INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X

Impact factor: 4.295

(Volume3, Issue1)

Available online at: www.ijariit.com

Novel Approach Of Diabetes Disease Classification By Support Vector Machine With RBF Kernel

Preeti Verma
M.Tech Student, Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India.
Vermapreeti008@gmail.com

Inderpreet Kaur
Assistant Professor, Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India.
inderpreet.kaur029@gmail.com

Jaspreet Kaur
Assistant Professor, Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India.
jaspreetkaur.rb@gmail.com

Abstract*— *Early diagnosis of any disease with less cost is always preferable. Diabetes is one such disease. It has become the fourth leading cause of death in developed countries and is also reaching epidemic proportions in many developing and newly industrialized nations. Diabetes leads to increase in the risks of developing kidney disease, blindness, nerve damage, blood vessel damage and heart disease also. In this research work, Support Vector Machine with RBF Kernel is used for finding out the classification accuracy of diabetes dataset. In the given method, the advance algorithm of SVM-RBF kernel is used; it contains some of the extended parameters for feature selection as well as the proposed correlation with SVM method obtains on UCI dataset. In this work, investigation is done on automatic approach to diagnose diabetes disease based on Support vector machine with RBF kernel and MLP (Multilayer perceptrons). The concept of data mining is used, in which the proposed SVM-RBF method obtains 88% accuracy on UCI diabetes dataset, which is better than other models.

Keywords*— *Diabetes, Nmachine Learning, Svm, Feature Selection.

I. INTRODUCTION

A classification of diabetes and other categories of glucose intolerance, based on contemporary knowledge of this heterogeneous syndrome, was developed by an international workgroup sponsored by the National Diabetes Data Group of the NIH. This classification and revised criteria for the diagnosis of diabetes were reviewed by the professional members of the American Diabetes Association, and similar versions were circulated by the British Diabetic Association, the Australian Diabetes Society, and the European Association for the Study of Diabetes. The ADA has endorsed the proposals of the international workgroup, and the Expert Committee on Diabetes of the World Health Organization has accepted its substantive recommendations [9]. Diabetes is two types Type I diabetes and Type II diabetes. In children and young adults Type I diabetes is usually diagnosed, and was previously known as juvenile diabetes [4]. Owing to the beta cells destruction of the pancreas Type I diabetes mellitus (IDDM) patients not producing insulin. Type II diabetes is the most common form of diabetes. Insulin endogenously is produced by Type II diabetes mellitus (NIDDM) patients and as compare to health subjects, this insulin effect and secretion are reduced [5]. For the diabetes currently cure does not exist, and the nearest possible normal values in the blood maintain their glucose levels, then take health care of the people affected is the only option The body unable to produce or properly use the hormone called insulin that unlocks the cells of the body and allows glucose to enter and fuel them in diabetes condition. To diagnose the diabetic patient there are many factors which need to be analysed and physician's job become difficult by these factor. Various techniques have been analysed in the past to classify diabetic patients and predict the accuracy. But in this works implement an efficient method for classification of patients for diabetes using soft computing technique. In this work is to improve the accuracy of diabetes dataset and reduce the error.

II. LITERATURE REVIEW

Baek Hwan Cho *et.*al[1] in this paper to accurately predict the onset of diabetic nephropathy, it applied various machine learning techniques to irregular and unbalanced diabetes dataset, such as support vector machine (SVM) classification and feature selection methods. Visualization of the risk factors was another important objective to give physicians intuitive information on each patient's clinical pattern. Methods and materials: It collected medical data from 292 patients with diabetes and performed pre-processing to extract 184 features from the irregular data. To predict the onset of diabetic nephropathy, we compared several classification methods such as logistic regression, SVM, and SVM with a cost sensitive learning method. It also applied several feature selection methods to remove redundant features and improve the classification performance. For risk factor analysis with SVM classifiers, it has developed a new visualization system which uses a nomogram approach.

Thomas A. Buchanan *et.*al [2] this paper is proposed currently conducting these assessments longitudinally in a cohort of Latino women with GDM. The present report details the relationship between important regulators of glucose tolerance during the index pregnancy and the presence of impaired glucose tolerance (IGT) or type 2diabetes within 6 months after delivery.

Henry S. Kahn *et.*al [3] this paper objective is to derive and validate scoring systems by using longitudinal data from a study that repeatedly tested for incident diabetes. Anthropometry, blood pressure, and pulse (basic system) plus a fasting blood specimen assayed for common analytes (enhanced system). Diabetes was identified in 18.9% of participants. Risk score integer points were derived from proportional hazard coefficients associated with baseline categorical variables and quintiles of continuous variables. This paper has a limitation the risk scoring systems had no question regarding previous gestational diabetes, and knowledge of parental diabetes may be uncertain. The analysed cohort was restricted by age and race; the systems may be less effective in other samples.

V Mohan *et.*al [4] The aim of this study was to develop and validate a simplified Indian Diabetes Risk Score for detecting undiagnosed diabetes in India. The risk score was derived from the Chennai Urban Rural Epidemiology Study (CURES), an ongoing epidemiological study on a representative population of Chennai. Phase 1 of CURES recruited 26,001 individuals, of whom every tenth subject was requested to participate in Phase 3 for screening for diabetes using World Health Organization (WHO) 2hour venous plasma glucose criteria [i.e. ≥200 mg/dl]. The response rate was 90.4% (2350/2600). The Indian Diabetes Risk Score [IDRS] was developed based on results of multiple logistic regression analysis. Internal validation was performed on the same data.

Maria Lönnrot *et.* Al [5] in this paper have studied the frequency of enterovirus infections in 21 DIPP cohort children who developed autoantibodies during follow-up and in their control subjects who were matched for the time of birth, sex, and HLA risk alleles. For comparison, adenovirus infections were analysed in the same children. Short sample intervals during the follow-up, together with a large panel of sensitive virological assays, create an optimal setting for the evaluation of the role of these infections within the framework of the present study design.

Martin Fiichtenbusch *et.*al [6] In this paper, examined the frequency of ICAs, GADAs, and IA2As in nondiabetic pregnant women to define the normal distribution of antibody levels during pregnancy. Then correlated the presence of ICAs, GADAs, and IA2As in the serum of women with GDM at delivery with the progression to type 1 diabetes postpartum. Different screening strategies with these autoimmune markers were further evaluated in an attempt to identify women at high risk for the disease.

Jason L. Vassy *et.* Al [7] in this paper tested three primary hypotheses. First, it hypothesized that an updated total genotype risk score (GRSt) with up to 65 T2D-associated risk loci improves the prediction of incident T2D in young and middle-aged adulthood compared with previously published scores with fewer loci. It examined genotype-only and also genotype-plus-clinical prediction models. Second, because b-cell dysfunction and IR represent two distinct pathways in the pathogenesis of T2D, hypothesized that separate GRS consisting of SNPs postulated to influence b-cell (GRSb) or IR (GRSIR) independently predict incident T2D. In subsidiary analyses, it investigated whether GRSb and GRSIR together exhibit a multiplicative effect on T2D risk and whether the association between T2D risk and GRSb or GRSIR varies between lean and obese individuals. Third, hypothesized that the relationships between incident T2D and GRSt, GRSb, or GRSIR do not differ between black and white individuals.

William H. Herman *et.*al [8] the objective of this paper is to develop a simple questionnaire to prospectively identify individuals at increased risk for undiagnosed diabetes. People with newly diagnosed diabetes (n = 164) identified in the Second National Health and Nutrition Examination Survey and those with neither newly diagnosed diabetes nor a history of physician diagnosed diabetes (n = 3,220) were studied. Major historical risk factors for undiagnosed non-insulin-dependent diabetes were defined, and classification trees were developed to identify people at higher risk for previously undiagnosed diabetes. The sensitivity, specificity, and predictive value of the classification trees were described and compared with those of an existing questionnaire.

III. PROBLEM FORMULATION

Diabetes is a major health problem not only in industrial but developing Countries as well and its incidence is inclining. It is a condition in which the body unable to produce or properly use the hormone called insulin that ''unlocks" the cells of the body and allows glucose to enter and fuel them. There are many factors which need to be analysed to diagnose the diabetic patient, and this makes the physician's job difficult. So we will implement an efficient method for classification of patients for diabetes using soft computing technique. Our major concern is to improve the accuracy of diabetes dataset. Various techniques have been analysed in the past to classify diabetic patients and predict the accuracy.

IV. OBJECTIVES

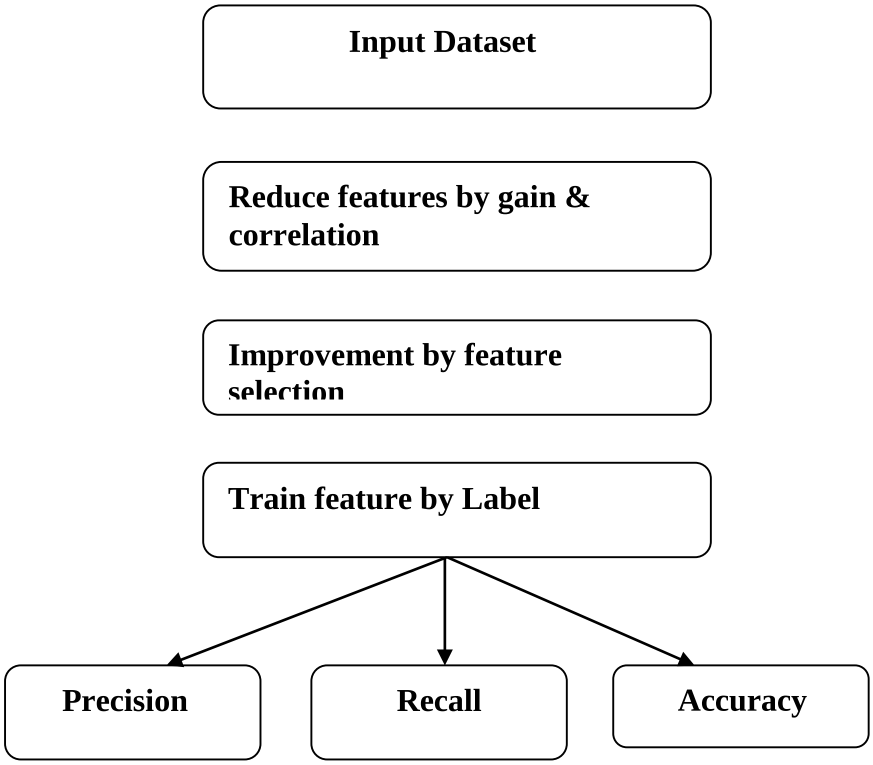
The foremost purpose is to Improve Default Prediction in Software module by using feature extraction and using SVM with RBF approach.

- Another aim is to study the prediction of Diabetes and Chronic Disease, after that Collect dataset of diabetes Disease patient.
- Additionally, one more goal is to propose and implement feature selection by information gain and then correlation of result with the convolution Support vector machine with RBF kernel.

- To implement the proposed technique with Support vector machine (SVM), Radial basis function (RBF) kernel, polynomial and multi perceptron kernel.
- To analyze the approach by precision, recall and then accurately compare them with existing methods.

V. PROPOSED METHODOLOGY

Step1: Input the collected dataset of patient with features.
Step2: Reduce the features by feature selection using information gain and correlation.
Step3: Improve the feature by feature extraction.
Step4: Train the features with label to SVM.
Step5: Analysis, precision, recall, accuracy.



VI. RESULT

Table 1

Classifier	Accuracy (intersection)	Precision(intersection)	Recall(intersection)	Error(intersection)
Svm Linear	**66**	**69**	**59.8**	**32**
SVM Polynomial	**67**	**65**	**67**	**33**
SVM Quadratic	**63**	**66**	**62**	**37**
SVM RBF	**80**	**65**	**70**	**20**

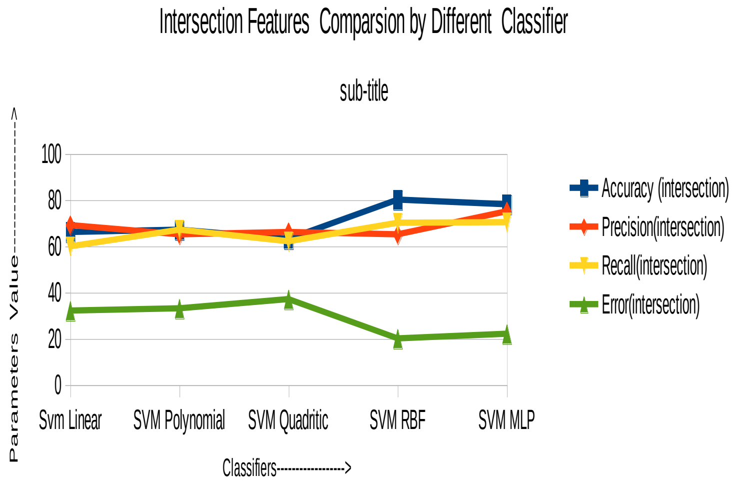


Figure 1: Comparison of intersection features by different classifier

Table 2

Classifier	Accuracy (union)	Precision(union)	Recall(union)	Error(union)
Svm Linear	**77**	**74**	**69**	**23**
SVM Polynomial	**63**	**73**	**60**	**37**
SVM Quadratic	**78**	**75**	**72**	**22**
SVM RBF	**88**	**83**	**70.23**	**12**
SVM MLP	**84**	**80**	**69.45**	**16**

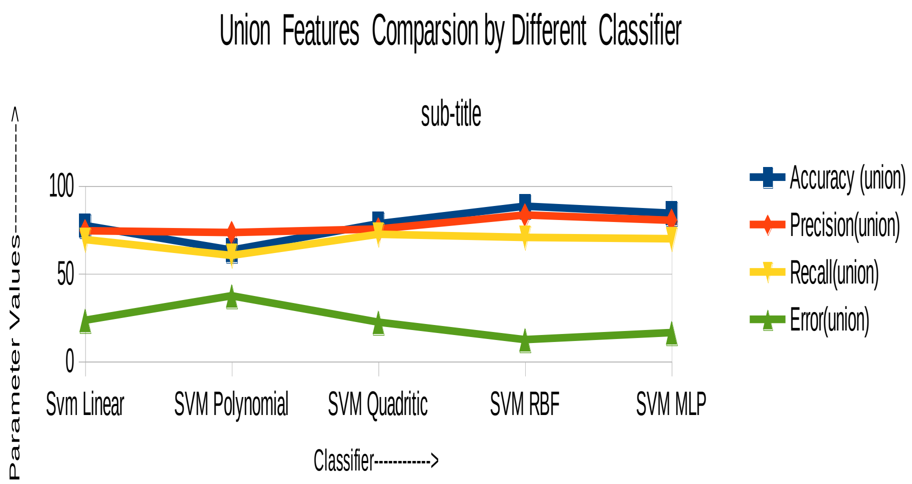


Figure 2: Union Features classifier comparison by different

Table 3

Classifier	Accuracy (all)	Precision(all)	Recall(all)	Error(all)
Svm Linear	**77.23**	**74**	**69**	**22**
SVM Polynomial	**67.89**	**73**	**60**	**32**
SVM Quadratic	**80.34**	**76**	**79**	**19**
SVM RBF	**90.23**	**83**	**80.34**	**9**
SVM MLP	**87.23**	**80**	**70**	**13**

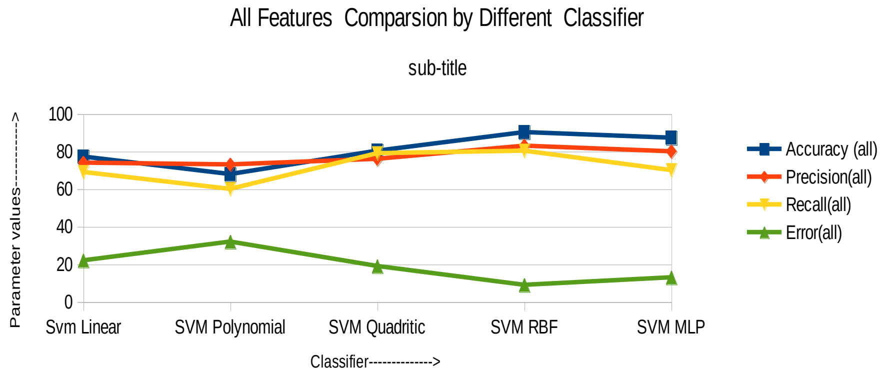


Figure 3: All feature comparison by different classifier

CONCLUSION

The classification accuracy SVM-RBF Kernel selected features are applied. It improves the accuracy of the model that is shown by the results. Moreover, Outcomes conclude that the proposed model is significantly more reliable, faster and takes less time as compared to the model proposed in previous works. This method can be coupled with medical software's to assist physicians to make more accurate decisions about Diabetes disease.

REFERNCES

[1] Cho, Baek Hwan, et al. "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods." *Artificial intelligence in medicine* 42.1 (2008): 37-53.
[2] Buchanan, Thomas A., et al. "Gestational diabetes: antepartum characteristics that predict postpartum glucose intolerance and type 2 diabetes in Latino women." *Diabetes* 47.8 (1998): 1302-1310.
[3]Kahn, Henry S., et al. "Two risk-scoring systems for predicting incident diabetes mellitus in US adults age 45 to 64 years." *Annals of Internal Medicine* 150.11 (2009): 741-751.
[4]Mohan, V., et al. "A simplified Indian Diabetes Risk Score for screening for undiagnosed diabetic subjects." *The Journal of the Association of Physicians of India* 53 (2005): 759-63.
[5]Lönnrot, Maria, et al. "Enterovirus infection as a risk factor for beta-cell autoimmunity in a prospectively observed birth cohort: the Finnish Diabetes Prediction and Prevention Study." *Diabetes* 49.8 (2000): 1314-1318.
[6]Füchtenbusch, Martin, et al. "Prediction of type 1 diabetes postpartum in patients with gestational diabetes mellitus by combined islet cell autoantibody screening: a prospective multicenter study." *Diabetes* 46.9 (1997): 1459-1467.
[7]Vassy, Jason L., et al. "Polygenic type 2 diabetes prediction at the limit of common variant detection." *Diabetes* (2014): DB_131663.
[8]Herman, William H., et al. "A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes." *Diabetes care* 18.3 (1995): 382-387.
[9]National Diabetes Data Group. "Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance." *Diabetes* 28.12 (1979): 1039-1057.
[10]Butler, Alexandra E., et al. "β-cell deficit and increased β-cell apoptosis in humans with type 2 diabetes." *Diabetes* 52.1 (2003): 102-110.

