a crucial outcome of data analytical tools. Additionally, there is translational value potentially achievable when disease patterns are discovered before the

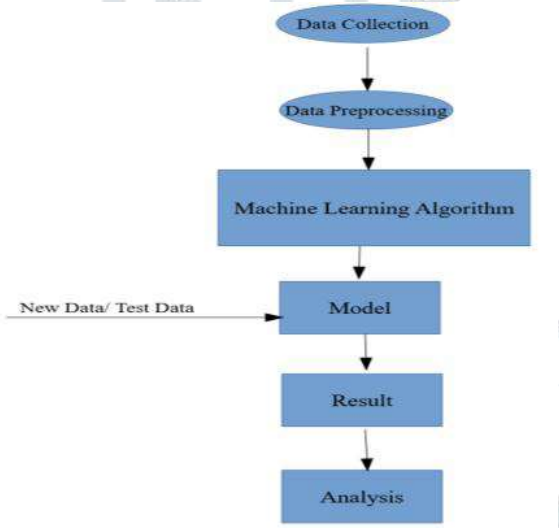
appearance of patient's symptoms, or when risk and/or outcome profiles are evaluated in patient populations for stratification purposes [7].

The high number of patients' information recorded in EHRs results in a large amount of stored data. In this context, one of the aims of biomedical informatics is to convert EHR into knowledge that could be exploited for designing a clinical Decision Support System (DSS). In this scenario, ML models are able to manage this enormous amount of data by predicting clinical outcomes and interpreting particular patterns sometimes unsighted by physicians [5]

The aim of this work is to exploit a ML methodology, named sparse balanced Support Vector Machine (SB-SVM), for discovering T2D using features extracted from a novel HER dataset, namely the FIMMG dataset. The proposed SB-SVM is able to manage high dimensional data by increasing the model interpretability and finding the most relevant features while dealing with the usual unbalanced class distribution. In the data analysis, among all the EHR features related to exemptions, examination and drug prescriptions, we have considered only those collected before T2D diagnosis, while excluding all features that have already revealed a T2D patient's follow- up [6]. Additionally, we considered a subset of subjects enclosed

From 60 −80 year's range, where the chronological age is not statistically relevant in order to discriminate T2D condition. The prime job of kidneys is to filter extra water and wastes from blood. The efficient functioning of this process is important to balance the salts and minerals present in our body. The rich balance of salts are necessary to control blood pressure, activate hormones, build red blood cells, etc. A high concentration of calcium leads to various bone diseases and cystic ovaries in women. Type-ii diabetes also may lead to sudden illness or allergy to certain medicines. Type-ii diabetes in many cases leads to permanent dialysis or kidney transplants.

The detailed review on literature shows the application of various machine learning algorithms to predict type-ii diabetes. This paper tries to predict type-ii diabetes using the classifiers like decision tree, random forest and support vector machine and also suggested best prediction model [8].



2. RELATED WORK

K G Alberti 1 , P Z Zimmet proposed he classification of diabetes mellitus and the tests used for its diagnosis were brought into order by the National Diabetes Data Group of the USA and the second World Health Organization Expert Committee on Diabetes Mellitus in 1979 and 1980. Apart from minor modifications by WHO in 1985, little has been changed since that time. There is however considerable new knowledge regarding the aetiology of different forms of diabetes as well as more information on the predictive value of different blood glucose values for the complications of diabetes. A WHO Consultation has therefore taken place in parallel with a report by an American Diabetes Association Expert Committee to re-examine diagnostic criteria and classification. The present document includes the conclusions of the former and is intended for wide distribution and discussion before final proposals are submitted to WHO for approval.

B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle proposed Experts consider health information technology key to improving efficiency and quality of health care. To systematically review evidence on the effect of health information technology on quality, efficiency, and costs of health care. The authors systematically searched the English-language literature indexed in MEDLINE (1995 to January 2004), the Cochrane Central Register of Controlled Trials, the Cochrane Database of Abstracts of Reviews of Effects, and the Periodical Abstracts Database. We also added studies identified by experts up to April 2005. Descriptive and comparative studies and systematic reviews of health information technology. Two reviewers independently extracted information on system capabilities, design, effects on quality, system acquisition, implementation context, and costs. 257 studies met the inclusion criteria. Most studies addressed decision support systems or electronic health records. Approximately 25% of the studies were from 4 academic institutions that implemented internally developed systems; only 9 studies evaluated multifunctional, commercially developed systems. Three major benefits on quality

were demonstrated: increased adherence to guideline-based care, enhanced surveillance and monitoring, and decreased medication errors. The primary domain of improvement was preventive health. The major efficiency benefit shown was decreased utilization of care. Data on another efficiency measure, time utilization, were mixed. Empirical cost data were limited.

R. Kaushal, K. G. Shojania, and D. W. Bates proposed Computerized physician order entry (CPOE) with clinical decision support (CDS) has been promoted as an effective strategy to prevent the development of a drug injury defined as an adverse drug event (ADE). To systematically review studies evaluating the effects of CPOE with CDS on the development of an ADE as an outcome measure. PUBMED versions of MEDLINE (from inception through March 2007) were searched to identify relevant studies. Reference lists of included studies were also searched. We searched for original investigations, randomized and nonrandomized clinical trials, and observational studies that evaluated the effect of CPOE with CDS on the rates of ADEs. The studies identified were assessed to determine the type of computer system used, drug categories being evaluated, types of ADEs measured, and clinical outcomes assessed. Of the 543 citations identified, 10 studies met our inclusion criteria. These studies were grouped into categories based on their setting: hospital or ambulatory; no studies related to the long-term care setting were identified. CPOE with CDS contributed to a statistically significant (P < or = .05) decrease in ADEs in 5 (50.0%) of the 10 studies. Four studies (40.0%) reported a non-statistically significant reduction in ADE rates, and 1 study (10.0%) demonstrated no change in ADE rates. Few studies have measured the effect of CPOE with CDS on the rates of ADEs, and none were randomized controlled trials. Further research is needed to evaluate the efficacy of CPOE with CDS across the various clinical settings.

R. Amarasingham, L. Plantinga, M. Diener-West, D. J. Gaskin, and N. R. Powe proposed despite speculation that clinical information technologies will improve clinical and financial outcomes, few studies have examined this relationship in a large number of hospitals. We conducted a cross-sectional study of urban hospitals in Texas using the Clinical Information Technology Assessment Tool, which measures a hospital's level of automation based on physician interactions with the information system. After adjustment for potential confounders, we examined whether greater automation of hospital information was associated with reduced rates of inpatient mortality, complications, costs, and length of stay for 167 233 patients older than 50 years admitted to responding hospitals between December 1, 2005, and May 30, 2006. We received a sufficient number of responses from 41 of 72 hospitals (58%). For all medical conditions studied, a 10-point increase in the automation of notes and records was associated with a 15% decrease in the adjusted odds of fatal hospitalizations (0.85; 95% confidence interval, 0.74-0.97). Higher scores in order entry were associated with 9% and 55% decreases in the adjusted odds of death for myocardial infarction and coronary artery bypass graft procedures, respectively.

S. T. Parente and J. S. McCullough proposed the potential of health information technology (IT) to transform health care delivery has spurred health IT adoption and will likely contribute to increased investments in coming years. Although an extensive literature shows the value of health IT at leading academic institutions, its broader value remains unknown. We sought to estimate IT's effect on key patient safety measures in a national sample. Using four years of Medicare inpatient data, we found that electronic medical records have a small, positive effect on patient safety. Although these results are encouraging, we suggest that investment in health IT should be accompanied by investment in the evidence base needed to evaluate it.

C. Chen, T. Garrido, D. Chock, G. Okawa, and L. Liang, proposed we examined the impact of implementing a comprehensive electronic health record (EHR) system on ambulatory care use in an integrated health care delivery system with more than 225,000 members. Between 2004 and 2007, the annual age/sex-adjusted total office visit rate decreased 26.2 percent, the adjusted primary care office visit rate decreased 25.3 percent, and the adjusted specialty care office visit rate decreased 21.5 percent. Scheduled telephone visits increased more than eightfold, and secure e-mail messaging, which began in late 2005, increased nearly sixfold by 2007. Introducing an EHR creates operational efficiencies by offering nontraditional, patient-centered ways of providing care..

3. PROPOSED MACHINE LEARNING MODELS

DECISION TREE

Decision Tree one of the algorithm, is used to solve regression and classification problems. The general objective of using Decision Tree is to create a model that predicts classes or values of target variables by generating decision rules derived from training data sets. Decision tree algorithm follows a tree structure with roots, branches and leaves. The attributes of decision making are the internal nodes and class labels are represented as leaf nodes. Decision Tree algorithm is easy to understand compared with other classification algorithms. The Decision trees algorithm consists of two parts: nodes and rules (tests). We construct the tree. In which each node reflect a test on an attribute the basic idea of this algorithm is to draw a flowchart diagram that contains a root node on top. All other (non-leaf) nodes represent a test until you reach a leaf node (final result). Decision tree algorithms have been widely used in data mining applications below are some important reasons that why decision trees are used in the area of data mining and classification:

Decision trees create user-friendly rules. They are considered one of easy to understand algorithms to the end user in Data mining. They show effective association among the dataset attributes and represent in an easy-to-understand form. Decision trees provide a clear indication of important attributes. Decision trees require less computation. They require less computation compared to other classification algorithms. When we implementing decision trees to detect breast cancer then leaf nodes are divided into two categories: Benign or Malignant. Rules will be established among the chosen data set attributes in order to determine if the tumor is benign or malignant.

Random Forest

Random forest algorithm constructs multiple decision trees to act as an ensemble of classification and regression process. A number of decision trees are constructed using a random subsets of the training data sets. A large collection of decision trees provide higher accuracy of results. The runtime of the algorithm is comparatively fast and also accommodates missing data. Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees. Random forest (RF) is a bagging ensemble approach proposed by Breiman that based on a machine learning mechanism called "decision tree". In a random forest, the "weak learners" in ensemble terms are decision trees. Random forest imposes the diversity of each tree separately by selecting a random feature. After generating a large number of trees, they vote for the most common class. The random forest algorithm can deal with unbalanced data, it is robust against overfitting, and its runtimes are quite a bit faster

Support Vector Machine

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and multivariate analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues, SVMs are one among the foremost robust prediction methods, being supported statistical learning frameworks or VC theory proposed by Vapnik and Chervonenkis. Given a group of coaching examples, each marked as belonging to at least one of two categories, an SVM training algorithm builds a model that assigns new examples to at least one category or the opposite , making it a non-probabilistic binary linear classifier (although methods like Platt scaling exist to use SVM during a probabilistic classification setting). An SVM maps training examples to points in space so on maximize the width of the gap between the 2 categories. New examples are then mapped into that very same space and predicted to belong to a category supported which side of the gap they fall. Type-II Diabetes has varying levels of seriousness. It usually gets worse over time though treatment has been shown to slow progression. If left untreated, Type-II Diabetes can progress to Diabetes and early cardiovascular disease..

Voting Classifier

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

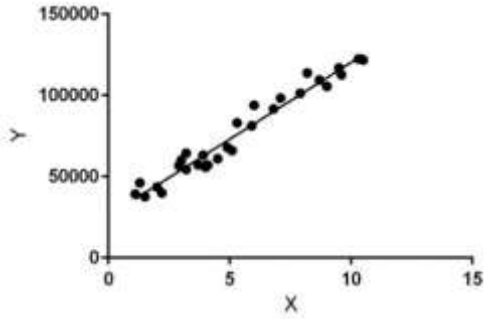
It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class. Voting Classifier supports two types of votings.

1. Hard Voting: In hard voting, the predicted output class is a class with the highest majority of votes i.e the class which had the highest probability of being predicted by each of the classifiers. Suppose three classifiers predicted the output class(A, A, B), so here the majority predicted A as output. Hence A will be the final prediction.
2. Soft Voting: In soft voting, the output class is the prediction based on the average of probability given to that class. Suppose given some input to three models, the prediction probability for class A = (0.30, 0.47, 0.53) and B = (0.20, 0.32, 0.40). So the average for class A is 0.4333 and B is 0.3067, the winner is clearly class A because it had the highest probability averaged by each classifier.

Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and

forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Classification Accuracy

Accuracy of the constructed classifier model can be calculation using the following equation.

TP+TN

Accuracy = (1)

TP+TN+FP+FN

Where,

TP = Observation is positive an predicted is also positive

TN=Observation is negative and predicted is also negative

FP = Observation is negative but predicted is positive

FN = Observation is positive but predicted is negative

Sparse Balanced SVM

The margin of the predicted SB-SVM response was mapped into interval by using a sigmoid function, without changing the SVM error function. The mapping was realized according to adding a post-processing step where the sigmoid parameters were learned with regularized binomial maximum likelihood. The computed probabilistic outputs of SB-SVM reflects the predicted response (yp) based on the threshold th = 0.5:

P(yp = T 2D|f) = 1 1 +ex p(A f +B) >th yp = T 2D else yp = Control

Several solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels. Differently, from the data-level solutions which include many different forms of re-sampling (e.g., random re-sample, directed re-sample, oversample with informed generation), we have decided to work at the algorithm level. In particular, we adjust the decision threshold th in the validation set in order to maximize the macro-recall metric while alleviating the effect of high unbalanced data. The prediction of the SB-SVM produces an uncalibrated value that is not a probability. The post-processing step allows to transform the output of the SB-SVM classifier (i.e., distance from the margin) into posterior probability. Thus, the posterior probability represents a salient information which can be integrated in a clinical DSS for supporting the early-stage diagnosis by revealing the confidence level of the performed prediction. The idea behind the employed methodology lies in the use of a parametric model to fit the posterior directly. Hence, the parameters A and B of the sigmoid function are adapted to give the best probability outputs. Differently from the employed sigmoid model has two parameters trained discriminatively, rather than one parameter

DATA SET AND ALGORITHM

The FIMMG_obs dataset is a subset of the FIMMG dataset, because not all the patients and the EHR fields have been selected for this study. The FIMMG_obs dataset contains a total of 968 patients and 3 main EHR fields.All the type 2 diabetes patients were excluded from the FIMMG_obs dataset.The number of different features for each main field is enclosed in square brackets:

- Demographic (Gender, Age) [2]
- Monitoring (Systolic and diastolic blood pressure, Height, Weight, BMI) [5]
- Clinical (Laboratory exams) [73]

The date of each laboratory exam and blood pressure measurement is reported for each patient. This aspect assumes a relevant significance, because it allows to trace up the patient's longitudinal clinical history from 2010 to 2018, by collecting a total of 2276 observations in accordance with the triglyceride-glucose (TyG) index measurements for each patient.

Dataset description	Count	Mean (Std)
Total patients	968	-
Observation period (years)	9	-
Total observations	2276	-
Fields	**Count**	**Mean (Std)**
Demographic	2	
Gender	-	
Male	473	-
Female	495	-
Age (years)	-	61(±18)
Monitoring	5	
Blood pressure ($mmHg$)	-	
Systolic	-	135(±16)
Diastolic	-	82(±9)
Height (cm)	-	160(±16)
Weight (Kg)	-	83(±17)
Body Mass Index (Kg/m^2)	-	32(±5)
# Laboratory exams	#	Laboratory exams
1 Thyroglobulin antibodies (TgAb)	38	Gfr using MDRD formula
2 Thyroperoxidase antibodies (AbTPO)	39	Hematocrit (HCT)
3 Albumin	40	Haemoglobin (HGB)
4 Alpha 1-fetoprotein (α1 fetoprotein)	41	Immunoglobulin A (IgA)
5 Alpha 1 globulin (α1 globulin)	42	Immunoglobulin G (IgG)
6 Alpha 2 globulin (α2 globulin)	43	Immunoglobulin M (IgM)
7 Alanine transaminase (ALT)	44	Lactate dehydrogenase (LDH)
8 Amylase	45	Lymphocytes
9 Aspartate aminotransferase (AST)	46	Lipase
10 Basophils	47	Bilateral mammography
11 Beta globulin (β globulin)	48	Mean cellular volume (MCV)
12 Beta 2 globulin (β2 globulin)	49	Microalbuminuria
13 Total bilirubin	50	Monocytes
14 Carbohydrate antigen 19-9 (CA 19-9)	51	Neutrophils
15 Calcium (Ca)	52	C-reactive protein (CRP)
16 Occult blood stool sample	53	Brain natriuretic peptide (BNP)
17 Carcinoembryonic antigen (CEA)	54	Platelets (PLT)
18 Creatinine clearance (Cockroft)	55	Potassium (K)
19 Chlorine (Cl)	56	Total proteins
20 HDL Cholesterol	57	Protein electrophoresis
21 LDL Cholesterol	58	Prostate specific antigen (PSA)
22 Total Cholesterol	59	Free prostate-specific antigen (free PSA)
23 Colonoscopy	60	Prothrombin time (PT)
24 Creatinine kinase (CK)	61	Erythrocytes (RBC)
25 Creatinine	62	Reticulocytes
26 Complete blood count (CBC)	63	Sodium (Na)
27 Eosinophils	64	Free triiodothyronine (T_3)
28 Hepatitis B surface antigen (HBsAg)	65	Free thyroxine (T_4)
29 Hepatitis C antibodies (HCV)	66	Thyrotropin (TSH)
30 Rheumatoid factor (RF)	67	Urea
31 Ferritin	68	Uric acid
32 Iron	69	Complete urine test
33 Vitamin $B9$ (folate)	70	Urine culture
34 Alkaline phosphatase (ALP)	71	Erythrocyte sedimentation rate (ESR)
35 Free/total prostate-specific antigen ratio (free/total PSA)	72	Vitamin $B12$ (cobalamin)
36 Gamma globulin (γ globulin)	73	Leukocytes (WBC)
37 Gamma glutamyl transferase (γGT)		

Fig 2: FIMMG Diabetes Disease Dataset

4. RESULTS

Models has been constructed using training data set (280 instances) which is 70% of original Type-II Diabetes data set. Constructed models have been validated using test data which is 30% of original data with respect to the parameter accuracy. Here, Accuracy has been calculated using confusion matrix .The best classifier model is the one with highest accuracy.

Accuracy of Decision tree Confusion Matrix has been generated by decision tree model for the test data (120 instances) with class (values: original Type-II Diabetes, Non original Type-II Diabetes)as the target variable. The confusion matrix clearly says that 7 instances are not classified properly and 113 instances have been classified accurately and the accuracy of this classifier model is **94.16%.**

Accuracy of SVM Confusion Matrix has been generated by SVM model for the test data(120 instances) with class as the target variable. The confusion matrix clearly says that 2 instances are not classified properly and 118 instances have been classified accurately and the accuracy of this classifier model is **98.33%**

Accuracy of Random Forest Confusion Matrix has been generated by SVM model for the test data (120 instances) with class as the target variable. The confusion matrix clearly says that 1 instances are not classified properly and 119 instances have been classified accurately and the accuracy of this classifier model is **99.36%.**

Accuracy of SB-SVM Confusion Matrix has been generated by SB-SVM model for the test data(120 instances) with class as the target variable. The confusion matrix clearly says that 2 instances are not classified properly and 118 instances have been classified accurately and the accuracy of this classifier model is **97.43%**

Accuracy of Linear Regression Confusion Matrix has been generated by Linear Regression model for the test data(120 instances) with class as the target variable. The confusion matrix clearly says that 2 instances are not classified properly and 118 instances have been classified accurately and the accuracy of this classifier model is **95.43%**

.

Paste your screenshots here

CONCLUSION

This project presented a prediction algorithm to predict Type-II Diabetes at an early stage. The dataset shows input parameters collected from the Type-II Diabetes patients and the models are trained and validated for the given input parameters. Decision tree, Random Forest and Support Vector Machine, Linear regression, SB-SVM learning models are constructed to carry out the diagnosis of Type-II Diabetes. The performance of the models are evaluated based on the accuracy of prediction. The results of the research showed that Random Forest Classifier model better predicts Type-II Diabetes in comparison to Decision trees and Support Vector machines and other algorithms. The comparison can also be done based on the time of execution, feature set selection as the improvisation of this research. The data distribution has properly covered the whole domain in Type-II Diabetes, but the general attributes towards Type-II Diabetes. The diagnosis of T2D at an early stage represents a key opportunity in order to prevent or significantly delaying devastating diabetes-related complications while alleviating the healthcare costs. The main contribution of this work is the introduction of the ML method, named SB-SVM, for discovering T2D in a novel collected EHR dataset (FIMMG dataset).We demonstrated the reliability of the proposed approach with respect to other ML and DL EHR based approaches widely employed in the state-of-the-art for solving this task. The proposed RF approach shows to be the best compromise between predictive performance and computation time. The SB-SVM is able to manage high dimensional data, by increasing the model interpretability and finding the most relevant features while dealing with the usual unbalanced class distribution. After training the model, it clearly shows that tree structures have higher accuracy than other classification algorithms, which can be justified from the distribution of the data set since the selected attributes have a clearer separation in the class attribute...

FUTURE WORK

We aim to validate our results by using big dataset or compare the results using dataset that contains the same features. Also, in order to help in reducing the prevalence of T2D, we plan to predict if a person with diabetes risk factors such as diabetes, hypertension, and family history of person will have T2D in the future or not by using appropriate dataset. In the future course of this study one can try to further improve the two-class classification accuracy by evaluating some hybrid or ensemble techniques, in addition to this a subset of features can be extracted from the complete medical data-set of T2D disease of parameters (features) without effecting the performance of the classification process, so that the financial burden a patient has to bear for undergoing various clinical tests can be reduced.

REFERENCES

[1] K. G. M. M. Alberti and P. Z. Zimmet, “Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHOconsultation.” Diabetic Medicine, vol. 15, no. 7, pp. 539–553, 1998.

[2] International Diabetes Federation, IDF Diabetes Atlas, 8th edn. Brussels,Belgium, 2017.

[3] WHO et al., Global report on diabetes. World Health Organization,2016.

[4] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth,S. C. Morton, and P. G. Shekelle, “Systematic review: impact of health information technology on quality, efficiency, and costs of medical care,”Annals of Internal Medicine, vol. 144, no. 10, pp. 742–752, 2006.

[5] R. Kaushal, K. G. Shojania, and D. W. Bates, “Effects of computerizedphysician order entry and clinical decision support systems on medi- cation safety: a systematic review,” Archives of Internal Medicine, vol.163, no. 12, pp. 1409–1416, 2003.

[6] R. Amarasingham, L. Plantinga, M. Diener-West, D. J. Gaskin, and N. R. Powe, "Clinical information technologies and inpatient outcomes:a multiple hospital study," Archives of Internal Medicine, vol. 169, no. 2,pp. 108–114, 2009.

[7] S. T. Parente and J. S. McCullough, "Health information technology and patient safety: evidence from panel data," Health Affairs, vol. 28, no. 2, pp. 357–360, 2009.

[8] C. Chen, T. Garrido, D. Chock, G. Okawa, and L. Liang, "The Kaiser Permanente Electronic Health Record: transforming and streamlining modalities of care," Health affairs, vol. 28, no. 2, pp. 323–333, 2009.

[9] D. Blumenthal, "Stimulating the adoption of health information tech- nology," New England Journal of Medicine, vol. 360, no. 15, pp. 1477–1479, 2009.

[10] D. Blumenthal and M. Tavenner, "The "meaningful use" regulation for electronic health records," New England Journal of Medicine, vol. 2010, no. 363, pp. 501–504, 2010.

[11] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," New England Journal of Medicine, vol. 375, no. 13, p. 1216, 2016.

[12] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.

[13] Z. C. Lipton, "The mythos of model interpretability," arXiv preprint arXiv:1606.03490, 2016.

[14] G. Sheikhi and H. Altınçay, "The cost of type ii diabetes mellitus: A machine learning perspective," in XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016. Springer, 2016, pp. 824–827.

[15] I. Kamkar, S. K. Gupta, D. Phung, and S. Venkatesh, "Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso," Journal of Biomedical Informatics, vol. 53, pp. 277–290,2015.

[16] B. H. Cho, H. Yu, K.-W. Kim, T. H. Kim, I. Y. Kim, and S. I.Kim, "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods," Artifi-cial Intelligence in Medicine, vol. 42, no. 1, pp. 37–53, 2008.

[17] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," Artificial Intelligence in Medicine, vol. 41, no. 3, pp. 251–262, 2007.

[18] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International Journal of Medical Informatics, vol. 97, no. Supplement C, pp. 120–127, 2017.

[19] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," Machine Learning, vol. 53, no. 1-2, pp. 23–69,2003.

[20] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes,"
Procedia Computer Science, vol. 82, pp. 115–121, 2016.

[21] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," BMC Medical Informatics and Decision Making, vol. 10, no. 1, p. 16, 2010.

[22] Y. Wang, P. F. Li, Y. Tian, J. J. Ren, and J. S. Li, "A shared decision making system for diabetes medication choice utilizing electronic health record data," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 5, pp. 1280–1287, 2017.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer,"SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[24] K.-M. Jung, "Support vector machines for unbalanced multicategory classification," Mathematical Problems in Engineering, vol. 2015, 2015.

[25] J. Bi, Y. Chen, and J. Z. Wang, "A sparse support vector machine approach to region-based image categorization," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 1121–1128.

[26] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," Journal of the American statistical Association, vol. 96, no. 456, pp. 1348–1360, 2001.

[27] H. Masnadi-Shirazi, N. Vasconcelos, and A. Iranmehr, "Cost-sensitivesupport vector machines," arXiv preprint arXiv:1212.0975, 2012.

[28] G. Lee and C. Scott, "Nested support vector machines," IEEE Transactions on Signal Processing, vol. 58, no. 3, pp. 1648–1660, 2010.

[29] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning,vol. 20, no. 3, pp. 273–297, 1995.

[30] J. Zhu, S. Rosset, R. Tibshirani, and T. J. Hastie, "1-norm support vector machines," in Advances in neural information processing systems, 2004, pp. 49–56.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso,"Journal of the Royal Statistical Society. Series B (Methodological), pp.267–288, 1996.

[32] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines." in International Conference on Machine Learning, vol. 98, 1998, pp. 82–90.

[33] T. Hastie, R. Tibshirani, and M. Wainwright, Statistical learning with sparsity: the lasso and generalizations. CRC press, 2015.

[34] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," IEEE Transactions on Signal Processing, vol. 57, no. 7, pp. 2479–2493, 2009.

[35] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward backward splitting," Multiscale Modeling & Simulation, vol. 4, no. 4, pp. 1168–1200, 2005.

[36] S. Sra, S. Nowozin, and S. J. Wright, Optimization for Machine Learning. The MIT Press, 2011.

[37] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, vol. 57, no. 11, pp. 1413– 1457, 2004.

[38] M. A. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," IEEE Journal of Selected Topics in Signal Processing, vol. 1, no. 4, pp. 586–597, 2007.

[39] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," Advances in Large Margin Classifiers, pp. 61–74, 2000.

[40] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," ACM Sigkdd Explorations Newsletter, vol. 6,no. 1, pp. 1–6, 2004.