PalArch's Journal of Archaeology of Egypt / Egyptology

FEATURE SELECTION FOR INSULIN RESISTANCE USING RANDOM FORESTS BASED APPROACH

Madam Chakradar,[1] Alok Aggarwal[2]
[1,2] School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India
chakradar10@hotmail.com,[1] alok.aggarwal@ddn.upes.ac.in[2]

Madam Chakradar,[1] Alok Aggarwal[2] -- Feature Selection for Insulin Resistance using Random Forests based Approach -- Palarch's Journal Of Archaeology Of Egypt/Egyptology 18(4), ISSN 1567-214x

Keywords Machine learning, Feature selection, Insulin resistance, Random Forests, CALERIE study.

Abstract

Type-2 diabetes mellitus (T2DM) is a significant problem considering that it is anticipated to reach over 693 million people by 2045. T2DM is the serious situation of insulin resistance, recognition, and quantification of insulin resistance calls for a specific blood examination which is made complex, lengthy, and most notably intrusive, making it not possible for regular day-to-day tasks of a human. With the advancement of current Artificial Intelligence approaches, the identification of insulin resistance might be executed without clinical procedures. In this job, insulin resistance is determined based on Machine Learning methods utilizing non-invasive strategies. Nineteen parameters are used for recognition of insulin resistance; such as age, sex, waist size, height, and so on as well as a mix of these specifications. Experiments are performed on the CALERIE dataset to determine the factors that impact insulin resistance. Each result of the function option technique is modelled with the help of a Random Forest Classifier. The suggested technique is validated by making use of a Stratified cross-validation test. Outcomes reveal that utilizing Logistic regression, Naïve Bayes, LDA, and also Random forests Classifier for recognition of insulin resistance, precision as much as 0.843 with AUC scores of 0.84 using Naïve Bayes classifier. [32] The major benefit of the suggested approach is that a person may forecast the insulin resistance and hence future probabilities of diabetic issues may be checked daily utilizing non-clinical methods. While the same is not virtually possible with clinical procedures.

Keywords Machine learning, Feature selection, Insulin resistance, Random Forests, CALERIE study.

Abbreviations CALERIE: Comprehensive Assessment of Long-Term Effects of Reducing Intake of Energy; LDA: Linear Discriminant analysis, AUC: Area Under the Curve.

Acknowledgment

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

1. Introduction

Insulin resistance disturbs the rate of sugar disposal in the body which increases insulin manufacturing resulting in hyperinsulinemia. The variety of patients suffering from T2DM has climbed dramatically across the whole globe to 8.5% of the world populace in 2014 as well as 2016 diabetes mellitus has been the source of 1.6 million fatalities, sustaining substantial human, social and economic losses. The prevalence of diabetes occurrence in 2019 is currently 9.3% worldwide which is supposed to get to 10.2% by 2030 and 10.9% by 2045 [1-2] Insulin resistance is risen levels of insulin in the blood causing weight gain because of weak insulin receptors. This cycle of fat reduction boosts before the insulin receptors start replying to the amount of blood glucose and insulin in the blood. Nonetheless, when it is not so then the rise in blood glucose degrees leads to hyperglycaemia [3] Having a recurring inequality in between insulin demand (boost in blood glucose degrees) and also insulin manufacturing, glycaemic degrees expand to levels constant with T2DM. Though there are great deals of different life risk elements associated with insulin resistance yet it is tough to examination by oneself daily without medical oversight [4-7]

Recent advancements expose Homa-IR (Homeostatic Design Assessment of) degree as an outstanding criterion to count on due to the life expectancy of this component from the bloodstream. Yet this examination needs a clinical technique. Consequently, the developments in the fields of machine learning (ML) as well as artificial intelligence (AI) where researchers intend to prevent such professional obstacles and also protect versus the pain associated with the removal of blood from the body with needles.

This work, it is explored what specifications might influence the incidence of insulin resistance. Just how much impact do these consist of on each specification and what calculations will fulfill the suitable accuracy and also make the approach significant to real-time? Insulin resistance has been determined for people making use of non-invasive techniques using machine learning procedures. Tests have been conducted on the CALERIE dataset. Recognition of the suggested method has been done with a Stratified cross-validation examination. With the recommended approach an individual can anticipate the insulin resistance, as well as hence potential possibilities of diabetic issues, which could be tracked daily utilizing non-clinical techniques. While the same is not virtually possible with medical procedures daily.

The remainder of the paper is arranged as follows; section 2 offers a brief literary works evaluation of the functions done by earlier researchers in the field of insulin resistance recognition with the help of machine learning approaches. Area 3 offers the approach covering dataset, function choice &

recognition. Outcomes are presented in section 4 with a brief conversation. Lastly, Section 5 concludes the work.

2. Literature Review

During this last one-decade, numerous forecast models utilizing ML algorithms have been proposed for insulin resistance as well as for this reason T2DM. These machine discovering strategies are mainly of 2 types; classification and regression formulas. Some of the most common algorithms are Assistance Vector Equipment (SVM), Linear Regression, Choice Trees, Artificial Neural Networks (ANN), and so on. [8-21, 44] Deep Discovering (DL) has additionally been used maintaining into factor to consider the increased dimension and intricacy of the data [22-26] Various consolidated techniques of both ML as well as DL have additionally been recommended [27-31]

Kandhasamy et al. [8] utilized numerous classifiers like KNN, SVM, and so on over the PIMA Indian dataset. 5-fold cross-validation (Curriculum Vitae) is applied to the dataset for validation objectives. It is shown that with KNN and Random Forest accuracy price of one hundred percent has been accomplished after pre-processing of data while an accuracy rate of 74% with J48 classifier without pre-processing. Tafa et al. [9] used an improved model of Naive Bayes and also SVM classifiers where both Ignorant Bayes and also SVM are incorporated for diabetes mellitus prediction. Eight attributes have been absorbed in the data with 402 individuals out of which 80 had T2DM.

Comparison of the recommended technique has been finished with the Naïve Bayes providing an accuracy of 94.52% and also SVM with an accuracy of 95.52% and it is revealed that the suggested incorporated method offers a precision of 97.6%. Mercaldo et al. [10] utilized six classifiers; JRiP, BayesNet, RandomForest, J48, Hoeffding Tree, and also Multilayer Perceptron. PIMA Indian dataset is used. 10-fold CV is applied in the dataset. Four qualities age, diabetes mellitus pedigree function, BMI, and plasma sugar concentration have been chosen. It is revealed that the highest possible performance has been accomplished with the Hoeffding Tree formula with accuracy worths of 0.757, F-measure of 0.759, and also recalls equal to 0.762.

Michele Bernardini et al. [12] have proposed an ML technique, TyG-er. Italian Federation of General Practitioners dataset is used. Non-conventional scientific factors like leukocytes, uricemia, and so on were spotted. Clients with typical to high-risk conditions were consisted of while T2DM individuals were excluded from this research study. A comparable research study was additionally performed in [32-36] using analytical evaluation. Konrad et al. [37] have suggested a machine learning-based strategy for estimating insulin resistance in kids with type 1 diabetes. The study was executed on 315 people aged between 7.6 to 19.7 years. Byoung et al. [38] have proposed T2DM prediction versions utilizing artificial intelligence methods with an EMR dataset. A total of 8454 patients who completed 5 years of follow-up without any history of diabetes and into treatment with a cardiovascular center were taken into consideration for the study. It is revealed that amongst various anticipating versions like LR, LDA, KNN, etc., the straight regression design revealed the most effective forecast performance with an AUC of 0.78.

In the majority of researches, a single information set has been made use of but few studies also took two datasets for predicting like in [11] 2 datasets Diabetes mellitus 130-US as well as Pima Indians dataset have been used as a consolidated dataset. 10-fold CV is applied to the data collection and also it is revealed that by utilizing a combined dataset, a better forecast of diabetes mellitus can be made with an AUC of 0.72.

Most insulin resistance identification methods are based upon intrusive approaches. Very couple of strategies have concentrated on non-invasive methods which are straightforward, quick, and also economical. The accuracy rate, nonetheless, is not appealing. These non-invasive methods need to be additionally explored for far better precision of insulin resistance and T2DM forecast [39]

Keeping aside the principle of non-invasively tracking the information, a device called continual glucose monitors has been offered. Which is a good device to track the sugar absorption price making use of which insulin resistance can be quantized into degrees. However, the drawback is it will certainly be connected to the client's arm for regarding 10-4 days as well as it interacts all the data to a mobile phone through Bluetooth [45] Purely with a non-invasive perspective, one more method called retinal microperimetry can also be made use of which is another growing study area presently [46] There is an additional non-invasively trackable technique that is recognized at genomic degrees from DNA testing which is called polygenic testing. Some researchers have effectively approximated the chance of individuals finding themselves with diabetes I and II in the future, as well as these scores, are polygenic threat scores [47].

3. Methodology

CALERIE dataset incorporates granular details of various scientific and anthropometric measurements. A lot of which is not required for the suggested work. Plenty of tools are made use of to determine just the crucial details which contain greatly Python & Scikit-learn and also their collections. Data from the CALERIE dataset is subsequently pre-processed by eliminating each dimension attained through intrusive approaches and also getting rid of various measurements that are not called for. After information pre-processing, all intrusive criteria are filtered and after that after the attribute selection procedure, lots of less considerable attributes are removed. This becomes a category issue whether a person has insulin resistance or not. The dataset is described as shown in the table.1 and graphically they are displayed from figures 1(a)-1(t) with their distributions. The X-axis of these graphs shown in the figures mentions their quality.

After the formula of the ratio is recognized, these values are scaled in a range of values between 0 and 1, called the scaling process. The feature choice action can now be acquired as the target aspect is projected. This is refined with all of the essential attributes and also has to be educated for layout production. By decreasing the predisposition in the data offered to education, a Stratified K-fold CV is used. These computations create variation estimates over training data along with their performance which is examined within the examination info. Figure 1 shows the block diagram of the recommended approach.

Machine learning is in a way of an analytical method but instead of addressing a problem with existing strategies based upon the information, machine learning enables us to give our tedious jobs to computer systems for feature recognition, attribute significance, mathematical modelling, visualization optimizing all intricacies in the presence of enormous datasets with statistical approach. Considering that the current version is a classification issue and also the initial and a standard design is the logistic regression model for any kind of binary category issue. To check out and also enhance performance criteria like precision, precision, recall, f1-score, etc, brand-new machines finding out formulas came into existence.

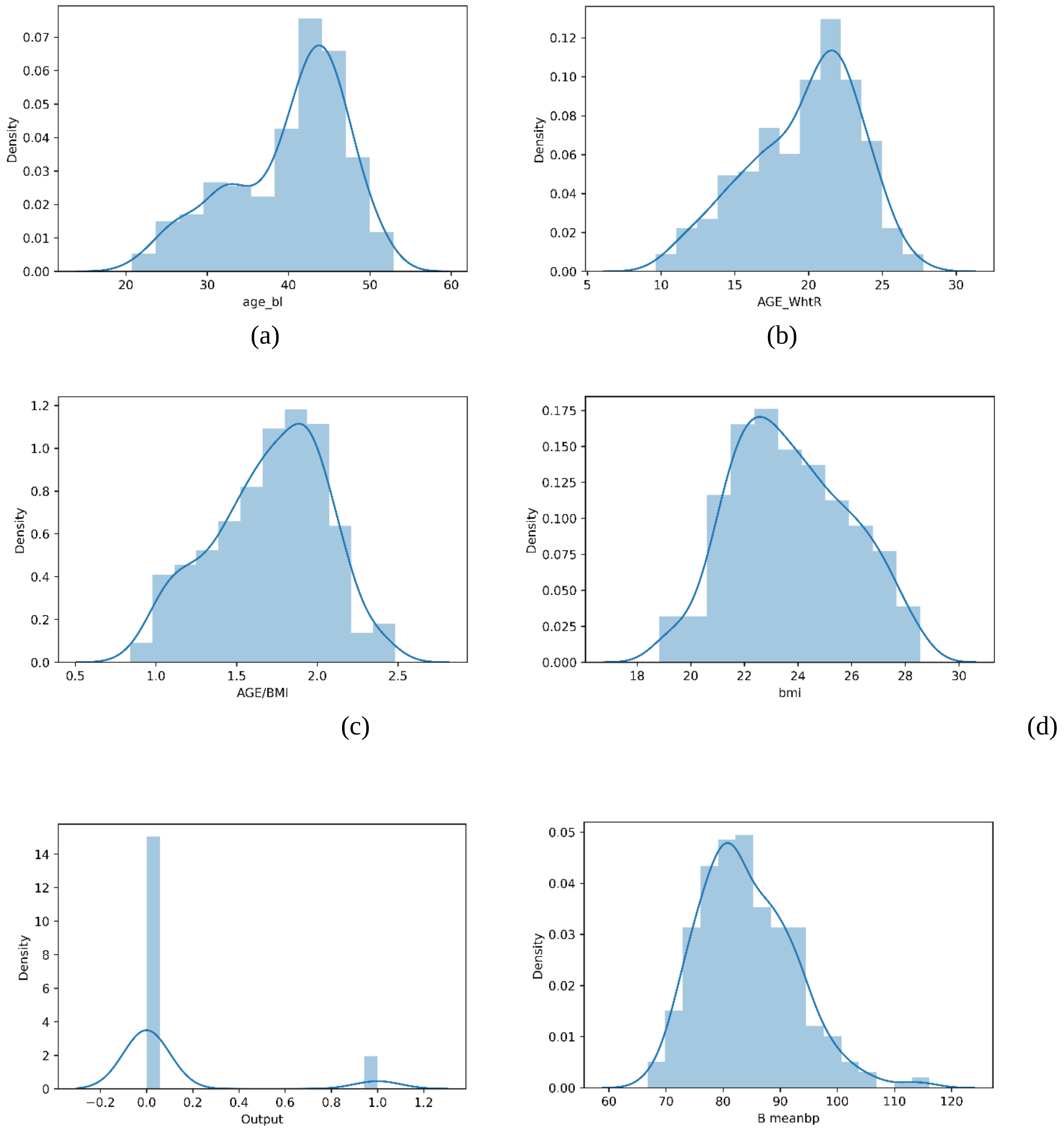
3.1 Dataset

CALERIE study is the randomized controlled trial for about 321 participants both males and females matured between 21 and also half a century with BMI between 22 as well as 27.9 kg/m2. The hypothesis of this study is discovered that the elastic modifications of most of the individuals will lead to exactly the certain same [40] This might be observed by decreasing power intake to 75% of individuals' standard intake. These elastic reactions included procedures like evasion as well as aging of age-related conditions such as diabetic issues etc. Besides the understanding of body temperature level as well as resting metabolic rate, this research study likewise focused on further lab analyses and also anthropometric evaluations. Whereas this study has been limited to 107 people with 33 guys and also 74 women over two years. Observing the target element forecast 37 insulin resistance, favorable cases were recognized as received Table 1. Inside our work for suggestions, Python language is used, and also data analysis is carried out with the assistance of the Scikit-Learn library inside the CALERIE dataset.

Table 1 Feature description.

S.no	Feature	Feature Description
1	GENDER	Gender
2	age_bl	Age
3	fma	Fat mass
4	ffma	Fat-free mass
5	1 clinwt	Clinical weight of the body
6	bmi	Body mass index
7	B meanwst	Mean waist circumference size
8	B meanumb	Mean waist circumference size around umbilical

9	B pulse	Pulse rate
10	B temp	Body temperature
11	B resp	Respiration rate
12	B meansbp	Mean systolic blood pressure
13	B meandbp	Mean diastolic blood pressure
14	B meanbp	Mean blood pressure
15	B pulseprs	Pulse pressure per second
16	WhtR	Waist to height ratio
17	AGE/BMI	age_bl / bmi
18	AGE_WhtR	Age * WhtR
19	Output	Output variable/serum C-peptide variable



(a)

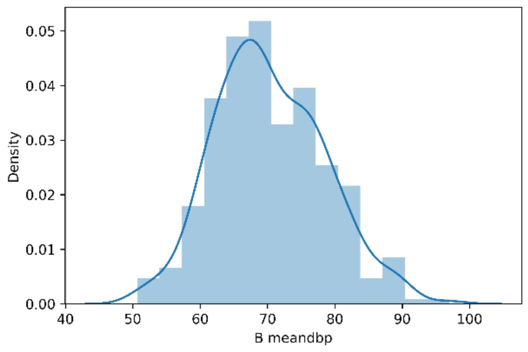
(b)

(c)

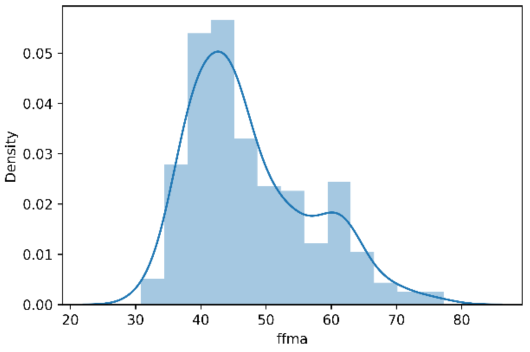
(d)

(e)

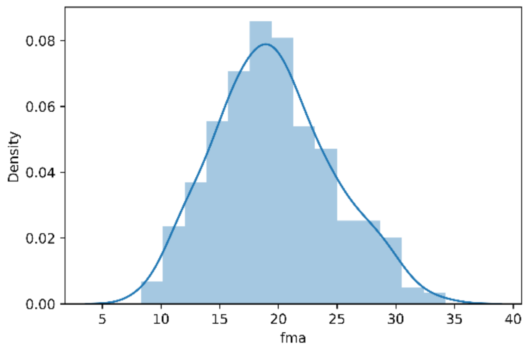
(f)


Density
0.05
0.04
0.03
0.02
0.01
0.00
40
50
60
70
80
90
100
B meandbp

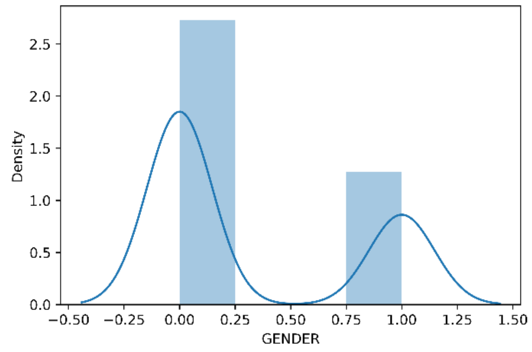
(g)


Density
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
80
ffma

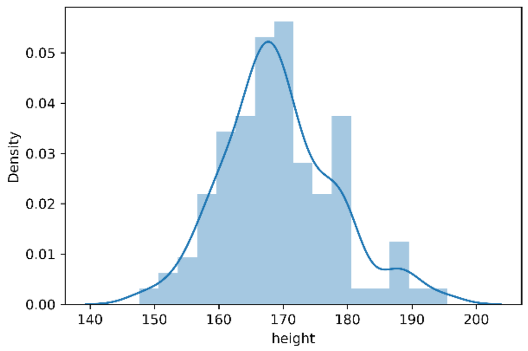
(h)


Density
0.08
0.06
0.04
0.02
0.00
5
10
15
20
25
30
35
40
fma

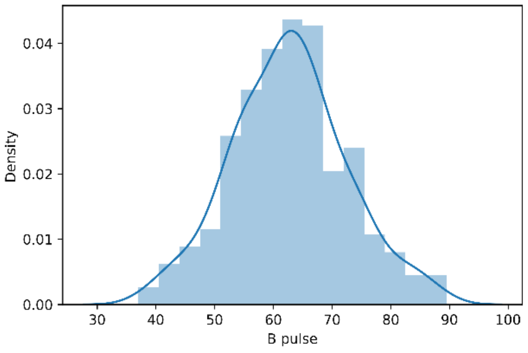
(i)


Density
2.5
2.0
1.5
1.0
0.5
0.0
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
GENDER

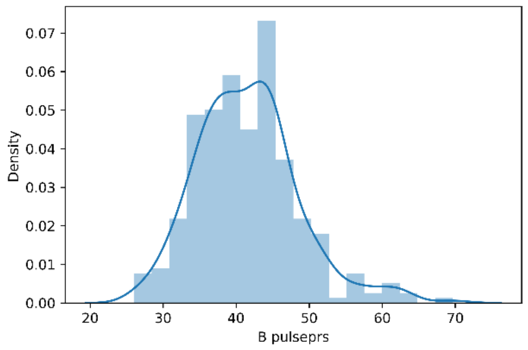
(j)


Density
0.05
0.04
0.03
0.02
0.01
0.00
140
150
160
170
180
190
200
height

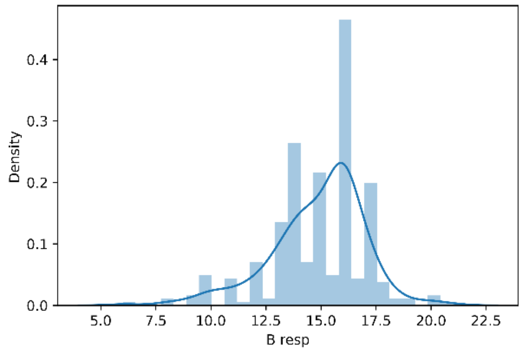
(k)


Density
0.04
0.03
0.02
0.01
0.00
30
40
50
60
70
80
90
100
B pulse

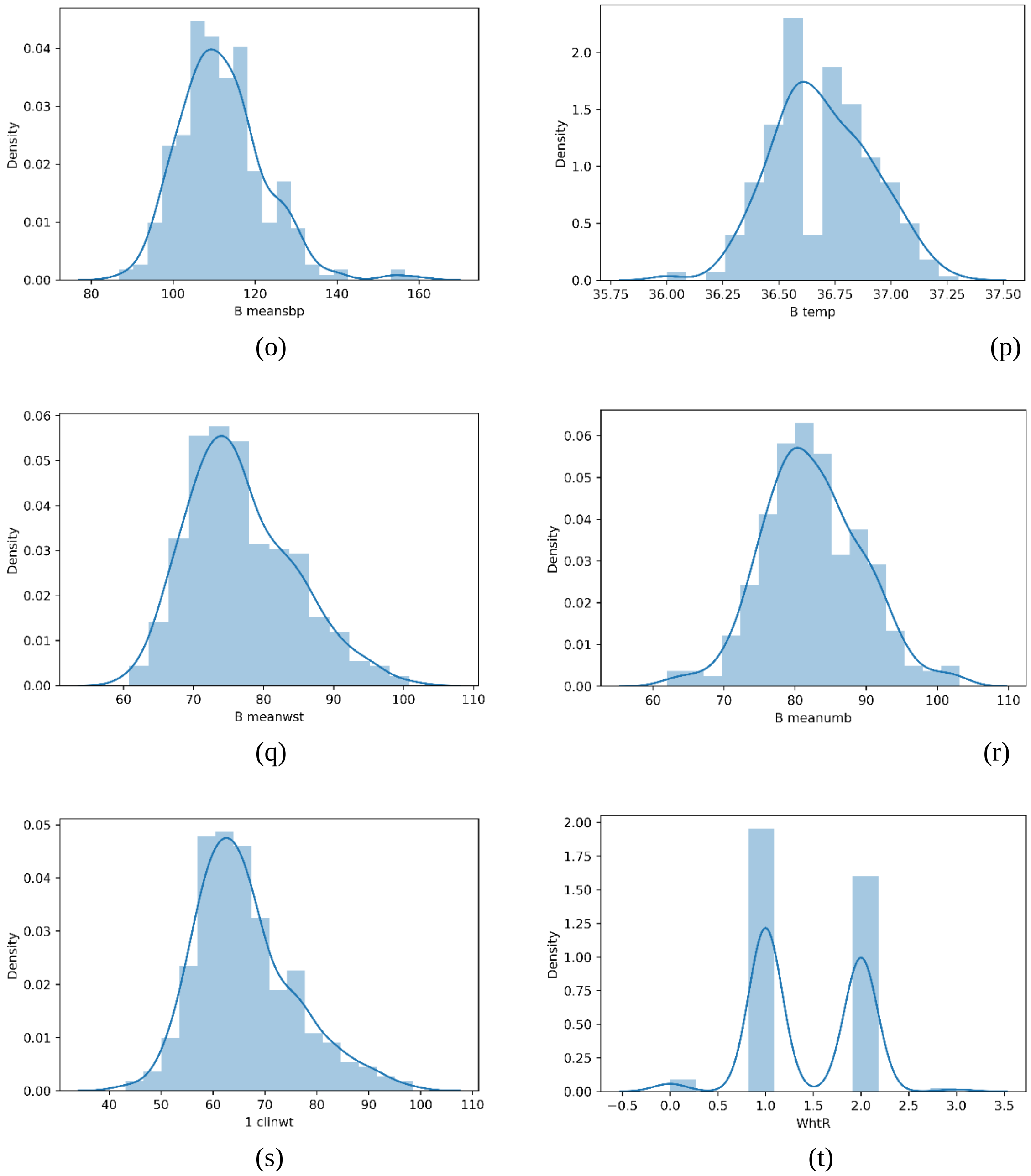
(l)


Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
B pulseprs

(m)


Density
0.4
0.3
0.2
0.1
0.0
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
B resp

(n)



(o)

(p)

(q)

(r)

(s)

(t)

Figures. 1a-1t Dataset distribution

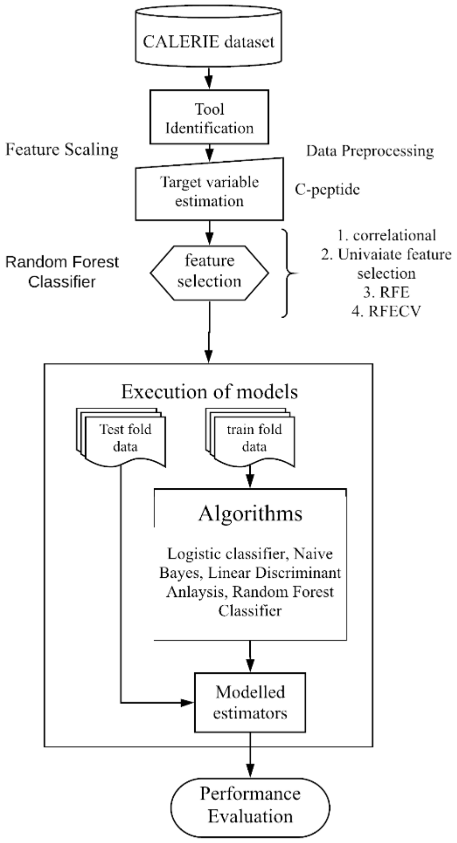


Figure.2 block diagram

3.2 Feature Selection

The feature/attribute selection is just one of the workouts to handle any type of regression or classification task. The data from the CALERIE analysis is pre-processed by eliminating the information accomplished via invasively tape-recorded information like blood sample reports, urology reports, and so on, called quality removal. In this way, a basic reduction of the database is achieved. After that, an extra reduction is required for which feature option methods are made use of. Target variables are created for function selection [41-42] These worths are scaled to a series of worths in between 0 and 1, 0 being normal as well as 1 being insulin immune. In the present paper blood serum, c-peptide levels are made use of as the metric for the target variable. Although there are various other replacements for the proportion of c-peptide which can additionally be determined like HOMA-IR, the ratio of triglycerides and HDL-c (high-density lipoprotein cholesterol) degrees from fasting blood sugar and insulin levels. The attribute option approach is currently finished by assessing the above-mentioned amounts, that is the target variable. The information is now all set with the needed attributes and needs to be trained. Stratified K fold cross-validation is performed for reducing predispositions in the information given for training. A Similar correlation heatmap is shown in figure 3 which is drawn according to Pearson's correlation technique. Generally, features that are more than 0.70 aligned with another one then they are supposed to collinear. If observed closely in figure 3 many instances will directly talk about the correlation among parameters in the heatmap. But picking the parameters with a freehand approach could end up getting in mistakes hence machine learning techniques are approached.

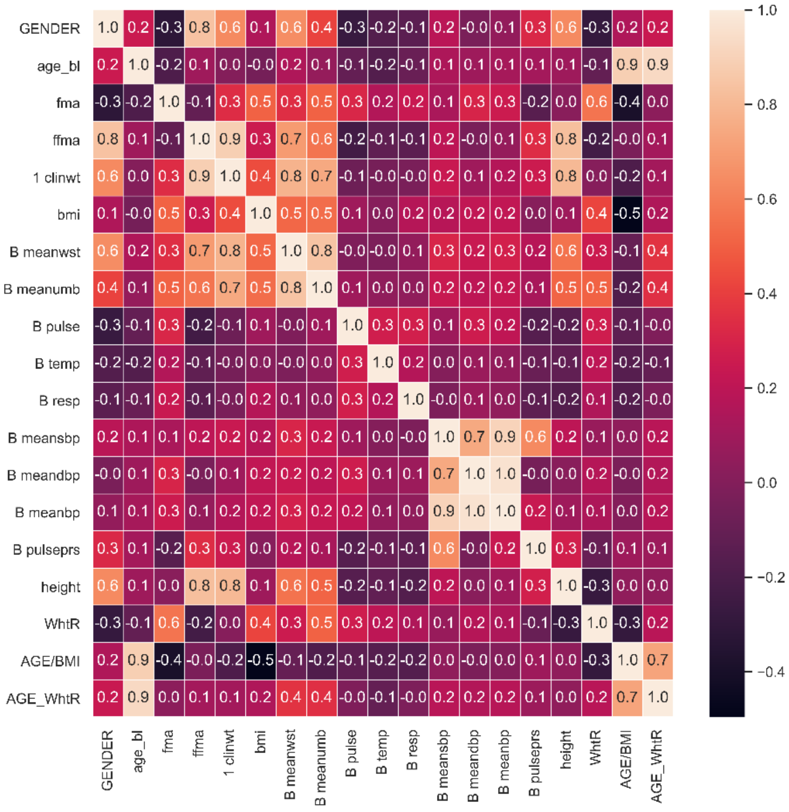


figure.3 correlation heatmap

3.2.3 Univariate Feature Selection

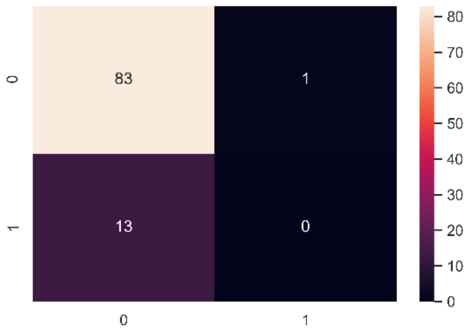
Univariate Feature examination starts with one function accompanied by an analytical Chi-squared test against the target variable/feature that reveals the analytical significance of the feature over the target component. After that after every function is included for the Chi-square test. For assigning importance to various features of the dataset Select K-best, RFE, RFEcv techniques were preferred and a random forest classifier was used to examine the performance of these techniques.

3.2.3.1 Select K-best Within this variation of univariate attribute analysis, a version is constructed by selecting k-best attributes inside the accuracies of various variations built over various abilities. figure 5 shows Chi-squared ratings with choose k-best features.

3.2.3.2 Recursive Feature Elimination (RFE) This method builds a variant by checking the precision and keep going down one attribute each time. Particular relevance is obtained from the style of precision by adding and eliminating the details same feature. This develops a ranking system where the lowest area specifies the optimum worth. figure 6 shows the standing according to RFE qualities.

3.2.3.2 Recursive Feature Elimination with Cross-validation (RFEcv) It is an in reverse compatible means of doing attribute elimination/selection. Initially, this technique develops a design by checking the precision and also keep going down one feature each time. Function relevance is developed by the deviation of the precision of the model while adding and removing the very same feature. This develops a ranking system where the most affordable ranking defines the greatest importance. Figure 7 programs ranking based upon RFE attributes.

To begin with, a random forests classifier was used to create the model by splitting the data into train and test as shown in figure 2 the block diagram explaining the methodology. So far, the CALERIE study's dataset is described in the article and the random forest approach is taken since the current problem is the classification between 0 and 1, therefore random forest generally is preferred but decision-making problems. The performance of this model reached an accuracy of 0.855 and the confusion matric for this experiment is shown in figure 4. This result is for the entire dataset of 19 presumed influential parameters. Therefore collecting 19 different parameters is a tedious task and there might be lesser or bad influential parameters in the dataset which could be downgrading the model. The process of removing such attributes from the dataset is called feature selection and the techniques that will be used are discussed in section 3.2.3.



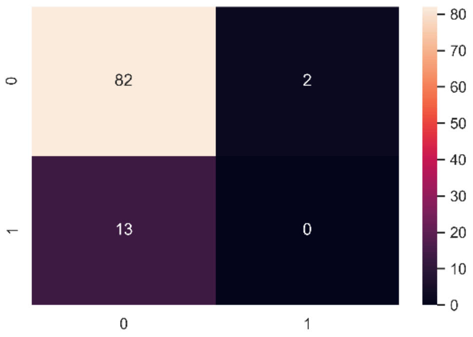


Figure. 5 confusion matrix of random forest classifier(RFC) figure.6 confusion matrix of RFC with select K-best

When the experiment is done using the RFE technique as explained in section 3.2.3.2, the technique produced few attributes based on the contribution to the model. All the features were given to them but to identify the appropriate amount of features that could achieve better results one has to cross-validate these features hence RFEcv. As observed from figure 7 a plot is drawn against the cross-validation scores of the models against all features on the x-axis. This could explain the next efficient process to identify the appropriate amount of features to achieve better performance of the model.

Based on the graph in figure 7 it will be efficient to prefer 6 features over 19 features and to identify these feature importance charts were drawn based on their feature importance scores. This feature importance score can be seen in figure 8 where the features are on the x-axis and their respective feature importance scores on the y-axis. According to figure 8, the feature attributes are fat mass, mean waist size, mean waist size over an umbilical cord, age*WhtR, body weight, and age of the body. With these parameters, machine learning techniques were run expecting better results.

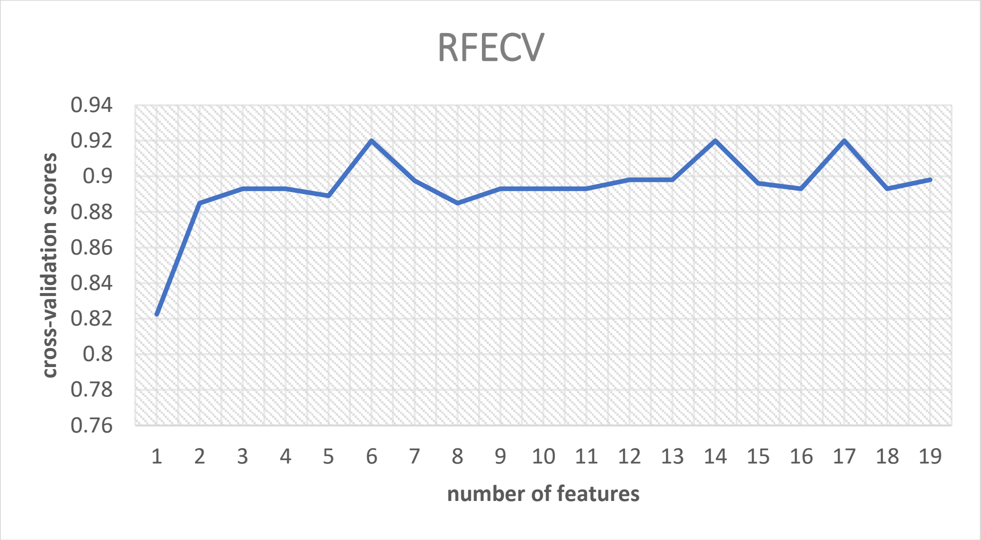


Figure.7 RFEcv number of features versus cross-validation scores

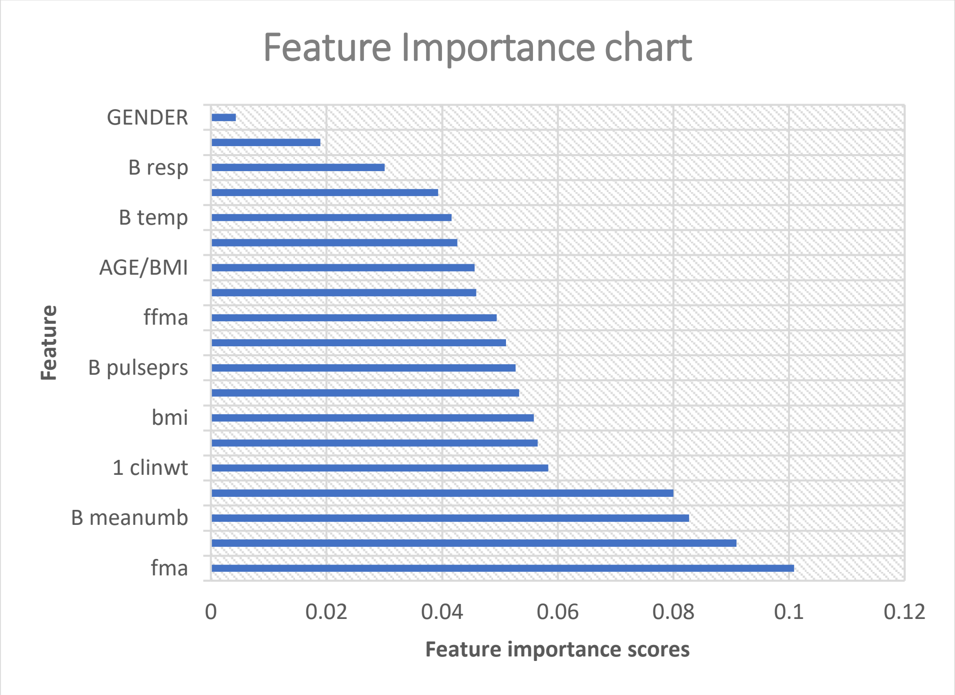


Figure.8 Feature importance chart

4. Results and Discussion

[49-50] Based on the literature and other supporting pieces of evidence few algorithms for machine learning are selected namely Logistic classifier, Naïve

Bayes classifier, Linear Discriminant Analysis, and random forest classifier. The results for these techniques can be seen in table no. 2. It explains which machine learning techniques to be used and their performance characteristics like Accuracy, Recall, precision, and F1-Score. These parameters help identify the appropriate to proceed for deployment. Based on the table Naïve Bayes classifier (NBC) produced better results and showed a significant improvement in identifying an individual with insulin resistance. This observed when compared on Figures 5 and 6 with figure 9 confusion matrix. Where NBC could successfully identify the true positive far better than previous algorithms.

Table no.2 performance characteristics of the models after feature selection

Model	Accuracy	AUC	Recall	Precision	F1-score
Logistic Classifier	0.8749	0.6636	0.0333	0.1000	0.0500
Naïve Bayes	0.8437	0.6984	0.2333	0.3500	0.2700
Linear Discriminant Analysis	0.8751	0.6685	0.0333	0.1000	0.0500
Random Forest Classifier	0.8530	0.5157	0.0333	0.0333	0.0333

Though logistic classifiers and linear Discriminant analysis secured better accuracy yet they failed to achieve better AUC and other performance characteristics which makes them less reliable models to trust for deployment. If drawn a confusion matrix for naïve Bayes classifier it outperforms previous best while performing random forest classifier without feature selection. The prediction gradient of true positive for naïve Bayes classifier grew almost 500% from 1 to 6 yet making some mistakes in decision making. The precision of 0.55 as shown in figure 10 as well as an extremely promising AUC score of 0.84 in figure 11 makes it the most viable model built for the dataset after feature selection.

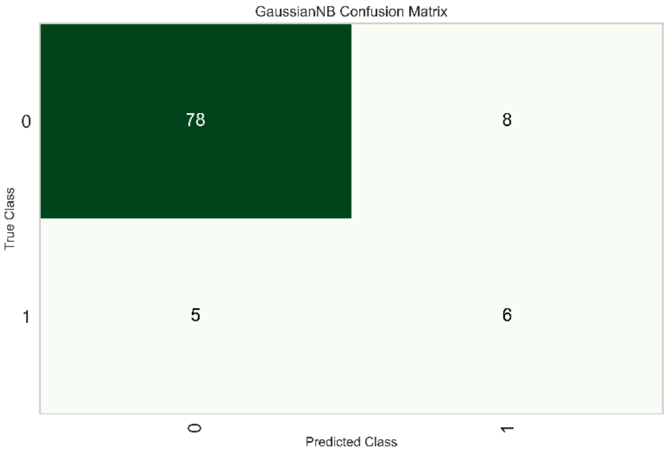


Figure.9 Confusion matrix of Naïve Bayes Classifier after feature selection

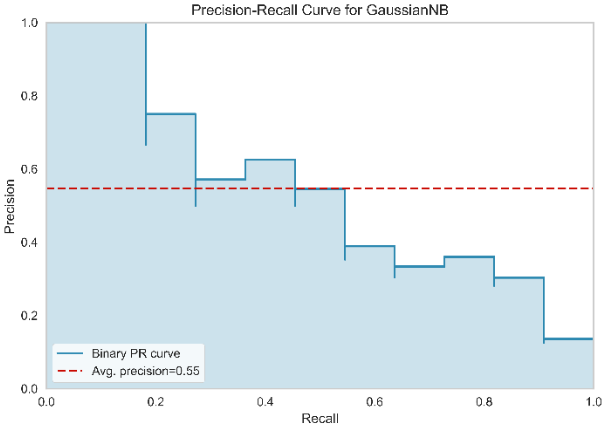


Figure.10 precision-recall curve for naïve Bayes

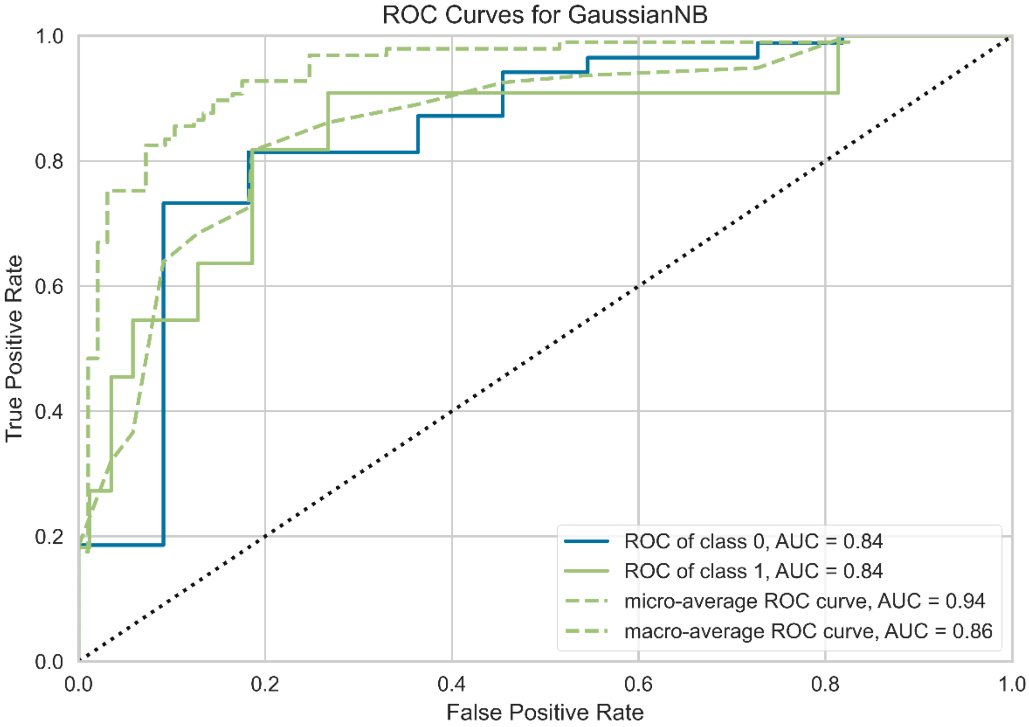


Figure.11 AUC scores

5. Conclusion

The existing strategy is anyway a brand-new research outcome but what separates it from various other existing strategies is, this needs all the inputs which can be tracked non-invasively with no state of the art of devices. In this age of common computing which motivates to recognize brand-new techniques which can be executed pervasively into day-to-day customer's life. Though CALERIE offers a great deal of data that is continuous for computational and also enhancement functions, the majority of non-invasively trackable criteria are developed into specific means including the classification variable. For that reason, the Target variable is categorized, as it can only inform whether the person has insulin resistance or otherwise, which means it can be used as an identification tool for tracking insulin resistance from age groups 21-50.

The proposed approach is created by finding out strategies for feature scaling, characteristic importance, feature selection which revealed 6 attributes/features of the dataset as one of the most important features with this proposed implementation. Machine learning techniques like Logistic classifier (LR), Naive Bayes classifier, LDA, Random Forests Classifier helped in creating as well as sustaining the trained model also it's precision. With the aid of these findings, it may be ended that checking early T2DM or tracking insulin resistance in healthy people with non-invasive methods is not improbable which is better examined for weight reduction tracking, diet quotation, etc. Though there is a lot of future scope for polygenic danger ratings for diabetes but for that procedure to meet successfully one needs to go via hereditary treatment. For that, the payments from biotechnology have discovered solutions like CRISPR-Cas9 which can potentially fix future problems [48]. The present work can be used to test and also verify the results of such gene treatment without puncturing the skin with a needle for blood.

REFERENCES

1. Valeska Ormazabal, Soumyalekshmi Nair, Omar Elfeky, Claudio Aguayo, Carlos Salomon, Felipe A. Zuñiga (2018) Association between insulin resistance and the development of cardiovascular disease, Cardiovascular Diabetology. 17:122 1-14, https://doi.org/10.1186/s12933-018-0762-4
2. Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Dominic Bright, Rhys Williams (2019) 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. 157: 107843 1-10, https://doi.org/10.1016/j.diabres.2019.107843
3. Freeman AM, Pennings N. Insulin Resistance. [Updated 2020 Jul 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan. https://www.ncbi.nlm.nih.gov/books/NBK507839/ Accessed october 5, 2020
4. Orison O. Woolcott, Richard N. Bergman (2019) 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. 9: 15279, https://doi.org/10.1038/s41598-019-51701-z
5. Global report on diabetes, WHO, ISBN 978 92 4 156525 7, 2016. https://www.who.int/diabetes/global-report/en/ (accessed March 1, 2020)
6. Xingxing Ren et al., (2016) Association between Triglyceride to HDL-C Ratio (TG/HDL-C) and Insulin Resistance in Chinese Patients with Newly Diagnosed Type 2 Diabetes Mellitus, PLOS ONE. 11, https://doi.org/10.1371/journal.pone.0154345

7. Zati Iwani Ahmad Kamil, Muhammad Yazid Jalaludin, Ruziana Mona Wan Mohd Zin, Fuziah Md. Zain (2017) Triglyceride to HDL-C Ratio is Associated with Insulin Resistance in Overweight and Obese Children, Scientific reports. 7: 40055, https://doi.org/10.1038/srep40055
8. Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Computer Science, 47, 45-51. doi:10.1016/j.procs.2015.03.182
9. Z. Tafa, N. Pervetica, B. Karahoda (2015) An intelligent system for diabetes prediction. In Proceedings of the 2015; 4th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 378–382
10. F. Mercaldo, V. Nardone, A. Santone (2017) Diabetes Mellitus Aected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Comput. Sci. 112 2519-2528
11. A. Negi, V. Jaiswal (2016) 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. 237-241
12. Michele Bernardini, Micaela Morettini,Luca Romeo, Emanuele Frontoni, Laura Burattini (2019) 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. 112: 103358
13. N. Yuvaraj, K.R. SriPreethaa (2017) Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust. Comput. 22 1-9
14. E.O. Olaniyi, K. Adnan (2014) Onset diabetes diagnosis using artificial neural network. Int. J. Sci. Eng, Res. 5 754-759
15. Z. Soltani, A. Jafarian (2016) A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II, Int. J. Adv. Comput. Sci. Appl. 7 89-94
16. Sarwar, A., & Sharma, V. (2013) Comparative analysis of machine learning techniques in prognosis of type II diabetes. Ai & Society, 29(1), 123-129. doi:10.1007/s00146-013-0456-0
17. M. Durairaj, G. Kalaiselvi (2015) Prediction of Diabetes using Back propagation Algorithm. Int. J. Innov. Technol. 1 21-25
18. M. Maniruzzaman, N. Kumar, M. Menhazul Abedin, M. Shaykhul Islam, H.S. Suri, A.S. El-Baz, J.S. Suri (2017) Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput. Methods Programs Biomed. 152 23-34
19. R. Mirshahvalad, N.A. Zanjani (2017) Diabetes prediction using ensemble perceptron algorithm. In Proceedings of the 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), Girne, Cyprus. 190-194
20. X. Sun, X. Yu, J. Liu, H. Wang (2017) Glucose prediction for type 1 diabetes using KLMS algorithm. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Liaoning, China. 1124-1128
21. D. Sisodia, D.S. Sisodia (2018) Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 132 1578-1585

22. A. Ashiquzzaman, A. Kawsar Tushar, M.D. Rashedul Islam, D. Shon, L.M. Kichang, P. Jeong-Ho, L. Dong-Sun, K. Jongmyon (2018) Reduction of overfitting in diabetes prediction using deep learning neural network. In: Kuinam J. Kim, Hyuncheol Kim, Nakhoon Baek (Eds.), IT Convergence and Security 2017,Lecture Notes in Electrical Engineering, Springer: Singapore. 449 35-43, https://doi.org/10.1007/978-981-10-6451-7_5
23. G. Swapna, K.P. Soman, R. Vinayakumar (2018) Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Comput. Sci. 132 1253-1262
24. A. Mohebbi, T.B. Aradóttir, A.R. Johansen, H. Bengtsson, M. Fraccaro, M. Mørup (2017) 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. 2896-2899
25. R. Miotto, L. Li, B.A. Kidd, J.T. Dudley (2016) Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci. Rep. 6: 26094
26. T. Pham, T. Tran, D. Phung, S. Venkatesh (2017) Predicting healthcare trajectories from medical records. A deep learning approach, J. Biomed. Inform. 69 218-229
27. A. Askarzadeh, A. Rezazadeh (2013) Artificial neural network training using a new efficient optimization algorithm. Appl. Soft Comput. 13 1206-1213
28. N.M. Rao, K. Kannan, X.Z. Gao, D.S. Roy (2018) Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution. Comput. Electr. Eng. 67 483-496
29. P. Rahimloo, A. Jafarian (2016) Prediction of Diabetes by Using Artificial Neural Network. Logistic Regression Statistical Model and Combination of Them, Bull. Société R. Sci. Liège. 85 1148-1164
30. N.S. Gill, P. A Mittal (2016) Computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. J. Theor. Appl. Inf. Technol. 87 1-10
31. M NirmalaDevi, S.A. Alias Balamurugan, U.V. Swathi (2013) 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. 691-695
32. H. Gylling, M. Hallikainen, J. Pihlajamäki, P. Simonen, J. Kuusisto, M. Laakso, T.A. Miettinen (2010) Insulin sensitivity regulates cholesterol metabolism to a greater extent than obesity. lessons from the METSIM Study, JLR (J. Lipid Res.). 51 2422-2427
33. E. Krishnan, B.J. Pandya, L. Chung, A. Hariri, O. Dabbous (2012) Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: a 15-year follow-up study. Am. J. Epidemiol. 176 108-116
34. 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 (2015) Glucose-dependent leukocyte activation in

patients with type 2 diabetes mellitus, familial combined hyperlipidemia and healthy controls, Metabolism. 64 213-217

35. D.J. Lee, J.S. Choi, K.M. Kim, N.S. Joo, S.H. Lee, K.N. Kim (2014) Combined effect of serum gamma-glutamyltransferase and uric acid on Framingham risk score. Arch. Med. Res. 45 337-342. https://doi.org/10.1016/j.arcmed.2014.04.004
36. Riaz S. (2015). Study of Protein Biomarkers of Diabetes Mellitus Type 2 and Therapy with Vitamin B1. Journal of diabetes research, 2015, 150176. https://doi.org/10.1155/2015/150176
37. Stawiski, K., Pietrzak, I., Młynarski, W., Fendler, W., & Szadkowska, A. (2018) NIRCa: An artificial neural network-based insulin resistance calculator. Pediatric diabetes, 19(2), 231–235. https://doi.org/10.1111/pedi.12551
38. Choi, B. G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., & Noh, Y. K. (2019) Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks. Yonsei medical journal, 60(2), 191–199. https://doi.org/10.3349/ymj.2019.60.2.191
39. Farran, B., AlWotayan, R., Alkandari, H., Al-Abdulrazzaq, D., Channanath, A., & Thanaraj, T. A. (2019) 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, 10, 624. https://doi.org/10.3389/fendo.2019.00624
40. William EKraus et al., (2019) 2 years of calorie restriction and cardiometabolic risk (CALERIE): exploratory outcomes of a multicentre, phase 2, randomised controlled trial. The lancet, diabetes and endocrinology. 7 673-683
41. A.G. Jones, A.T. Hattersley (2013) The clinical utility of C-peptide measurement in the care of patients with diabetes. Diabetic medicine : a journal of the British Diabetic Association. 30 803-17, https://doi.org/10.1111/dme.12159
42. K.D. Pagana, T.J. Pagana, T.N. Pagana (2019) Mosby’s Diagnostic & Laboratory Test Reference. 14th ed. St. Louis, Mo: Elsevier.
43. Xueying Zheng, Bin Huang, Sihui Luo, Daizhi Yang, Wei Bao, Jin Li, Bin Yao, Jianping Weng, Jinhua Yan (2017) A new model to estimate insulin resistance via clinical parameters in adults with type 1 diabetes. Diabetes Metabolism Research and Reviews. 33. https://doi.org/10.1002/dmrr.2880
44. Liu Y. (2020) Artificial Intelligence-Based Neural Network for the Diagnosis of Diabetes: Model Development. JMIR medical informatics, 8(5), e18682. https://doi.org/10.2196/18682
45. Rodbard D. (2017) Continuous Glucose Monitoring: A Review of Recent Studies Demonstrating Improved Glycemic Outcomes. Diabetes technology & therapeutics, 19(S3), S25–S37. https://doi.org/10.1089/dia.2017.0035
46. Andreea Ciudin, Olga Simó-Servat, Cristina Hernández, Gabriel Arcos, Susana Diego, Ángela Sanabria, Óscar Sotolongo, Isabel Hernández, Mercè Boada, Rafael Simó (2017) Retinal Microperimetry: A New Tool for Identifying Patients With Type 2 Diabetes at Risk for Developing Alzheimer Disease. Diabetes, 66 (12) 3098-3104; DOI: 10.2337/db17-0382

47. Udler, M. S., McCarthy, M. I., Florez, J. C., & Mahajan, A. (2019). Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. Endocrine reviews, 40(6), 1500–1520. https://doi.org/10.1210/er.2019-00088
48. Balboa, D., Prasad, R. B., Groop, L., & Otonkoski, T. (2019). Genome editing of human pancreatic beta cell models: problems, possibilities and outlook. Diabetologia, 62(8), 1329–1336. https://doi.org/10.1007/s00125-019-4908-z
49. Klén, R., Karhunen, M. & Elo, L.L. Likelihood contrasts: a machine learning algorithm for binary classification of longitudinal data. Sci Rep 10, 1016 (2020). https://doi.org/10.1038/s41598-020-57924-9
50. Payal Patil, Dr. Prof. Swati Shinde (2020) PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFICATION ALGORITHMS: NAÏVE BAYES, DECISION TREE AND K-STAR. Journal of Critical Reviews, 7 (19), 1160-1164. doi:10.31838/jcr.07.19.144