Recursive Feature Elimination with Ridge Regression (L2) Machine Learning Hybrid Feature Selection Algorithm for Diabetic Prediction using Random Forest Classifer.

K venkatachalam (venkatme83@gmail.com)
VIT Bhopal University
P Prabhu
Christ University
B saravana Balaji
Lebanese French University
Mohamed Abouhawwash
Mansoura University
R Rajadevi
Kongu Engineering College

Research Article

Keywords: ridge regression, recursive feature elimination, random forest, machine learning, Feature selection.

Posted Date: July 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-742641/v1

Recursive Feature Elimination with Ridge Regression (L2) Machine Learning Hybrid Feature Selection Algorithm for Diabetic Prediction using Random Forest Classifer.

K Venkatachalam[1], P Prabhu[2], Saravana Balaji B[3], Mohamed Abouhawwash*,[4,5],R Rajadevi[6]

[1]Senior Assistant Professor, School of Computer Science and Engineering,VIT Bhopal University,Bhopal.
[2]Department of Computer Science, CHRIST (Deemed to be University), Hosur Road, Bengaluru.
[3]Assistant Professor, Dept of Information Technology, Lebanese French University.
[4]Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt;
[5]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, 48824 USA;
Assistant Professor(SL),Dept of information Technology, Kongu Engineering College,perundurai
venkatme83@gmail.com, prabu.p@christuniversity.in, saravanabalaji.b@gmail.com, saleh1284@mans.edu.eg, abouhaww@msu.edu,rajdevi.it@kongu.edu

Abstract:

In day today life, diabetes illness is increasing in count due to the body not able to metabolize the glucose level. The prediction of the right diabetes patients is an important research area that many researchers are proposing the techniques to predict this disease through data mining and machine learning methods. In prediction, feature selection is one of the key concept in preprocessing so that the features that are relevant to the disease will be used for prediction. This will improve the prediction accuracy. Selecting right features among the whole feature set is a complicated process and many researchers are concentrating on it to produce the predictive model with high accuracy. In this proposed work, the wrapper based feature selection method called Recursive Feature Elimination (RFE) is combined with Ridge regression (L2) to form a hybrid L2 regulated feature selection algorithm to overcome the overfilling problem of the data set. Over fitting is the major problem in feature selection which means that the new data are not fit to the model since the training data is small. Ridge regression is mainly used to overcome the overfitting problem. Once the features are selected using the proposed feature selection method, random forest classifier is used to classify the data based on the selected features. The proposed work is experimented in PIDD data set and the evaluated results are compared with the existing algorithms to prove the accuracy effect of the proposed algorithm. From the results obtained by proposed algorithm, the accuracy of predicting the diabetes disease is high compared to other existing algorithms.

Key terms: ridge regression, recursive feature elimination, random forest, machine learning, Feature selection.

1. INTRODUCTION:

Supervised learning methods can be divided into classification and regression problems. The continuous problem can be predicted easily using regression method. The dataset is collection of information with samples and parameters. If we have fewer samples with more number of parameters the ridge regression can be efficiently used to get best solution. Its important to understand bias and variance in machine learning context. Bias is model plots the inline nearby samples. The differences between datasets are fits is called variance. If the lines in graph are squiggly proceeds then it is said to be high variance and if line proceeds in straight then it is said to be low variance. When we introduce sum of square in linear regression then it is said to ridge regression.

Some feature based model is trained in machine learning algorithms [16]. Accuracy of selecting the feature for new data is very less. The main problems in new data is underfitting and overfitting.

1.1 PROBLEM FORMULATION

When the training model is fitted with huge data features, then the model gets overfitting problem. Model is trained with less data features, then the machine learning will be biased [17, 18]. It leads to underfitting problem. If the model is trained with more number of features then the high variance will be come into picture. It leads to less efficiency in identification of suitable model. This problem is defined as overfitting. This is a high model issue.

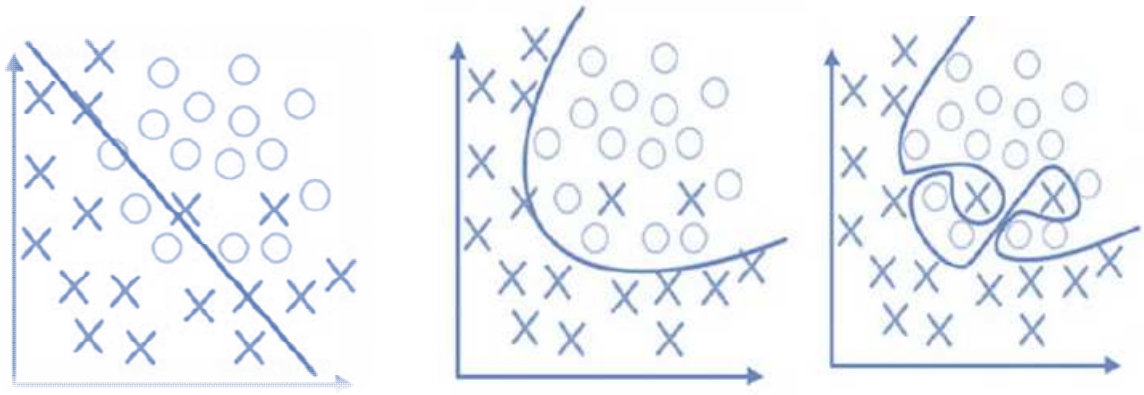

Figure 1 : a) underfitting *b) actual fitting* *c) over fitting*

Problem in few data feature processing to train the model leads to predict wrong unknown data as shown in figure a. some feature selection models are suggested to reduce underfitting model [19,20].

Two terms bias and variance come into overfitting and underfitting model is explained in figure 1(a,b,c) . Bias is a process used to match the correct value by calculating the differences. Variance is the prediction model for given features at various realization views.

Further over fitting model can be overcome by selecting more features and bias function. High bias machine learning algorithms requires to solve the problems. In our article, we combine ridge regression with recursive feature elimination model to reduce the overfitting problems. When training error is less than testing error, then there is overfitting problem. The conventional methods L1,L2 regularization are used to minimize overfitting problems. But the outcome efficiency is very

less. To improve the accuracy of the feature selection we merge recursive feature elimination model with L2 regularization. Then the output feature is further classified using random forest algorithm.

The remaining section of the paper is organized as following: chapter 2 studied about literature survey of existing implementations. Chapter 3 explains about proposed model in the article. Next chapter 4 discusses about the outcome of our proposed research model. Finally chapter 5 concludes the work.

2. Literature Review

The paper [1] listed the survey on feature selection algorithms such as K-NN, K means, Naïve Bayes etc., This paper use the common diabetic data set and the results of the algorithms are analyzed and they suggested the best algorithm based on the performance accuracy. As a survey result, they conclude that Branch ad Bound algorithm gives high level of accuracy compared to the other eight algorithms such as Naïve Bayes, SVM, C4.5, kNN, K means, Randomized Hill climb and Simulated Annealing.

A survey of various feature selection methods are listed in paper [2]. This paper introduced feature selection based on Genetic Algorithm to detect and diagnose the biological issues. This paper gave detailed description about the types of feature selection algorithms such as filter based, wrapper and embedded feature selection algorithms. They experimented the result of the algorithms on five bench mark datasets from UCI repository. They conclude that among the three feature selection method wrapper based methods are perform well to reduce the features. This paper also discuss about the challenges in feature selection.

The paper [3] proposed a feature selection algorithm based on L1 (Lasso) and classification on microarray cancer data by the use of Random Forest. They experimented the proposed algorithm on eight standard dataset of microarray cancer dataset. The learning proficiency of the classifier is explored using the learning curve model called fivefold cross validation during the training phase. The comparative result of the proposed shows the best accuracy level than the recent research works. The evaluation is performed with the help of the accuracy, recall, precision, f measure and confusion matrix.

The objective of the paper[4] is to propose a prediction model with high sensitivity and selectivity. The prediction on diabetes mellitus disease of Canadian patients based on the patients lab results while they visit to the medical services. The proposed model based on logistic regression and Gradient Boosting Machine approaches. To improve the sensitivity of the algorithm, the authors are implementing the method with adjusted threshold and class weight. The proposed model is compared with Decision Tree and Random forest in terms of AROC. The proposed GBM and logistic regression performed better accuracy of 87% and 84% compared to other existing algorithms.

The paper [5] focuses on the regularization of the embedded feature selection algorithms such as ridge regression, Lasso regression and combination of both. They evaluated the algorithms using five large dimensional datasets. They experimented and compare them in terms of sparsity, correlation and execution time. From the result observation of the proposed work, L21 performs better in sparse dataset, SC have high BSR, with non sparse data sets, LL gave BSR high, EN have higher BSR rate than L21, for dense data sets LL, L1 SVM, and EN give best result. In terms of execution time EN performs better than other algorithms. So the based on the dataset the feature selection methods also differs.

In the paper [6] the feature selection algorithms are embedded in the SVM learning. They proposed two algorithms such as L1 regularization of weight and RFE extension. These algorithms are used to classify the multi class problems. They proved the efficient optimization technique based on information gain. The study [7] use decision tress, Neural network and random forest to predict diabetes mellitus disease. The model is examined by five fold cross validation. This experiment selects 68994 healthy persons as training data. To overcome the data unbalance, they experimented five times the original data and average of these five experiments take as final result. To reduce the dimensionality of the features PCA and mRMR are used. While all 14 attributes are used, random forest give best accuracy of 80% than others.

To identify the set of prognostic genes, the paper [8] proposed a novel random forests algorithm called Random Survival Forest algorithm. This proposed algorithm form many binary trees each

constructed based on deterministic techniques. It uses several split criteria such as log rank, log-rank score, conserve and random. According to the predictor variable values, each observation is assigned as leaf node or terminal node. This gene selection based method is combined to multivariate correlations in microarray data set. And this research work performed well in terms of simulation and real data.

The proposed algorithm of paper [9] use Recursive Feature Elimination and PCA as feature reduction technique and deep neural network and Artificial Neural Network as classifier to design an expert system to predict the diabetes disease. The proposed work is compared with other existing machine learning algorithms in terms of accuracy, sensitivity and specificity. The analysis conclude that RFE performs better in feature reduction and compare to ANN, deep neural network classify the data with high accuracy. The paper [10] gives survey about the diabetes prediction using feature selection methods and classification. They analyzed F score, GA, SVM and ANN and they conclude their experimental analysis as embedded feature selection method called f score are performed better than other algorithms.

The paper [11] predict the diabetes patient based on two steps. First to select the relevant attributes, weighting methods are used and second as classification based on AdaBoost, Gradient Boosted and Random Forest algorithms. The experimental results shows the better accuracy using stability selection and AdaBoost algorithms. Paper [12] user Fisher's score, Recursive Feature Elimination and decision tree to select features and Random Forest, Regression, SVM and MLP to predict the diabetes disease. They experimented using PIDD data set and the result with high accuracy of 98% is obtained while Random Forest is used as a classifier with 19 features. Then the features are reduced with feature selection algorithm to five as new data set. The proposed work obtained high accuracy than others.

Paper [13] identified the insulin resistance using non invasive approaches of machine learning techniques. Experimented the work with CALERIE data set with 18 parameters such as age, gender,height etc., The selected attributes of feature selection is given as input to the classification algorithms such as logistic regression, CART, SVM,LDA,KNN etc., the analysis results shows high accuracy of 97% to identify the insulin resistance while using logistic regression and SVM.

Paper [14] proposes an SVMRFE model by the modification of SVM. It just rank the genes of the data based on the discriminatory power and the gene not participated are removed. They form the gene regulatory network that have the gene that are used to identify the diabetes disease have the top rank. This proposed method is experimented with type II diabetes. The genes involved in the proposed study will cause of the disease.

In the paper [15] the standard SVM with RFE were studied. In this work the correlation bias reduction (CBR) are combined with feature selection process. They conducted the experiments on breath analysis dataset. Comprehensive attributes are selected and classification peroformed. The proposed SVM-RFE with CBR perform better than the original SVM with RFE algorithms. An ensemble version of the proposed study is the suggestion of the next work to improve the stability of the proposed work.

3. PROPOSED HYBRID L2-RFE METHODOLOGY:

Diabetes mellitus is an illness caused by the body not able to metabolize the glucose level. In short future there will be an increase in count of the diabetes patient. There are many researchers or proposing predictive models to predict the diabetes mellitus disease at earlier stage. We are in the still in research to find good predictive model in terms of high accuracy to predict the diabetes disease. Feature Selection is an important preprocessing step to find the relevant features for classification. In this proposed work, the wrapper based feature selection method called Recursive Feature Elimination (RFE) is combined with Ridge regression (L2) to form a hybrid L2 regulated feature selection algorithm to overcome the overfilling problem of the data set. Once the features selected using the proposed method, random forest classifier is used to classify the data based on the selected features. The overall architecture of the proposed work is shown in Fig 2. Diabetes Data is used as an input. The proposedwork contains two part such as, Feature selection using L2 regulated RFE and Classification using Random Forest.

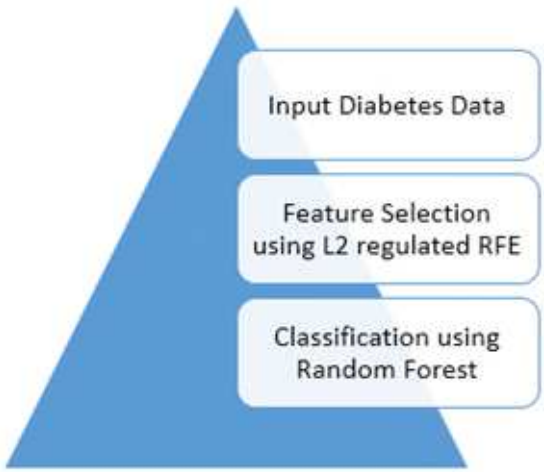

Fig 2 Overall architecture of the proposed work

3.1 L2 Regulated Recursive Feature Elimination:

L2 regulated Feature selection is one of the embedded feature selection methods that uses wrapper based ridge regression with Recursive Feature Elimination as a feature selection algorithm. overfitting is the major problem in Feature selection which means the new data are not fit to the model since the training data is small. Ridge regression is mainly used to overcome the overfitting problem that are widely combine with the linear feature selection model to select the relevant features from the data set for further processing. RFE is select the features of the dataset by recursively executing the process to select the smaller amount of features based on the coefficient value or important of the features. So that the least important attributes are pruned from the data set. This process is repeatedly considering to prune the maximum amount of irrelevant data to produce the minimum amount of relevant data for classification. The steps involved in L2RFE includes,

1. Fitting the model using Ridge Regression
2. Rank the important features
3. Discard the least important feature
4. Refit the model until find the relevant features.

L2 regulated feature selection model shown in eqn 3 is formulated by add the squared magnitude of coefficient as the penalty shown in eqn 1 to the loss functioneqn 2.

$$p = \lambda \sum_{j=1}^{d} |w_j|^2 \quad (1)$$

$$LF = \sum_{i=1}^{n} \left(y_{i,j} - \sum_{j=1}^{d} w_{i,j} * X_i\right)^2 \quad (2)$$

$$\text{L2RFE} = \sum_{i=1}^{n} \left(y_{i,j} - \sum_{j=1}^{d} w_{i,j} * X_i\right)^2 += \lambda \sum_{j=1}^{d} |w_i|^2 \quad (3)$$

Where p= penalty,

λ= control variable

$|w_j|$= coefficient of the jth feature sample

This proposed L2 regulated Feature selection is applied on Diabetes data to select the relevant features. This is an efficient technique to select the relevant and optimal features that are having the weight as minimal. This L2RFE used λ to control the features in order to select the number of important features based on the value. If λ value is smaller then few features will be selected and if λ value is high then larger number of features will be selected. Choosing the value for λ is important. High value for λ leads to increase the weight that causes under fitting. If the weight of the features is zero then that features are considered to be irrelevant and non-zero weight features are considered to be relevant. This algorithm enforces to select the minimum valued feature to have zero coefficient as the optimal features for further processing.

The main advantage of this ridge regression based RFE is to overcome the overfitting problem. To solve the overfitting problem feature scaling is much important step. Eqn 1 is used for feature scaling between 0 and 1.The transformed scale value of the features is

$$f_{new} = (X_i - \mu)/\sigma \quad (4)$$

Where,

X_i= ith feature of the dataset

μ=feature vector Mean

σ=feature vector standard deviation

3.2 Proposed hybrid Algorithm :

The proposed L2-RFE algorithm steps are stated as follows. Since the Recursive Feature Elimination is backward selection, low value of λ are reduced recursively until the optimum number of features selected. This algorithmic steps are programmatically implemented in Scikit-learn machine learning in python.

Step 1: Start

Step 2: Input Diabetes data set

Step 3: for all features 1:n

Step 4: Fit the features to the proposed **L2FRE** model using eqn 3.

Step 5: end for

Step 6: the resultant features are transformed using Eqn 4

Step 7: Feature scaling using eqn 1

Step 8: scaled features fit into Random forest classifier using algorithm2

Step 9: shown the Predicted result

Step 10: Stop

3.3 Proposed Workflow on Feature Selection:

The workflow of the proposed L2-RFE based feature selection to predict the diabetes mellitus is shown in Figure 3. The diabetes mellitus data are given as input to the algorithm. Each feature in the dataset is fit to the proposed L2RFE model to reduce the features having low important. Selected relevant features are scale down using eqn 4 to overcome the overfitting problem. The data set is divided into test set and training set. The scaled features are fitted into the Random forest Classifier to predict the diabetes disease.

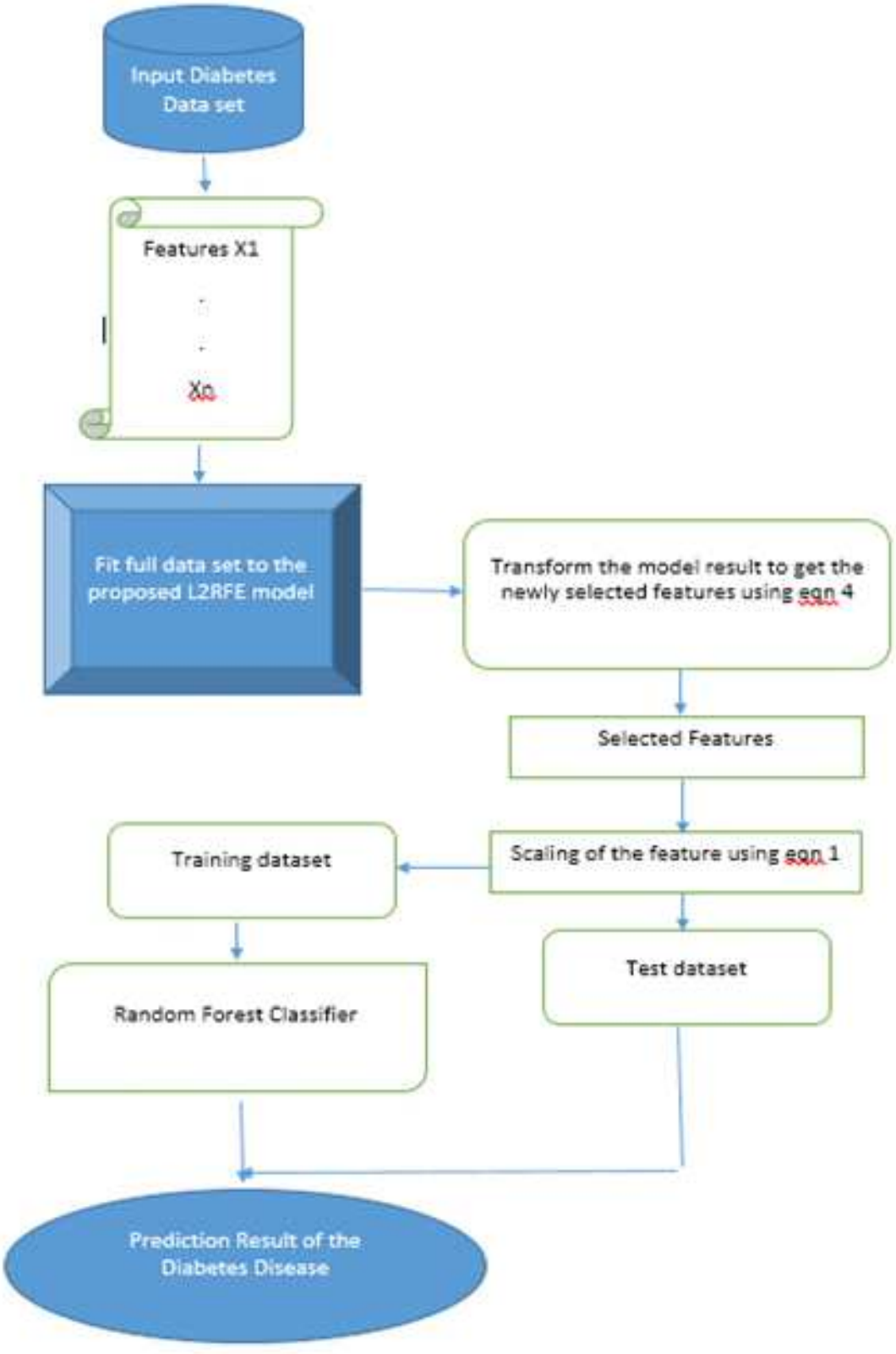

Fig 3: Workflow of the proposed work

3.4 Random Forest Classifier:

Random forest classifier is an unbiased classification model which consists of group of decision tress with average noise. This is a one such a model to give predictive accuracy higher in binary classification problem. The entire data set is divided as sub data set and each sub data set are trained in d number of decision tress shown in Fig 3. Each decision trees are trained separately and predict the result of the sub data set. Final prediction is based on the majority of the subset predictive result. The probability of each subset of a predictive class is expressed in eqn 5,

$$prob(c|f) = p_1 + p_2 + \cdots + p_n \sum_{i=1}^{n} \left(p_i(c|f)\right) \quad (5)$$

Where,

c = class

f = features

$p_1 \ldots p_n$ = probability of each each feature and class $(c|f)$

n= number of sub data set

3.5 RF Algorithm :

Input : Raw Diabetes Mellitus Data set

Output : Predictive report of the Diabetes disease

1. Randomly select k features from the total number of n features.
2. Compute the node d based on best split algorithm for all k features.
3. Split the d node into child node using best split
4. Repeat step 1 to 3 until n reached
5. Repeat step 1 to 4 until n number of trees created
6. Predict the class label of all decision trees and compute the vote to find the maximum vote for the class label
7. The most frequently predicted label is the final prediction result.

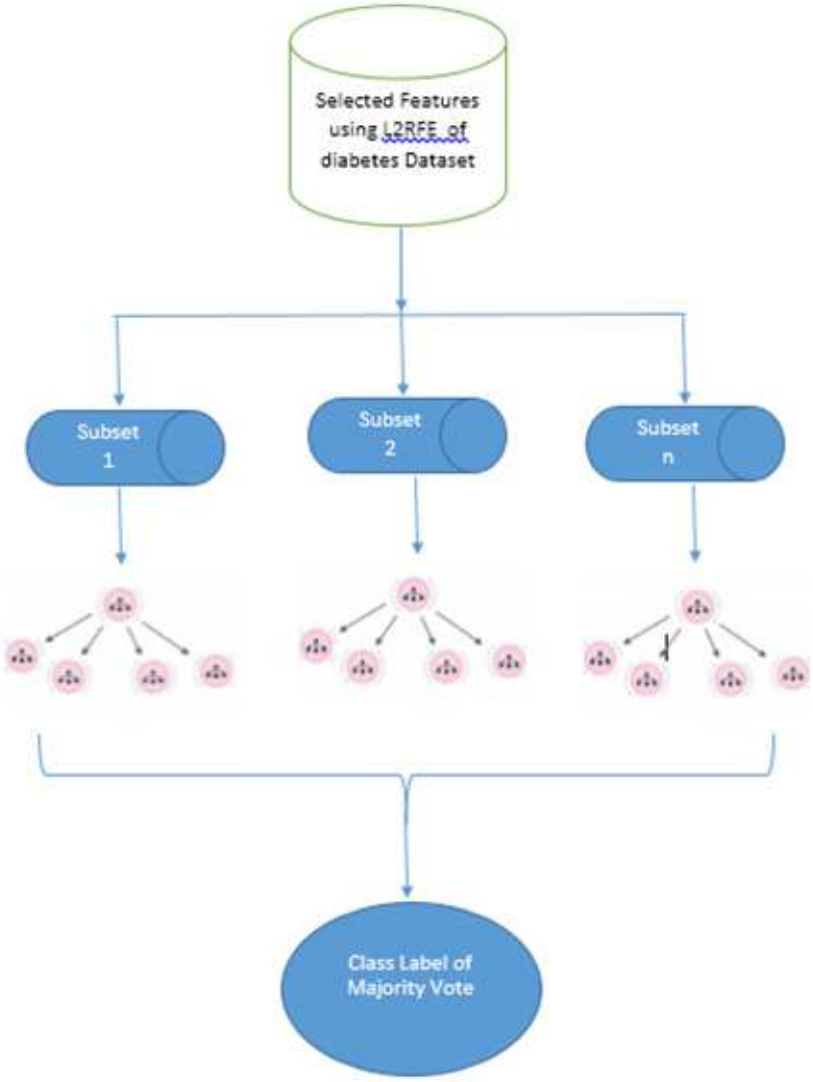

Fig 4: Work flow of Random Forest Classifier

4. EXPRIMENTAL RESULTS AND DISCUSSIONS

4.1 DATASET EVALUATION

The proposed work on hybrid L2 regulated Recursive Feature Elimination based feature selection on diabetes prediction using random forest classification is experimented in Pima Indians Diabetes Data Set (PIDD) data set. Which is open source data set available in UCI repository [5]. The data set consist of 768 samples including one class attribute to indicate the diabetes positive and negative. There are 267 positive samples and 500 negative samples in the data set. The attributes are shown in Table 1

Table 1: Attributes of PIDD

S.No	Attribute	Description
1	Age	Age of a person
2	Gender	Male or female
3	Plasma glucose fasting	
4	Plasma glucose post prandial	
5	Pregnancy	Pregnancy count of women
6	Blood glucose level	Plasma glucose concentration a 2 h in an oral glucose tolerance test
7	Blood pressure	Diastolic blood pressure (mm Hg)
8	Skin thickness	Triceps skin fold thickness (mm)
9	Insulin	2-h serum insulin (mu U/ml)
10	BMI (body mass index	Body mass index (weight in kg/(height in m)2)
11	DPF	Diabetes pedigree function
12	Serum creatinine	Test measures the level of creatinine in the blood
13	Serum sodium	sodium content is in your blood
14	Serum potassium	Potassium content in blood
15	HBAIC	Hemoglobin A1c, a blood pigment that carries oxygen

The proposed work is implemented using python machine learning library called Scikit-learn and the experimented results are evaluated using the evaluation metrics called Accuracy, Sensitivity, Specificity, F1 Measure, Recall, Precision, Matthews Correlation Coefficient (MCC) and AUC. The result of our proposed work is compared with existing algorithms such as Naive Bayes, Support Vector Machine(SVM), C4.5, Branch and Bound, K-Nearest Neighbor, Simulated Annealing and Randomized Hill Climb. The selected attributes of each algorithm is listed in Table 2.

Table 2: Selected Attributes of various algorithm

S.No	**Algorithm**	**Seleted Attribute**
1	NB	pregnant count,blood glucose level, insulin,BMI,DPF
2	SVM	pregnant count,blood glucose level, skin thickness, insulin,BMI,DPF, Age
3	C4.5	DPF,pregnant count
4	BB	DPF,pregnant count
5	k-NN	DPF,pregnant count
6	SN	DPF,pregnant count
7	proposed L2RFE-RF	DPF,blood glucose level

4.2 Evaluating Criteria:

Basics for all evaluating methods are True Positive, False Negative, True Negative and False Positive. The evaluating metrics calculations are shown in Eqn 6 to Eqn 13.

$$SN = \frac{TP}{TP+FN} \tag{6}$$

$$SP = \frac{TN}{TN+FP} \tag{7}$$

$$ACC = \frac{TN+TP}{TN+FN+TP+FP} \tag{8}$$

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F - Measure = \frac{2*Precison*Recall}{Precision+Recall} \quad (12)$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (13)$$

4.3 Experimental Results:

The proposed algorithm evaluated results are shown in Table 3 in terms of the selected attributes. The proposed algorithm selects 2 attributes such as DPF and blood glucose level. This is compared with the same proposed with one attribute called blood glucose level. The pictorial representation is shown in Fig 5.

Table 3: Proposed L2RFE-RF Evaluated Result based Selected Attributes

Algorithm	Sensitivity	Specificity	ACC	MCC	Precison	Recall	F-Measure	AUC
DPF+Blood glucose	0.98	0.97	0.99	0.86	0.93	0.91	0.9234	0.897
Blood Glucose	0.87	0.89	0.87	0.78	0.83	0.76	0.79	0.685

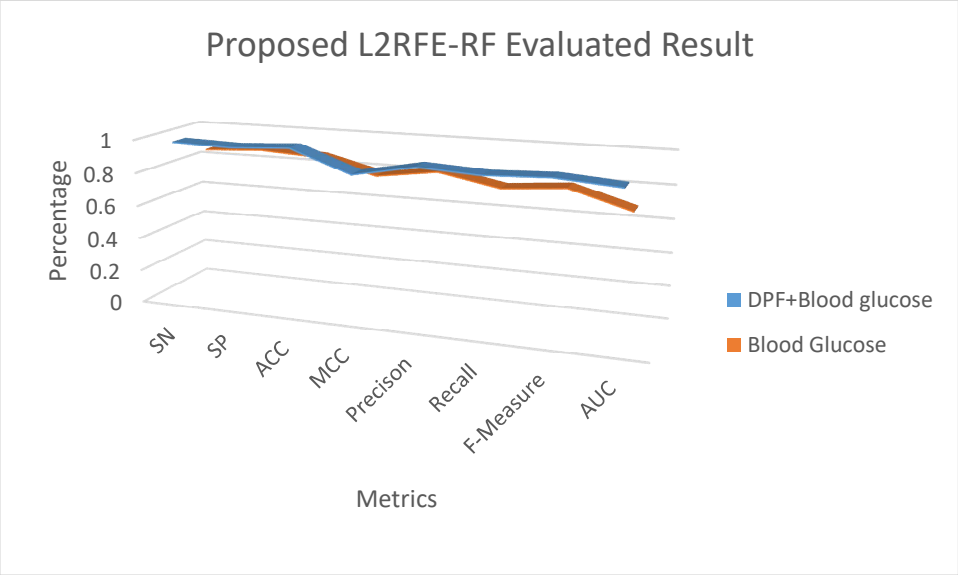

Fig 5: L2 RFE-RF RESULT

The evaluated result obtain high accuracy on all kind in the proposed work with two attributes compare to the proposed with one attribute. Even though the attribute count reduced one attribute selection does not provide correct prediction compared to two attribute selection.

The depicted decision tree based on the features of PIDD shown in fig 6. The root node is blood glucose level. Based on this next sub trees are generated with the respective root node. The next sub tree root node is insulin level and the class value of this is 0 means negative. The same way each sub tree is evaluated and the majority of the class label of the sub trees give the overall class prediction of the data set. For this sample 50% of the samples get positive and 50% get negative result.

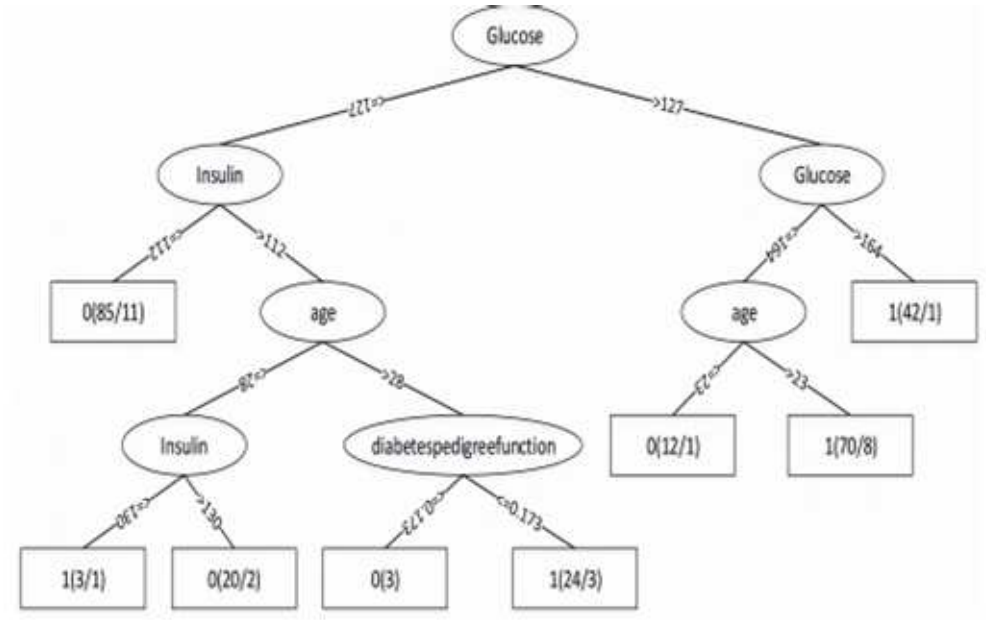

Fig 6: Features with Class attribute of PIDD using L2-RFE-RF

The comparative result with existing algorithms is shown in Table 4. Based on the selected attributes of each algorithm the results are compared using the metrics. The resultant pictorial representation is shown in Fig 7. In all the cases, our proposed algorithm obtains Sensitivity as 100%, Specificity as 97%, Accuracy as 100%, MCC as 86%, Precision as 100%,Recall as 91%, F Measure as 92% and AUC as 0.97. The next best method after our proposed is Branch and Bound with the accuracy of 96%.

Table 4: Comparative Study of various algorithms on DIDD

Approaches	**Sensitivity**	**Specificity**	**ACC**	**MCC**	**Precision**	**Recall**	**F-Measure**	**AUC**
NB	0.67	0.74	0.78	0.29	0.73	0.76	0.34	0.674
SVM	0.652	0.418	0.81	0.234	0.652	0.654	0.652	0.443
C4.5	0.652	0.745	0.95	0.564	0.698	0.652	0.634	0.617
BB	0.7343	0.875	0.963	0.839	0.874	0.846	0.753	0.505
k-NN	0.603	0.4567	0.67	0.017	0.555	0.675	0.632	0.522
SN	0.615	0.532	0.732	0.018	0.555	0.465	0.559	0.539
proposed L2RFE-RF	1	0.97	1	0.86	1	0.91	0.9234	0.97

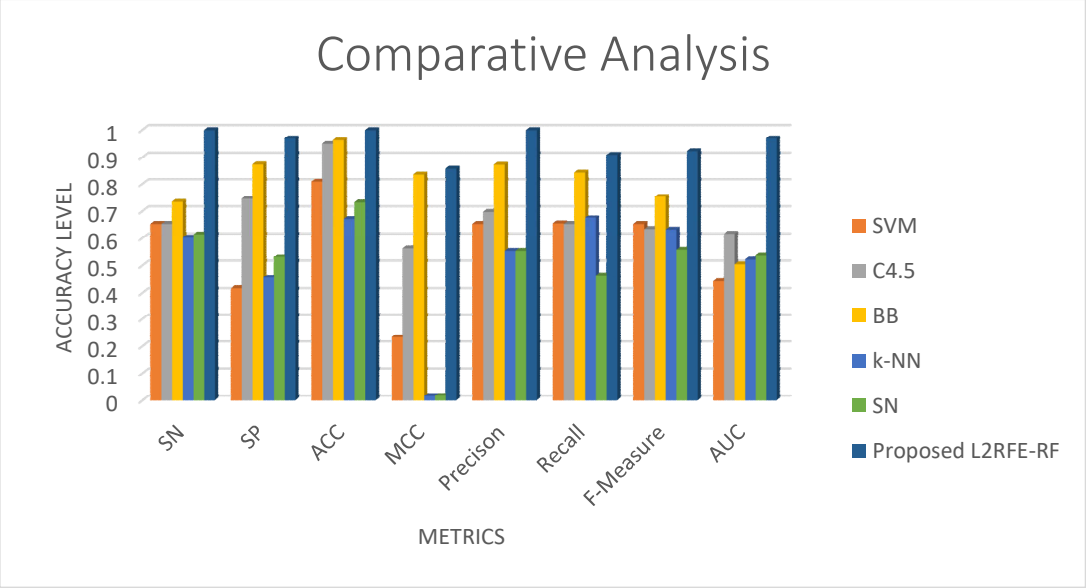

Figure 7: Comparative analysis of various algorithm with respect to Proposed L2RFE-RF

The ROC curve of the proposed algorithm with two and one attributes and the two best accuracy method obtained from Figure 8 are compared. This shows the high level accuracy with two attributes of our proposed algorithm.

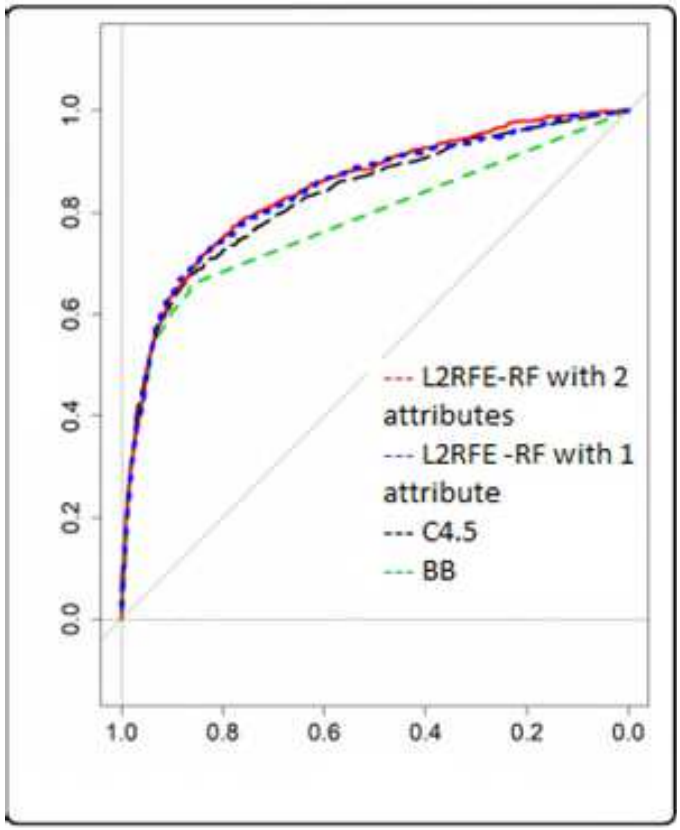

Fig 8: ROC curve of proposed algorithm

From the experimental results and discussion on our proposed algorithm called hybrid L2RFE with Random Forest is one of the best machine learning method to predict diabetes disease. By finding suitable attribute and classifiers are the important part in prediction. The proposed algorithm give better prediction since it select two relevant attributes such as DPF and glucose level using the proposed feature selection algorithm called ridge regression (L2) based Recursive Feature Elimination. The selected suitable features are then classified using Random forest classifier for classification. Hence, our proposed algorithm utilize best suitable features with best classification algorithm on PIDD data set to predict the diabetes disease.

CONCLUSION AND FUTURE WORK:

Feature selection is very challenging research area recently focused in big data, data mining etc. it is concluded that our proposed L2-RFE model produce high accuracy than comparing with existing models like SVM , KNN etc. L1 regularization does not have analytical solution where as L2 processes have analytical calculation. Recursive feature elimination helps to eliminate the worst unfit data from the feature data. It loops still it find the best solution and feature selection. Output feature is further classified by random forest classifier algorithm for getting best accuracy in feature selection. In future different machine learning algorithms can be implemented with L2 for accurate feature selection process.

Ethics declarations

Conflict of interest
All authors of the paper declare that they have no conflict of interest.

Acknowledgement This study was not funded by any grant.

Human and animal rights
This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent
Informed consent was not required as no humans or animals were involved.

Author's Contribution

K Venkatachalam: Conceptualization, Methodology, Formal analysis, Supervision, Writing - original draft, Writing - review & editing.
Prabhu P: Project administration, Investigation, Writing - review & editing.
Saravana Balaji B: Software, Visualization, Writing - original draft.
Mohamed Abouhawwash: Investigation, Data Curation, Validation, Resources, Writing - review & editing.
R Rajadevi: Validation, Writing - Review & Editing

REFERENCES

1. R. Lomte, S. Dagale, S. Bhosale and S. Ghodake, "Survey of Different Feature Selection Algorithms for Diabetes Mellitus Prediction," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697452.

2. SavinaColaco, Sujit Kumar, Amrita Tamang and Vinai George Biju, "A Review on Feature Selection Algorithms", Advances in Intelligent Systems and Computing, Emerging Research in Computing, Information, Communication and Applications ERCICA 2018, Volume 2 .p:133-153
3. B. H. Shekar and GueshDagnew, "L1-Regulated Feature Selection in Microarray Cancer Data and Classification Using Random Forest Tree", Advances in Intelligent Systems and Computing, Emerging Research in Computing, Information, Communication and Applications ERCICA 2018, Volume 2 .p:65-87
4. Lai, H., Huang, H., Keshavjee, K. *et al.* Predictive models for diabetes mellitus using machine learning techniques. *BMC EndocrDisord* 19, 101 (2019). https://doi.org/10.1186/s12902-019-0436-6DOI: https://doi.org/10.1186/s12902-019-0436-6
5. Nand Sharma, PrathameshVerlekar, Rehab Ashary, Sui Zhiquan, "Regularization and feature selection for large dimensional data", Machine Learning (cs.LG); Numerical Analysis (math.NA); Optimization and Control , arXiv:1712.01975, 2019
6. Olivier Chapelle and SathiyaKeerthi. Multi-class feature selection with support vector machines, 2008
7. QuanZou*, KaiyangQu, YameiLuo, Dehui Yin, Ying Ju and Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Front. Genet., 06 November 2018 | https://doi.org/10.3389/fgene.2018.00515
8. Herbert Pang, Stephen L. George, Ken Hui, Tiejun Tong, "Gene selection using iterative feature elimination random forests for survival outcomes",IEEE/ACM Trans ComputBiolBioinform. 2012 Sep-Oct; 9(5): 1422–1431. doi: 10.1109/TCBB.2012.63
9. J. Vijayashree, J. Jayashree, "AN EXPERT SYSTEM FOR THE DIAGNOSIS OF DIABETIC PATIENTS USING DEEP NEURAL NETWORKS AND RECURSIVE FEATURE ELIMINATION", International Journal of Civil Engineering and Technology (IJCIET) Volume 8, Issue 12, December 2017, pp. 633–641, Article ID: IJCIET_08_12_069
10. Khyati K. Gandhi, 2. Prof. Nilesh B. Prajapat, "Study of Diabetes Prediction using Feature Selection and Classification", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3 Issue 2, February – 2014

11. Akyol, K. and BahaŞen. “Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms.” *International Journal of Modern Education and Computer Science* 10 (2018): 10-16.
12. JiaqiHou,Yongsheng Sang,Yuping Liu,Li Lu, “Feature Selection and Prediction Model for Type 2 Diabetes in the Chinese Population with Machine Learning”, CSAE 2020: Proceedings of the 4th International Conference on Computer Science and Application EngineeringOctober 2020 Article No.: 103 Pages 1–7https://doi.org/10.1145/3424978.3425085
13. Madam Chakradar1 Alok Aggarwal2, “A Machine Learning Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR”, International Journal of Emerging Trends in Engineering Research , volume 8, no 5,may 2020.Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter95852020.pdf https://doi.org/10.30534/ijeter/2020/95852020
14. Atul Kumar, D.JeyaSundara, Sharmila, Sachidanand Singh, “SVMRFE based approach for prediction of most discriminatory gene target for type II diabetes”, Elsevier,Genomics DataVolume 12, June 2017, Pages 28-37
15. 15. Ke Yan, David Zhang, “Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination”, June 2015 Sensors and Actuators B Chemical 212. DOI: 10.1016/j.snb.2015.02.025

16. J. Carner, A. Mestres, E. Alarcón and A. Cabellos, "Machine learningbased network modeling: An artificial neural network model vs a theoretical inspired model," 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), Milan, 2017, pp. 522-524.
17. H. Zhang, L. Zhang and Y. Jiang, "Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems," 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 2019, pp. 1-6.
18. J. Zaech, D. Dai, M. Hahner and L. V. Gool, "Texture Underfitting for Domain Adaptation," 2019 IEEE Intelligent T ransportation Systems Conference (IT SC), Auckland, New Zealand, 2019, pp. 547 -552.
19. I. Bilbao and J. Bilbao, "Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks," 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, 2017, pp. 173-177.
20. M. Molinier and J. Kilpi, "Avoiding Overfitting When Applying Spectral-Spatial Deep Learning Methods on Hyperspectral Images with Limited Labels," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 5049 -5052.