ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter95852020.pdf
https://doi.org/10.30534/ijeter/2020/95852020

A Machine Learning Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR

Madam Chakradar[1] Alok Aggarwal[2]
[1,2] School Of Computer Science, University Of Petroleum & Energy Studies, Dehradun, India.
mchakradar@ddn.upes.ac.in[1], alok.aggarwal@ddn.upes.ac.in[2]

ABSTRACT

Type-2 diabetes mellitus (T2DM) is a significant concern since it is anticipated to reach over 693 million individuals by 2045. Identification and quantification of insulin resistance requires a particular blood test which is complicated, time-consuming and most importantly invasive, making it not feasible for routine day-to-day activities of a human being. With the development of recent machine learning approaches, identification of insulin resistance could be performed without clinical procedures. In this work insulin resistance is identified based on machine learning approaches using non-invasive techniques. Eighteen parameters are used for identification of insulin resistance; such as age, gender, waist size, height, etc. and a combination of these parameters. Experiments are conducted on the CALERIE dataset. Each output of the feature selection method is modelled over different calculations such as logistic regression, CARTs, SVM, LDA, KNN, etc. The proposed approach is verified using Stratified cross-validation test. Results show that using Logistic regression, SVM and few other versions for identification of insulin resistance, accuracy up to 97% has been achieved with standard deviation of 1% compared to 66% with Bernardini et al. [5] & Stawiski et al. [11], 61% Zheng et al. [35] and 83% Farran et al. [32]. Major benefit of the proposed approach is that a person may forecast the insulin resistance and thus future odds of diabetes may be monitored on a daily basis using non-clinical approaches. While the same is not practically possible with clinical procedures.

Key words: Insulin resistance, machine learning, type 2 diabetes mellitus, non-invasive parameters, Homa-IR

1. INTRODUCTION

Insulin resistance weakens the rate of glucose disposal from the body that increases insulin production leading to hyperinsulinemia. This may also result in metabolic abnormalities leading to hyperglycaemia, hypertension, and cardiovascular disorders etc. [38-39]. An overriding result of insulin resistance for approximately 10-15 years is T2DM. The incidence of T2DM has risen dramatically across the world to 8.5% of the populace in 2014 and in 2016 diabetes was the root cause of 1.6 million deaths, incurring enormous human, economical and societal expenses. The worldwide diabetes incidence in 2020 is now anticipated to be 9.4% that is assumed to reach 10.2% by 2030 and further extrapolated to attain 10.9% by 2045 [40]. Insulin resistance is usually recognized as a condition due to excessive body fat and hereditary causes.

Identification of insulin resistance is a complex task since it needs a complex installation, painful clinical procedure for an individual. Surrogate measuring techniques for insulin resistance are amongst other methods are also helpful. Recent advancements demonstrate Homa-IR degree is a major parameter to rely on since the lifespan of the element from the blood is more compared to other aspects. Nevertheless, these methods require clinically procedures to the individual.

Insulin resistance is escalated amounts of insulin in the blood leading to weight gain because of feeble insulin receptors. This cycle of weight reduction proceeds until the insulin receptors begin responding to the sum of blood sugar and insulin in the blood. If it doesn't then the increase in blood sugar levels leads to hyperglycaemia [38]. With continuing mismatch between insulin need (increase in blood sugar levels) and insulin production, glycaemic levels increase to amounts consistent with T2DM. Even though there are many different life risk factors related to insulin resistance yet it's tough to try by oneself on a daily basis without medical supervision. As a result of the improvements in the fields of machine learning and AI where investigators aim to get rid of such clinical obstacles and prevent trauma connected to the body on account of the extraction of blood by needles.

In this work, it is analysed that what parameters can affect the occurrence of insulin resistance. Just how much impact does this include on every parameter and what calculations will satisfy the proper precision and create the technique pertinent to real time use. Insulin resistance is identified based on machine learning approaches using non-invasive techniques. Eighteen parameters are used for identification of insulin resistance; such as age, gender, waist size, height, etc. and a combination of these parameters. Experiments are conducted on the CALERIE dataset. Each output of the feature selection method is modelled over different calculations such as logistic

regression, CARTs, SVM, LDA, KNN, etc. The proposed approach is verified using Stratified cross-validation test. Results show that using Logistic regression, SVM and few other version identifications of insulin resistance, accuracy up to 97% has been achieved with standard deviation of 1%.

Rest of the paper is organized as follows, section 2 gives the brief of literature review of the works done by earlier researchers in the field of insulin resistance identification with the help of machine learning approaches. Section 3 gives the methodology covering dataset, feature selection & identification. Results are presented in section 4 with a brief discussion. Proposed approach has been compared with other works in section 5. Finally, section 6 concludes the work.

2. LITERATURE REVIEW

Machine learning approaches are broadly of two kinds for insulin resistance and hence T2DM. These are classification and regression algorithms. Deep Learning (DL) has also been utilized keeping the greater size and sophistication of information into account [21]-[25], [42]. Joint strategy of machine learning and profound learning has also been proposed by few researchers in [26]-[30], [43].

Used various classifiers such as KNN, SVM, etc. 5-fold cross-validation is put on the dataset for validation purposes. It's shown that using KNN and Random Forest precision rate of 100% was attained after pre-processing of information while a precision rate of 74 percent with J48 classifier with no pre-processing. Tafa et al. [2] utilized a better version of SVM and Naive Bayes classifiers in which the two SVM and Naive Bayes are incorporated for diabetes forecast. Eight attributes are taken from the information with 402 patients from that 80 had T2DM. The suggested strategy is contrasted with the performance of SVM and Naive Bayes providing a precision of 95.52percent and 94.52% respectively and it's revealed that the proposed integrated strategy gives a precision of 97.6%. 10-fold cross-validation is put from the dataset. Four characteristics era, diabetes pedigree function, BMI, and Plasma Glucose concentration are chosen. It's revealed that the maximum functionality was attained with all the Hoeffding Tree algorithm together with precision worth of 0.757, F-measure of 0.759 and recalls equivalent to 0.762.

Michele Bernardini et al. [5] have suggested a high-interpretable machine learning strategy, TyG-er, for predicting where the insulin resistance state is encoded. Non-conventional clinical variables such as leukocytes, uricemia, etc., were discovered and given an insight in the very best combination of clinical variables for discovering early sugar tolerance deterioration. Patients with regular to high-risk ailments were contained while T2DM patients had been excluded from this research. A similar study was done in [6-10] utilizing statistical evaluation. The research was conducted on 315 patients aged between 7.6 to 19.7 decades. It's shown that for a pediatric population now used version for estimating GDR isn't true and this suggested GDR estimation version reliant on ANN provides a optimized forecast of GDR for clinical and research purposes. Byoung et al. [12] have suggested T2DM prediction models with machine learning methods together with the EMR dataset. A total of 8454 patients that completed five decades of follow up without a history of diabetes and to therapy in a cardiovascular heart were considered for the analysis. A total of 28 factors were pulled from the EMRs. 10-fold cross-validation is put on the data collection for validation functions. It's revealed that one of many predictive models such as LR, LDA, KNN, etc. linear regression model gave the maximum forecast performance using an AUC of 0.78.

In Nearly All studies, a Single data set was used but few researchers also took two datasets for creating a prediction just like in [4]. 10-fold cross-validation is put on the data collection and it is demonstrated by utilizing a joint dataset, diabetes forecast may be more dependable using the accuracy of 72%. In Summary, the majority of insulin resistance identification methods derive from invasive approaches. Not Many methods have focused on non-invasive approaches Which are easy, quick, and inexpensive. The Truth rate, however, Isn't quite appealing. These non-invasive approaches have to be further investigated for much better precision of insulin resistance and T2DM forecast [32].

3. METHODOLOGY

CALERIE dataset includes granular information of numerous clinical and anthropometric measurements. Many of which are not required for the projected work. Various tools containing mostly Python & Scikit-learn and their libraries are utilized to identify the essential data required for the proposed work. Information is then pre-processed by eliminating all of the measurements achieved through invasive methods and eliminating a variety of measurements that are not required. After information pre-processing, all invasive parameters have been filtered and then following the attribute selection procedure, many less important attributes are cut down. The goal factor from this dataset is called a Homa-IR level element [32-34]. This is a classification issue whether the person has insulin resistance or not.

The target variable in this work is a value that explains the increased Homa-IR (explained in section 3.2). When the formula of the ratio is accomplished, these values have been scaled in a range of values between 1 and 0, known as the scaling procedure. The attribute selection step, formerly mentioned, can currently be achieved as the target variable is projected. The information is now ready with all the essential features and must be trained for design creation. By reducing the bias in the information given to instruction, a Stratified K-fold cross-validation is performed. These calculations produce model estimations over training information and their performance is assessed over the evaluation information.

3.1 Dataset

The target of this work is to understand the adaptive changes such as in a number of other animal research done previously

like mice and rats during calorie restriction in the kind of energy consumption, whereas this work concentrates on people. In the proposed work people; including both men and women; in the age group ranging from 21-50 years have been taken. The hypothesis of the study is seen that the elastic changes in most of the participants are going to lead to exactly the same. This can be observed by falling energy consumption to 75% of those population baseline ingestion. These flexible reactions involved procedures like aging and avoidance of age-related ailments like diabetes etc. Aside from the comprehension of body temperature and resting metabolic rate this work also focused on additional laboratory tests and anthropometric tests. These tests were done on weeks 1, 6, 3, 9, 12, 18, and reasoned at 24th month. Whereas this work was narrowed down to 107 people with 33 men and 74 women over 2 years. Following the target factor prediction 37 insulin resistance, positive instances were identified. Within our work for information, technology Python is used and information evaluation is completed with the aid of this Scikit-Learn library within the CALERIE dataset. Block diagram of the proposed approach is shown in figure 1.

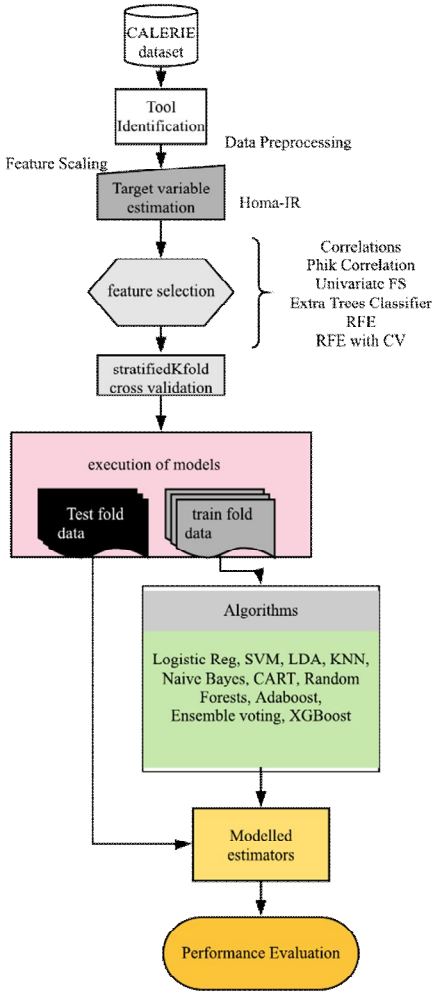

Figure 1: Block diagram of the proposed approach

3.2 Feature selection

Feature selection is amongst the principal exercises to handle almost any regression or classification problem. The information from CALERIE study is pre-processed by eliminating all of the measurements attained through invasively recorded information such as blood sample accounts, urology reports etc. called as characteristic removal. By means of this information pre-processing, a fundamental reduction of this database is accomplished. After which an additional trim down is required for which attribute selection techniques are utilized. For feature selection, target factors are devised. The target variable is chosen to function as Homa-IR test amounts since it reveals a sharp correlation between insulin resistance and T2DM [33]. This Homa-IR is the mathematical relationship between fasting blood glucose and fasting blood insulin which makes it worth for insulin resistance identification. These values have been scaled to a range of values between 1 and 0 where 0 being regular and 1 as insulin resistance. The feature selection process is completed by evaluating the above-mentioned amounts, that's the target variable. Using various techniques such as correlations, Phi K (φk) correlation, Univariate attribute choice, Extra trees classifier, RFE (recursive feature elimination), RFE using CV (cross-validation), embedded arbitrary woods, Lasso with logistic regression and lightGBM characteristic significance, goal factor is warranted contrary to all parameters. The information is now ready with all the essential features and must be trained. StratifiedKfold cross-validation is performed for reducing biases from the information provided for coaching.

3.2.1 Correlational Analysis

Correlation analysis is a statistical analysis technique utilized to analyse the strength of connections involving two numerically measured factors. This analysis is helpful to ascertain whether there are potential connections between variables/parameters. Figure 2 shows Pearson's correlational heat-map chart where intense red shows highly related, intense blue negative correlation, and white for impartial correlational mapping. Figure 3-5 shows Kendall's (τ), Spearman's (P) and Phi k (φk) correlational heat-map chart respectively. Phi k (φk) correlation is increased between 1 and 0, colored between intense red to blue. It's very small to no (negative) correlational among most of the data points from the dataset. Additionally, IR (Homa-IR goal factor) has quite few correlational data points in most charts. Normally, logistic regression analysis is opted because the output is Boolean in character and is a classification issue.

3.2.2 Extra tree classifier

While identifying the attributes correlational mapping could be integrated. Accuracy may be utilized as an instrument of reference. A decision tree provides the validity of this result based on selected attributes. This classifier is an excess tree attribute classifier where each decision tree inside this classifier is built based on the first training set. At each test node, every decision tree is provided random k features from

the whole dataset. According to this output signal best features are divided depending on the conclusion trees mathematically with Gini Index. This technique gives better comprehension as a random sample of attributes contributes to numerous de-correlated decision trees. All attributes having a score greater than 0.05 are chosen to make the model. Figure 6 shows the Extra trees classifier visualization.

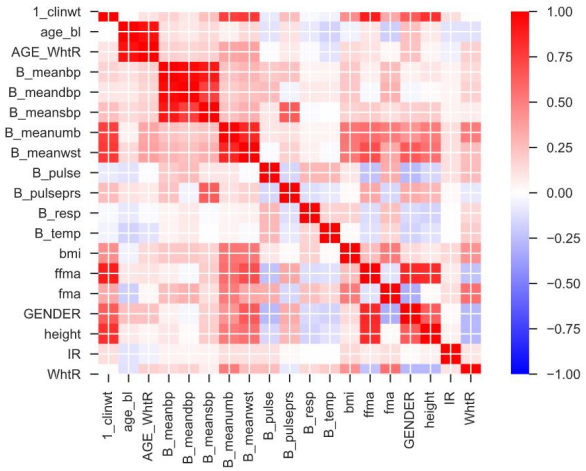

Figure 2: Pearson's correlation heat-map chart

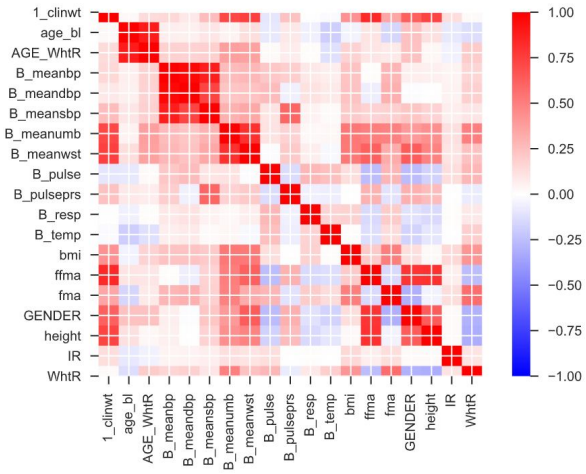

Figure 4: Spearman's (P) correlation heat-map chart

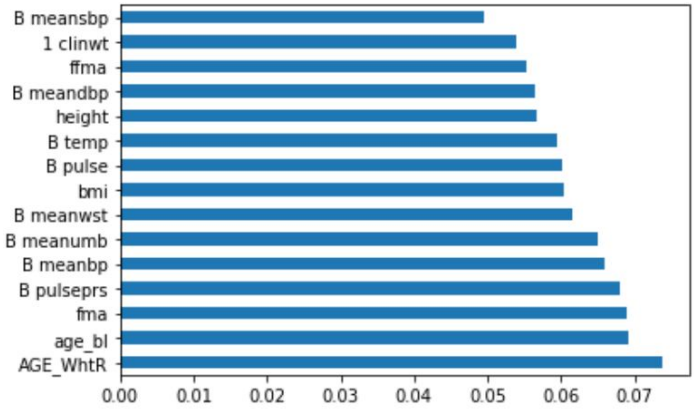

Figure 6: Extra trees classifier visualization

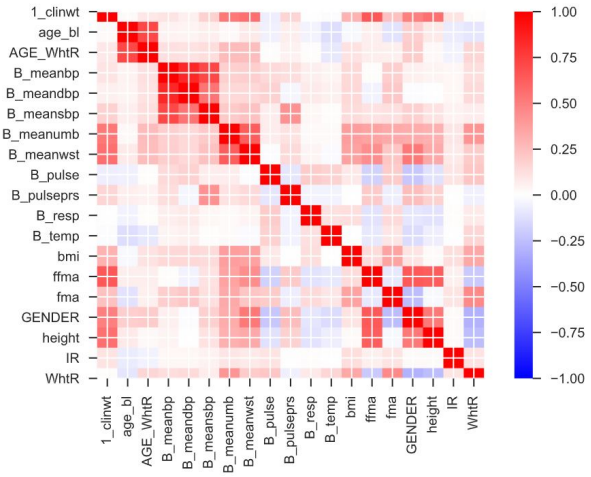

Figure 3: Kendall's (τ) correlational heat-map chart

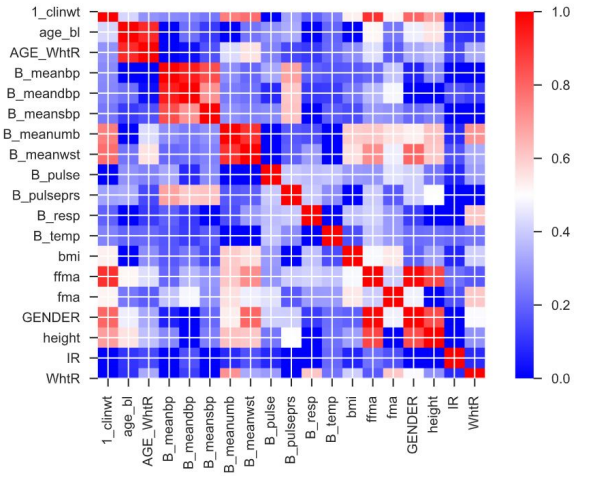

Figure 5: Phi k (φk) correlational heat-map chart

3.2.3 Univariate Feature Diagnosis

Univariate feature evaluation starts with a single attribute followed with a statistical chi-squared evaluation against the target variable/feature which shows the statistical importance of the feature over the goal factor. Then after every attribute will be inserted to execute the Chi-square test.

Select K-best: In this edition of univariate characteristic evaluation, a model is constructed by choosing k-best features within the accuracies of different models constructed over various capabilities. Figure 7 shows chi-scores with pick k-best capabilities.

Recursive Feature Elimination (RFE): This is a backward compatible manner of doing attribute elimination/selection. Originally, this technique builds a version by assessing the precision and maintain dropping one attribute at a time. Feature importance is devised by the deviation of the precision of this design when removing and adding exactly the same feature. This makes a ranking system in which the lowest position specifies the maximum value. Figure 8 shows rank based on RFE attributes.

RFE using cross-validation (RFEcv): It is a version of RFE where versions are always constructed by adding one attribute

at a time. Then an optimal choice of attributes is incorporated dependent on cross-validation of the precision of every model. Figure 9 shows the results for RFE with cross validation.

S.no	Features	Score
1	Pulse	7.302763
2	body weight	7.191259
3	fat mass	6.771772
4	age	4.440471
5	fat free mass	4.171184
6	mean waist size	3.344128
7	mean waist size(uml	3.33656
8	bmi	1.484742
9	AGE*WhtR	0.866144
10	mean systolic BP	0.801871
11	waist to height ratio	0.737427
12	mean BP	0.58869
13	height	0.490333
14	mean diastolic BP	0.482209
15	pulse pressure	0.320313
16	sex	0.075453
17	respiration rate	0.000645
18	body temperature	0.000012

Figure 7: Chi-squared scores with select k-best features

S.no	Feature	Ranking
1	age	1
2	waist to height ratio	1
3	respiration rate	1
4	AGE*WhtR	1
5	mean waist size	1
6	body temperature	1
7	bmi	1
8	sex	1
9	height	1
10	mean diastolic BP	2
11	mean waist size(umb)	3
12	fat mass	4
13	fat free mass	5
14	body weight	6
15	pulse	7
16	AGE/BMI	8
17	mean BP	9
18	pulse pressure	10
19	mean systolic BP	11

Figure 8: Ranking based on RFE features

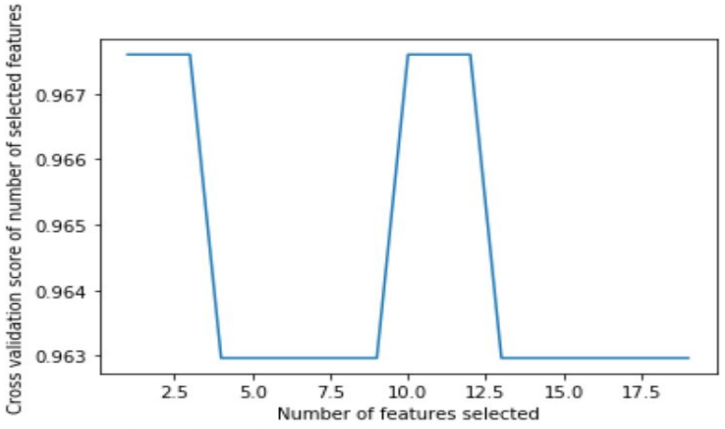

Figure 9: Results for RFE with cross validation

3.3 Feature Identification

CALERIE dataset is a mixture of all potential clinical and physical laboratory benefits. For identifying the significance of each attribute against the goal factor, as mentioned in section 3.2, various techniques are used. These techniques were segmented into eight classes as attributes 1 to attributes 8 with each class having its specific technique. These techniques based in their approaches lead to various features regarding both attributes in addition to quantity. Characteristics based on attribute selection techniques are displayed in Table 1.

Table 1: Features based on feature selection techniques

S. No.	Features category	Feature Selection type	Features
1.	Features 1	i) Pearson correlation (fig. 2) ii) Spearman's (P) correlation (fig. 3) iii) Kendall's (τ) correlation (fig. 4) iv) Phi k (φk) correlation (fig. 5)	All parameters are taken
2.	Features 2	Chi squared Select K-best (fig. 7)	All parameters are taken
3.	Features 3	Extra Trees Classifier (fig. 6)	Fat mass, mean waist size, pulse pressure, pulse, mean waistsize (umb), body weight, height, mean bp, mean diastolic bp, WhtR, age, AGE_WhtR, sex, body temperature.
4.	Features 4	Recursive Feature Elimination (RFE) (fig. 8)	age, fat free mass, bmi, body weight, sex, mean waistsize (umb), Body

			temperature, respiration rate, mean systolic bp, WhtR, AGE_WhtR, height, fat mass, mean waistsize, pulse, mean diastolic bp, mean bp, pulse pressure.
5.	Features 5	RFE with cross validation (fig. 9)	Bmi
6.	Features 6	Embedded Random forests	age, fat free mass, bmi, mean waistsize (umb), AGE_WhtR, height, fat mass, mean waist size, pulse, mean diastolic bp, mean bp.
7.	Features 7	Lasso with Logistic Regression	age, fat free mass, bmi, body weight, mean waistsize (umb), WhtR, AGE_WhtR, fat mass, mean waist size, pulse.
8.	Features 8	Light GBM classifier	All parameters are taken

3.4 Stratified K-fold Cross Validation

For the desired accuracy and variance in design implementation suggested model has to be validated. The validation procedure quantifies hypothesized relationships between factors. In the present situation evaluation of operation for several machine learning models, an evaluation has to be carried out on hidden data where under-fitting/over-fitting/well-generalized condition of this suggested model could be examined. To confirm the effectiveness of almost any machine learning or any mathematical version, cross-validation is employed as a much better instrument. It is likewise a resampling process used to assess a version in the event of restricted data.

If the version of problem is a classifier and goal factor is binary/multiclass then 'Stratified K-fold method is a better choice. Stratified K-fold cross validation is used in this situation since this function is a binary classification of the target variable that is predicated on the Homa-IR levels. This ratio estimates if the man or woman is afflicted by insulin resistance.

4. RESULTS AND DISCUSSION

Results of features 1 to features 8 with 10-Stratified K-Fold cross-validation (CV) are shown in tables 2a and 2b. Algorithm that suited the best for these features is SVM, Logistic regression (LR), LDA, XGBoost etc. which has an accuracy of 0.97 with a standard deviation of 0.01.

Table 2a: Results of features 1 to features 4 with 10-Stratified K-Fold Cross Validation (CV)

Sr. No.	Algorithm	Accuracy			
		Features 1	**Features 2**	**Features 3**	**Features 4**
1.	Logistic Regression	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.96 (+/- 0.01)
2.	Random Forest	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)
3.	Naïve Bayes	0.88 (+/- 0.09)	0.88 (+/- 0.09)	0.90 (+/- 0.09)	0.88 (+/- 0.09)
4.	AdaBoost	0.94 (+/- 0.01)	0.94 (+/- 0.01)	0.95 (+/- 0.02)	0.94 (+/- 0.01)
5.	SVM	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)
6.	LDA	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.96 (+/- 0.01)
7.	KNN	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.96 (+/- 0.01)
8.	CART	0.92 (+/- 0.03)	0.91 (+/- 0.05)	0.92 (+/- 0.04)	0.91 (+/- 0.05)
9.	Ensemble	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)
10.	XGBoost	0.96 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.95 (+/- 0.01)

Table 2b: Results of features 5 to features 8 with 10 Stratified K-Fold Cross Validation (CV)

Sr. No.	Algorithm	Accuracy			
		Features 5	**Features 6**	**Features 7**	**Features 8**
1.	Logistic Regression	0.97 (+/- 0.01)	0.96 (+/- 0.01)	0.97 (+/- 0.01)	0.96 (+/- 0.01)
2.	Random Forest	0.95 (+/- 0.03)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)
3.	Naïve Bayes	0.97 (+/- 0.01)	0.90 (+/- 0.09)	0.90 (+/- 0.09)	0.88 (+/- 0.09)
4.	AdaBoost	0.96 (+/- 0.01)	0.95 (+/- 0.02)	0.95 (+/- 0.02)	0.94 (+/- 0.01)
5.	SVM	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)
6.	LDA	0.97 (+/- 0.01)	0.96 (+/- 0.01)	0.97 (+/- 0.01)	0.96 (+/- 0.01)
7.	KNN	0.96 (+/- 0.01)	0.96 (+/- 0.01)	0.97 (+/- 0.01)	0.96 (+/- 0.01)
8.	CART	0.95 (+/- 0.03)	0.92 (+/- 0.04)	0.92 (+/- 0.04)	0.92 (+/- 0.03)
9.	Ensemble	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)
10.	XGBoost	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.96 (+/- 0.01)

Complete the mixtures among feature selection and model production, many calculations such as logistic regression, SVM, LDA, XGBoost etc demonstrated best results with the precision of 97% and 1 percent variance with several inputs but RFE with cross validation made it feasible with only BMI. The proposed version may classify a person with insulin resistance with advanced machine learning methods by preventing invasive procedures that require extraction of blood from the subject. Although body height and weight

measurement can be carried out by many ways because this over time could be improved through non-invasive parameters. Chances of insulin resistance could be diminished by controlling anthropometric measurements by calorie count, fitness centre regimes, etc. This analysis could help individuals to track their probability of insulin resistance that is a preventative measure for T2DM.

5. COMPARISON OF THE PAST WORK WITH THE PROPOSED APPROACH

Zheng et al. [35] estimated the rate of blood glucose diffusion which provides the total amount of time necessary for the body to absorb all of the glucose in blood. A minimal absorption rate implies high odds of insulin resistance and visa-versa. 61.6% precision was attained with a population size of 36. 44% precision was attained with MARSplines using a median error of 3.6% and 66% precision with ANN using a median error of 0.6%. Bernardini et al. [5] employed a population size of 968 for insulin resistance quote. This technique includes a focused set parameter together with the ratio of triglycerides to sugar ratio peeled from a present technique simply requires BMI. Insulin resistance changes digital health record. 66% accuracy was attained with Ensemble Random by Stawiski et al. [11] but with a much better P-value. But all these methods utilized parameters which are invasively gathered through blood sample reports, urine samples, etc. However, Farran et al. [32] suggested a solution based on Kuwait medical documents to extract non-invasive parameters such as anthropometric parameters, family history of hypertension and diabetes etc. which gave 83% precision and the information enhanced precision slightly diminished as much as 71%.

Initial 3 methods are invasive whereas the last one as non-invasive. Proposed work is modelled from non-invasive parameters that are economical, easy, rapid, and don't need family history and medical background from any player as with Farran et al. [32]. Employing various newest machine learning methods this work proposes to use SVM/LR/LDA/XGBoost approaches to achieve a precision of 97% with a standard deviation of 1%. Table 3 shows the comparison of this proposed work with other contemporary works.

Table 3: Comparison of insulin resistance (IR) identification

Authors	Population	Parameters Status	Parameters	Algorithms and Models	Results
Zheng et al. [35]	36	Invasive	HbA1c, Diastolic BP, WHR.	lnGDR = 4.964 − 0.121 × HbA1c (%) − 0.012 × diastolic blood pressure (mmHg) − 1.409 × WHR	R2=0.616, p<0.01
Stawiski et al. [11] NIRCa	315	invasive	Waist size, Triglycerides, HbA1c.	1. MARSplines 2. ANN	• R2=0.44, p<0.0001, median error=3.6% • R2=0.66, median error= 0.6%
Bernardini et al. [5] TyG-er	968	invasive	uricemia, leukocytes, gamma-glutamyltransferase and protein profile.	Ensemble Random Forests	R2=0.666, p<0.05
Farran et al. [32]	1837	Non-invasive (with history reports)	Age, BMI, family history of diabetes, hypertensive status, family history of hypertensive, sex.	1. KNN 2. LR 3. SVM (AUC scores)	A. 3 years data 1. 0.83 2. 0.74 3. 0.73 B. 5 years data 1. 0.83 2. 0.72 3. 0.68 C. 7 years data 1. 0.79 2. 0.72 3. 0.71
Proposed work	321	Non-invasive (without history)	BMI	1. Logistic Regression 2. SVM 3. XGBoost 4. LDA 5. XGBoost	• Accuracy=0.97, (+/-0.01) For all

6. CONCLUSION

A non-invasive technique is proposed to identify and monitor insulin resistance in a healthier person aged between 21-50 years. Eighteen parameters are used for identification of insulin resistance; such as age, gender, waist size, height, etc. and a combination of these parameters. Experiments are conducted on the CALERIE dataset. Proposed approach is developed on most recent machine learning approaches for attribute scaling, feature selection, attribute significance which shown BMI as the important element for this execution. State of the art machine learning algorithms such as SVM, Logistic regression (LR), Ensemble voting classifier, XGBoost etc. helped in producing and validating machine learning model and its accuracy. The proposed approach is verified using Stratified cross-validation test. Results show that using Logistic regression, SVM and few other versions for identification of insulin resistance, accuracy up to 97% has been achieved with standard deviation of 1% compared to 66% with Bernardini et al. [5] & Stawiski et al. [11], 61% Zheng et al. [35] and 83% Farran et al. [32]. With the support of such findings it may be reasoned that tracking premature type 2 diabetes or monitoring insulin resistance in healthy people with non-invasive methods is not far-fetched which is further researched for weight loss monitoring, diet regiments etc.

ACKNOWLEDGEMENT

Financial support from University of Petroleum & Energy Studies (UPES), Dehradun, India for conducting this work is gratefully acknowledged.

REFERENCES

[1] J.P Kandhasamy, S. Balamurali." **Performance Analysis of Classifier Models to Predict Diabetes Mellitus**", *Procedia Comput. Sci.*, vol. 47, pp. 45–51, 2015. https://doi.org/10.1016/j.procs.2015.03.182

[2] Z. Tafa, N. Pervetica, B. Karahoda." **An intelligent system for diabetes prediction**", *In Proceedings of the 2015; 4th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 14–18 June 2015; pp. 378–382.

[3] F. Mercaldo, V. Nardone,A. Santone. "**Diabetes Mellitus Aected Patients Classification and Diagnosis through Machine Learning Techniques**"**,** *Procedia Comput. Sci.* 2017, 112, 2519–2528. https://doi.org/10.1016/j.procs.2017.08.193

[4] A. Negi, V. Jaiswal." **A first attempt to develop a diabetes prediction method based on different global datasets**"**,** *In Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC),*Waknaghat, India, 22–24 December 2016; pp. 237–241. https://doi.org/10.1109/PDGC.2016.7913152

[5] Michele Bernardini, Micaela Morettini, Luca Romeo, Emanuele Frontoni, Laura Burattini." **TyG-er: An ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic Health Records**"**,** *Computers in Biology and Medicine,* 15 July 2019.

[6] H. Gylling, M. Hallikainen, J. Pihlajamäki, P. Simonen, J. Kuusisto, M. Laakso, T.A. Miettinen. " **Insulin sensitivity regulates cholesterol metabolism to a greater extent than obesity: lessons from the METSIM Study",** *JLR (J. Lipid Res.)* 51 (8) (2010) 2422–2427.

[7] E. Krishnan, B.J. Pandya, L. Chung, A. Hariri, O. Dabbous. " **Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: a 15-year follow-up study**"**,** *Am. J. Epidemiol.* 176 (2) (2012) 108–116. https://doi.org/10.1093/aje/kws002

[8] M.A. de Vries, A. Alipour, B. Klop, G.J.M. van de Geijn, H.W. Janssen, T.L. Njo, N. van der Meulen, A.P. Rietveld, A.H. Liem, E.M. Westerman, W.W. de Herder, M.C. Cabezas. " **Glucose-dependent leukocyte activation in patients with type 2 diabetes mellitus**", *familial combined hyperlipidemia and healthy controls,* Metabolism 64 (2) (2015) 213–217.

[9] D.J. Lee, J.S. Choi, K.M. Kim, N.S. Joo, S.H. Lee, K.N. Kim. " **Combined effect of serum gamma-glutamyltransferase and uric acid on Framingham risk score**", *Arch. Med.* Res. 45 (4) (2014) 337–342. https://doi.org/10.1016/j.arcmed.2014.04.004

[10] S. Riaz. " **Study of protein biomarkers of diabetes mellitus type 2 and therapy with vitamin B1"**, *J. Diabetes Res.* 2015 (2015) 10. Article ID: 150176. https://doi.org/10.1155/2015/150176

[11] Stawiski, K., et al. "**NIRCa: An artificial neural network-based insulin resistance calculator**"**,** *Pediatr Diabetes,* 2018. 19(2): p. 231-235.

[12] Choi, B. G. *et al*. " **Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks**"**,** *Yonsei Med. J.* 60, 191–199. https://doi.org/10.3349/ymj.2019.60.2.191

[13] N. Yuvaraj, K.R. SriPreethaa. " **Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster**"**,** *Clust. Comput,* 2017, 22, 1–9.

[14] E.O. Olaniyi, K. Adnan." **Onset diabetes diagnosis using artificial neural network**"**,** *Int. J. Sci. Eng,* Res. 2014, 5, 754–759.

[15] Z. Soltani, A. Jafarian." **A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II**"**,** *Int. J. Adv. Comput. Sci.* Appl. 2016, 7, 89–94.

[16] A. Sarwar, V. Sharma." **Comparative analysis of machine learning techniques in prognosis of type II diabetes**"**,** *AI Soc,* 2014, 29, 123–129.

https://doi.org/10.1007/s00146-013-0456-0

[17] Durairaj, M.; Kalaiselvi. G.“ **Prediction of Diabetes using Back propagation Algorithm**”, Int. J. Innov. Technol. 2015, 1, 21–25.

[18] M. Maniruzzaman, N. Kumar, M. Menhazul Abedin, M. Shaykhul Islam, H.S. Suri, A.S. El-Baz, J.S. Suri. “ **Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm**”, *Comput. Methods Programs Biomed*. 2017, 152, 23–34.

[19] R. Mirshahvalad, N.A. Zanjani.“ **Diabetes prediction using ensemble perceptron algorithm**”**,** *In Proceedings of the 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN),* Girne, Cyprus, 16–17 September 2017; pp. 190–194. https://doi.org/10.1109/CICN.2017.8319383

[20] X. Sun, X. Yu, J. Liu, H. Wang.“ **Glucose prediction for type 1 diabetes using KLMS algorithm**”**,** *In Proceedings of the 2017 36th Chinese Control Conference (CCC),* Liaoning, China, 26–28 July 2017; pp. 1124–1128.

[21] A. Ashiquzzaman, A. Kawsar Tushar, M.D. Rashedul Islam, D. Shon, L.M. Kichang, P. Jeong-Ho, L. Dong-Sun, K. Jongmyon.“ **Reduction of overfitting in diabetes prediction using deep learning neural network In IT Convergence and Security**”**,** *Lecture Notes in Electrical Engineering*; Springer: Singapore, 2017; Volume 449.

[22] G. Swapna, K.P. Soman, R. Vinayakumar.“ **Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals**”**,** *Procedia Comput. Sci.* 2018, 132, 1253–1262. https://doi.org/10.1016/j.procs.2018.05.041

[23] A. Mohebbi, T.B. Aradóttir, A.R. Johansen, H. Bengtsson, M. Fraccaro, M. Mørup.“ **A deep learning approach to adherence detection for type 2 diabetics**”**,** *In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC*), Jeju, Korea, 11–15 July 2017; pp. 2896–2899.

[24] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley. “ **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**”**,** *Sci. Rep.* 2016, 6, 26094.

[25] T. Pham, T. Tran, D. Phung, S. Venkatesh.“ **Predicting healthcare trajectories from medical records: A deep learning approach**”**,** *J. Biomed. Inform.* 2017, 69, 218–229. https://doi.org/10.1016/j.jbi.2017.04.001

[26] A. Askarzadeh, A. Rezazadeh.“ **Artificial neural network training using a new efficient optimization algorithm**”**,** *Appl. Soft Comput.* 2013, 13, 1206–1213.

[27] N.M. Rao, K. Kannan, X.Z. Gao, D.S. Roy.“ **Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution**”, *Comput. Electr. Eng.* 2018, 67, 483–496.

[28] P. Rahimloo, A. Jafarian.“ **Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them**”, *Bull. Société R. Sci. Liège 2016,* 85, 1148–1164.

[29] N.S. Gill, P. A Mittal.“ **computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease**”**,** *J. Theor. Appl. Inf. Technol.* 2016, 87, 1–10.

[30] M NirmalaDevi, S.A. Alias Balamurugan, U.V. Swathi.“ **An amalgam KNN to predict diabetes mellitus**”, *In Proceedings of the 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, India,* 25–26 March 2013; pp. 691–695. https://doi.org/10.1109/ICE-CCN.2013.6528591

[31] Prof William EKraus et al.” **2 years of calorie restriction and cardiometabolic risk (CALERIE): exploratory outcomes of a multicentre, phase 2, randomised controlled trial**”**,** *the lancet,diabetes and endocrinology,* Volume 7, Issue 9, Pages 673-683, September 2019.

[32] Farran, Bassam et al. “ **Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait**”**.** *Frontiers in endocrinology vol. 10 624*, 11 Sep. 2019, doi:10.3389/fendo.2019.00624

[33] Gutch Manish, Kumar Sukriti, Razi Syed Mohd, Gupta Kumar Keshav, Gupta Abhinav.“ **Assessment of insulin sensitivity/resistance**”, *Indian Journal of Endocrinology and metabolism* vol. 19,1 (2014): 160-164. https://doi.org/10.4103/2230-8210.146874

[34] Morimoto, A., Tatsumi, Y., Soyano, F., Miyamatsu, N., Sonoda, N., Godai, K., Ohno, Y., Noda, M., & Deura, K. (2014).“ **Increase in homeostasis model assessment of insulin resistance (HOMA-IR) had a strong impact on the development of type 2 diabetes in Japanese individuals with impaired insulin secretion: the Saku study**”**,** PloS one, 9(8), e105827. https://doi.org/10.1371/journal.pone.0105827.

[35] Xueying Zheng, Bin Huang, Sihui Luo, Daizhi Yang, Wei Bao, Jin Li, Bin Yao, Jianping Weng, Jinhua Yan.“ **A new model to estimate insulin resistance via clinical parameters in adults with type 1 diabetes**”, *Diabetes Metabolism Research and Reviews*, volume 33, issue 4, May 2017.

[36] Orison O. Woolcott, Richard N. Bergman.“ **Relative Fat Mass as an estimator of whole-body fat percentage among children and adolescents: A cross-sectional study using NHANES**”, *Scientific reports, Nature Research,* (2019) 9:15279 https://doi.org/10.1038/s41598-019-51701-z

[37] Sisodia, D.; Sisodia. D.S.“ **Prediction of Diabetes using Classification Algorithms**”, *Procedia Comput. Sci.* 2018, 132, 1578–1585. [CrossRef]

[38] Freeman AM, Pennings N. **Insulin Resistance**, *StatPearls (Internet)*, Treasure Island, 2019.

[39] Valeska Ormazabal, Soumyalekshmi Nair, Omar Elfeky, Claudio Aguayo, Carlos Salomon & Felipe A. Zuñiga.“ **Association between insulin resistance and the development of cardiovascular disease**”, *Cardiovascular Diabetology*, volume 17, Article number: 122 2018.

[40] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Dominic Bright, Rhys Williams.“ **Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition**”, *Diabetes research and clinical practice*, volume 157, 107843, november 01, 2019.

[41] **Global report on diabetes**, WHO, ISBN 978 92 4 156525 7 2016 (Accessed on 1st March 2020).

[42] Edward B. Panganiba,“ **Microcontroller-based Wearable Blood Pressure Monitoring Device with GPS and SMS Feature through Mobile App**”, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 6, pp. 32-35, 2019. https://doi.org/10.30534/ijeter/2019/02762019

[43] David Aregovich Petrosov, Roman Alexandrovich Vashchenko, Alexey Alexandrovich Stepovoi, Natalya Vladimirovna Petrosov,“ **Application of Artificial Neural Networks in Genetic Algorithm Control Problems**”, *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 1, pp. 177-181, 2020. https://doi.org/10.30534/ijeter/2020/24812020