iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71252

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Detection Diabetes Mellitus Based on Anthropometric Parameters and Machine Learning Techniques

Pratiksha Gajbhiye[1], Dr. Muzaffar Khan[2], Dr. Ahmed Sajjad Khan[3]
[1]*Student, Electronics & Communication Engg., Anjuman College of Engineering & Technology, Nagpur, India*
[2]*Assistant Professor, Electronics & Communication Engg., Anjuman College of Engineering & Technology, Nagpur, India*
[3]*Professor, Electronics & Communication Engg., Anjuman College of Engineering & Technology, Nagpur, India*

Abstract: ***This paper explores machine learning techniques to assess the anthropometric measures most commonly linked to type 2 diabetic mellitus (T2DM). According to recent data, T2DM, which is mostly associated with visceral or abdominal obesity and metabolic abnormalities, is more common in patients with metabolic syndrome. . Identifying those who are at high risk for type 2 diabetes is vital given the disease's prevalence and serious complications. Anthropometric assessment techniques are one of the most straightforward and non-invasive ways to detect individuals at risk for diabetes, even though one-third of patients with the disease have not been identified, and this number is rising. Diabetes is predicted by the Waist-to-Height Ratio (WHtR), Body Adiposity Index (BAI), A Body Shape Index (ABSI), Body Mass Index (BMI), and Waist Circumference (WC). According to one study, the best indicators are WC and BMI. Few researches have compared the values of these anthropometric indices and their relationship with the prevalence of diabetes, despite the fact that there have been several studies on the topic. Additionally, visceral and subcutaneous fat cannot be detected using conventional anthropometric techniques. We chose to design a cohort study to investigate additional and novel anthropometrical measures for assessing their association with diabetes, given the rising prevalence of the disease, the dearth of accurate measurements for its diagnosis, and the controversy surrounding the findings of earlier studies. In order to identify patients at risk of acquiring diseases, we also employed machine learning techniques, which are revolutionary and efficient ways to organize a huge number of indicators while creating powerful predictive models.***

Keywords: Machine learning, anthropometric parameters, diabetes mellitus, Random Forest, Support Vector Machines (SVM), Logistic Regression.

I. INTRODUCTION

Diabetes mellitus is a chronic metabolic disease affecting the human body by converting blood sugar into energy. If left undiagnosed or untreated, it leads to severe complications such as cardiovascular disease, kidney failure, and neuropathy. Diabetes mellitus is a group of metabolic disorders characterized by high blood sugar levels over a prolonged period. This high blood sugar is caused by either the pancreas not producing enough insulin, or the body's cells not responding properly to the insulin produced. Insulin is a hormone that regulates blood sugar levels by allowing glucose to enter cells for energy. When insulin is insufficient or ineffective, glucose builds up in the bloodstream, leading to various health problems.

Additional diagnostic methods, including fasting blood glucose tests and HbA1c measurements, are invasive, costly, and often inaccessible in low-resource or remote regions. This creates a critical gap in early detection, particularly in underserved populations where diabetes prevalence is rising alarmingly. Diabetes mellitus, a condition affecting 1 in 10 adults globally, is no longer just a medical challenge—it's a societal one. While blood tests remain the gold standard for diagnosis, their reliance on labs and trained personnel leaves millions in underserved communities without access to timely screening. Imagine a world where a tape measure and a smartphone could predict diabetes risk as effectively as a blood test. Anthropometric parameters are physical measurements like height, weight, BMI, waist circumference, hip circumference, even things like skinfold thickness.

Anthropometric parameters are measurements of the human body's size, shape, and composition. They are commonly used in health and nutrition assessments to evaluate growth, development, and overall health status. Anthropometric parameters—such as Body Mass Index (BMI), waist circumference, waist-to-hip ratio (WHR), and skinfold thickness—offer a promising alternative for diabetes detection. These measurements are non-invasive, cost-effective, and simple to collect, making them ideal for large-scale preventive healthcare initiatives. However, interpreting these parameters manually for diabetes prediction remains challenging due to complex interactions between

metabolic and body composition factors. Extensive research has established strong correlations between these parameters and diabetes risk, particularly due to their association with visceral fat and insulin resistance. Using body measurements could make screening more accessible, especially in resource-limited areas. Anthropometric parameters are foundational to non-invasive diabetes screening, especially in resource-limited settings. When combined with machine learning, they enable scalable, early detection and empower preventive healthcare strategies. Addressing ethnic variability and integrating dynamic data will further enhance their utility.

Type 2 diabetes mellitus (T2DM) is the most common diabetes category characterized as hyperglycemia due to insulin resistance or insufficient insulin production in the body. Recently researcher reported prediction models used lab-test-based measurements to diagnose if a person has diabetes or not with promising accuracy. A more preliminary diagnosis solution without any lab test measurement is more demanded. Traditional lab-tests based diabetic detection methods are time-consuming and expensive is done frequently. In general, clinicians take approximate prediction and diagnosis of diabetes mellitus of patients by taking oral glucose tolerance, fasting blood sugar, or random blood sugar tests Therefore, research on anthropometric measurements' features and their impact on diabetes detection models should be performed. In all data-driven diabetes detection classification solutions, machine learning is one of the most popular options due to its classification function using statistical approaches, without requirement of high computation power. The relationship between possible risk factors in causing diabetes is certainly nonlinear and statistical approaches are useful in studying nonlinear relationship. A more preliminary diagnosis solution without any lab test measurement is more demanded. Therefore, research on anthropometric measurements' features and their impact on diabetes detection models should be performed.

II. LITERATURE REVIEW

Maryam Saberi-Karimian et al. conducted a study and found that in both men and women, the anthropometric parameters like Waist-to-height ratio (WHtR), Body Adiposity Index (BAId), A Body Shape Index (ABSI), Body Mass Index (BMI), and Waist Circumference (WC) are predictors contributing to diabetes. WC was the strongest predictors of T2DM. Also, both in men and women BAI, and WC were the most associated factors with T2DM. DT analysis identified BAI, HC, and WC as the best variables to categorize data in men and women. Therefore, it was recommended to use these anthropometric factors for predicting the risk and screening of T2DM. The study suggests a causal relationship between anthropometric measurements and T2DM. In this study, they used new analyzing methods including machine learning algorithms such as decision tree to arrange T2DM predictors. The limitation of the study is that nearly half of all individuals with T2DM were older adults (aged $\geq$ 65 years) but we only included subjects aged between 35 and 65, so we might have lost many potential cases of T2DM that could affect data analysis.

Wee, B.F et al. conducted comparative study of machine learning-based Diabetes Detection Models, the Deep Learning-based models were able to perform better, with an average accuracy of 86.7% achieved in this diabetes classification function. The other popular deep learning-based models: CNN, DNN, and MLP. In general, they achieved an average accuracy of 84.37%, 98.1%, and 81.49% respectively. The research gap identified were selection of data collection and standardization of datasets for machine learning or deep learning-based diabetes classification tools are also topics that should be investigated in inventing a modern data-driven diabetes classification tool. Wei J, et al. found associations between visceral adiposity index (VAI), body shape index and diabetes in adults were inconsistent. They assessed the predictive capacity of VAI and body shape index for diabetes by comparing them with body mass index (BMI) and waist circumference (WC). They used the data of 5838 Chinese men and women aged ≥18 years from the 2009 China Health and Nutrition Survey. Multivariate logistic regression analysis was performed to examine the independent associations between Chinese VAI (CVAI) or body shape index and diabetes. The predictive power of the two indices was assessed using the receiver-operating characteristic (ROC) curve analysis, and compared with those of BMI and WC. Both CVAI and body shape index were positively associated with diabetes. The odds ratios for diabetes were 4.9 (2.9–8.1) and 1.8 (1.2–2.8) in men, and 14.2 (5.3–38.2) and 2.0 (1.3–3.1) in women for the highest quartile of CVAI and body shape index, respectively. The area under the ROC (AUC) and Youden index for CVAI was the highest among all four obesity indicators, whereas BMI and WC are better indicators for diabetes screening. Higher CVAI and body shape index scores are independently associated with diabetes risk. CVAI has a higher overall diabetes diagnostic ability than BMI, WC and body shape index in Chinese adults. BMI and WC, however, are more appealing as screening indicators considering their easy use.

Hosseini Net al. in their study identified key indicators strongly associated with diabetes: BRI, BAI, and MAC in males, and BMI, BRI, and MAC in females. Among the machine learning models tested (SVM, ANN, KNN), KNN performed the best, with superior accuracy, F1-score, precision, and sensitivity (71.08% for males, 79.87% for females). The ROC curve and AUC further supported these results.

When compared to a previous study using logistic regression (LR) and decision trees (DT), which found WC, BIA, Demispan, and HC as significant predictors, the current study's KNN model achieved higher accuracy (93.28% for males, 93.49% for females), outperforming the earlier DT model (77.59% for males, 79.77% for females).The best predictors: BRI, BAI, MAC (males); BMI, BRI, MAC (females).The other improvement over prior research found KNN significantly outperformed previous DT and LR models in accuracy.

Lugner, M et al. in his study used XGBoost machine learning to predict 10-year type 2 diabetes (T2D) risk using UK Biobank data from 448,277 participants (aged 40–69, no prior diabetes). The key finding of this study found HbA1c (strongest predictor), followed by BMI, waist circumference, blood glucose, family history of diabetes, gamma-glutamyl transferase, waist-hip ratio, HDL cholesterol, age, and urate and other biological factors (e.g., HbA1c, BMI) outperformed lifestyle/socioeconomic factors (e.g., diet, physical activity).

III. METHODOLOGY

Figure (a) shows flowchart of proposed model for prediction of diabetes using anthropometric parameters and machine learning techniques, which can work online and offline mode, also provides the additional feature such as to provide a pre-trained classifier on a database. The classifier trained on a larger database is suitable to generalize the performance of the classifier. The various steps are as follows:

1) The input data is applied to proposed diagnostic model in required format. The data contain both qualitative and quantitative data.
2) The quantitative data is normalized while qualitative contain nominal data such as Male or Female, in case ordinal data contains interval between category which may not be uniform such high, medium and low. The data is reprocessed at this stage.
3) The next stage, include statistical analysis and feature selection process to reduce dimensionality.
4) Feature selection module is followed by classifier design, training and testing. Variety of classifier were used at this stage.
5) The final stage is used to predict class label based on input data and validate.

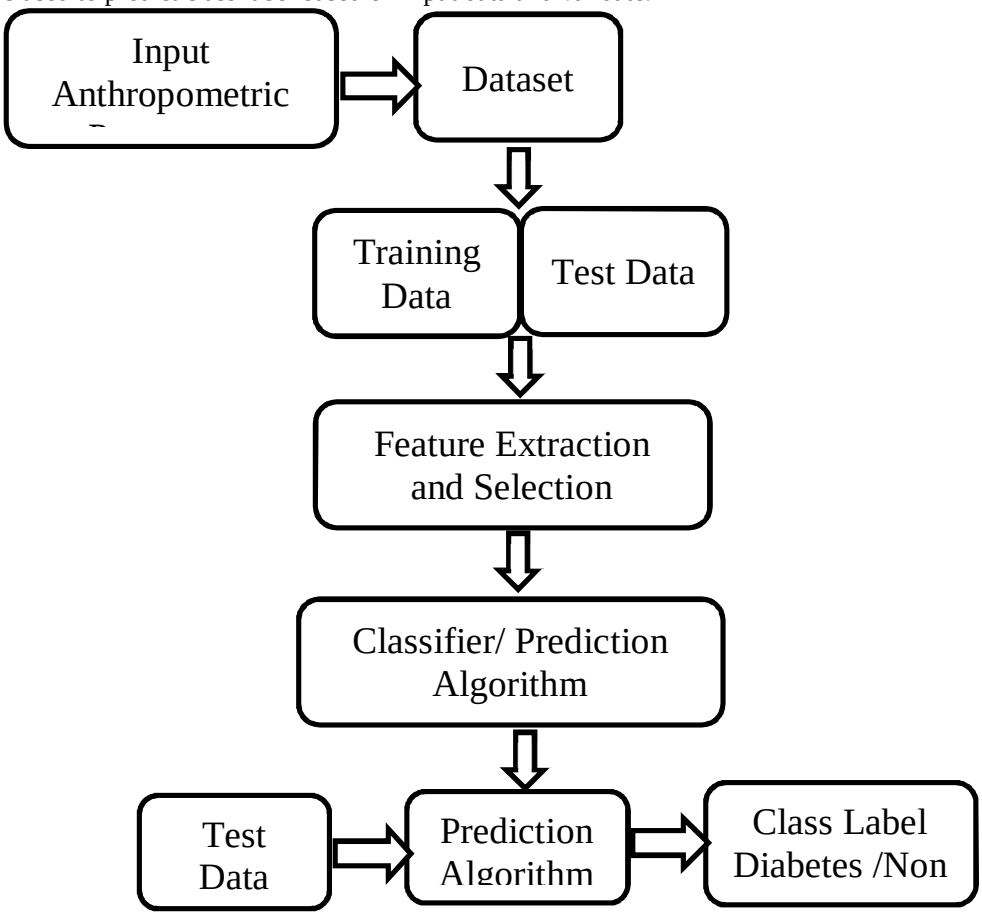


Fig (a): Flowchart of working methodology

A. Data Collection

The dataset used in this study was obtained from reference which is open source database of trial data for diabetes with corresponding anthropometric parameters, containing 10,000 data records obtained from 10,000 subjects. The dataset contains 4835 normal subjects and 5165 subject suffering from type 2 diabetes. The anthropometric parameters recorded are as given below in table 1 with their normal range.

Table (1): Description of Qualitative and Quantitative data.

Categorical Data (Qualitative)		
S.No	Parameter	Normal range/category
1	Ethnicity	Hispanic, Asian, White, Black
2	Smoking Status	Never, Former, Current
3	Family History of Diabetes	Nominal (Yes/No)
4	Previous Gestational Diabetes	Nominal (Yes/No)
5	Physical Activity Level	Sedentary, Moderate, Active
Numerical Data (Quantitative)		
6	BMI (Body Mass Index)	
7	Waist Circumference	cm/inches
8	Fasting Blood Glucose	mg/dL or mmol/L
9	HbA1c	
10	Blood Pressure Systolic	Mm HG
11	Blood Pressure_ Diastolic	mmHg
12	Cholesterol Total	mg/dL
13	Cholesterol HDL	mg/dL
14	Cholesterol LDL	mg/dL
15	GGT(Gamma-Glutamyl Transferase	U/L
16	Serum Urate	mg/dL
17	Dietary Intake Calories	kcal/day
18	Alcohol Consumption	grams/day or drinks/week

B. Classification Model

We considered three classification models for prediction of diabetes using anthropometric parameters; the models are logistic regression, decision tree and random forest.

1) *Logistic Regression:* Logistic Regression is a supervised machine learning algorithm used for binary classification (though it can be extended to multiclass). Unlike linear regression, which predicts continuous values, logistic regression predicts the probability that an instance belongs to a particular class. Logistic regression uses the sigmoid function to map predicted values to probabilities between 0 and 1:

$$\sigma(z) = {}^{1}/_{(1 + e^{-z})}$$

Where $z=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta nx_n$ (linear combination of inputs)

Output σ(z)is the probability P(y=1|x) $P(y=1|x)$. The classification is based on decision boundary if σ(z)≥ 0.5predict 1 else 0. Logistic regression uses Log Loss cost as a cost function. The logistic regression best works when relationship between features and log-odds is linear.

2) Decision Tree: The Decision Tree algorithm is a simple classic supervised learning model that works surprisingly well. The classifiers discussed above all expect attribute values to be presented at the same time. A decision tree uses a treelike graph that represents a flow-chart-like structure in which the starting point is the root. Each internal node of the tree represents a test on an attribute or subset of attributes.

Selecting the Best Attribute**:** The first step in building a decision tree is selecting the best attribute to split the dataset. This is done using criteria such as information gain or Gini index. The attribute that results in the highest gain (or lowest impurity) is chosen as the root node.

Recursive Splitting: Once the best attribute is selected, the data is split into subsets based on the value of that attribute. This process is repeated recursively for each subset, creating branches and sub-branches until the data is perfectly split or meets a stopping criterion (e.g., maximum depth or minimum number of instances per leaf.)

Stopping Criteria: The decision tree continues splitting the data until one of the following conditions is met:

- All instances in a node belong to the same class.
- The maximum depth of the tree is reached.
- The number of instances in a node falls below a specified threshold.

Tree Pruning: After the decision tree is fully grown, it may be necessary to prune it to remove overfitting. Pruning helps simplify the model by removing branches that do not contribute much to the accuracy of the model on unseen data.

3) Random Forest: It is a popular ensemble learning method used for classification and regression. It works by constructing multiple decision trees during training and outputs the average prediction (for regression) or majority vote (for classification) of the individual trees. Random forest uses Bootstrap Aggregation approach uses random subsets of the training data are selected with replacement. At each split in a decision tree, only a random subset of features is considered. For predicting class Random Forest uses Voting/Averaging Predictions from various tree classifier. The main advantage of it reduces overfitting as compared to single decision trees. The key tuning parameters are as follow:

Number of estimators: Number of trees (more trees → better performance but slower).
Max_depth: Maximum depth of each tree.
Max_features: Number of features considered for splitting (e.g., sqrt(n_features)).
Min_samples_split: Minimum samples required to split a node.
Bootstrap: Whether to use bootstrapping (default: True).

IV. RESULT

Model Result: We had tested variety of model such as 'Logistic Regression, Decision Tree and Random Forest using k fold validation where k is set to 5. For Logistic Regression (LR) maximum iteration was set 1000, the classification accuracy achieved by LR in terms of F1 Score of 0.963707. While Decision Tree tuning parameter such as maximum of depth of split set to 5 with minimum sample split of 10. The classification accuracy achieved by Decision Tree F1 score 1. In case of Random Forest tuning parameter such as number of estimators set to 200, maximum depth of 10 and minimum sample split of 5.The classification accuracy achieved by Random Forest F1 score 1 as shown in Table 2.

Sr No	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
1	Logistic Regression	0.963969	0.970754	0.956763	0.963707	0.994798
2	Decision Tree	1.00	1.00	1.00	1.00	1.00
3	Random Forest	1.00	1.00	1.00	1.00	1.00

Table 2: Result of Classification using 20 features

The highest accuracy of classification achieved by Decision tree and Random Forest, the choice of classifier depend on computing recourse, Random Forest is computationally expensive compare to Decision Tree, but avoid overfitting of data.

The confusion matrix in table 3 represents the performance of a logistic regression classifier in predicting diabetes status (diabetic vs. non-diabetic).

		NO diabetes	Diabetes
Actual	NO diabetes	1752	52
	Diabetes	78	1726
Predicted			

Table 3: Confusion matrix Logistic regression

This logistic regression model demonstrates excellent performance for diabetes prediction, with balanced precision and recall. Its simplicity and interpretability make it ideal for medical applications, though ensemble methods (XGBoost) may offer marginal gains in accuracy.

		NO diabetes	Diabetes
Actual	NO diabetes	1804	0
	Diabetes	0	1804
Predicted			

Table 4: Confusion matrix Decision tree and Random Forest

The above table 4 shows the confusion matrix representing the performance of a Decision tree and Random forest classifier in predicting diabetes status (diabetic vs. non-diabetic). Both the classifiers provided same accuracy, precision, recall value, F1 score and ROC AUC.

V. CONCLUSION

The future of diabetes detection using anthropometrics parameters and Machine Learning lies in personalized, accessible, and explainable AI solutions that can seamlessly integrate into healthcare systems. By leveraging emerging technologies and expanding datasets, these models can revolutionize early diagnosis and preventive care globally.

It presents a cost-effective, non-invasive, and scalable approach to early diagnosis and risk assessment. By leveraging easily measurable body metrics such as BMI, waist circumference, waist-to-hip ratio, and skinfold thickness, combined with advanced ML algorithms, this method offers a promising alternative to traditional blood-based tests, particularly in low-resource settings.

1) High Accuracy: Machine learning models (e.g., Random Forest, SVM, Neural Networks) can effectively predict diabetes risk using anthropometric and demographic data.
2) Early Detection: Enables identification of high-risk individuals before clinical symptoms appear, allowing for timely lifestyle interventions.
3) Accessibility: Reduces dependency on invasive tests, making screening feasible in remote areas with limited healthcare infrastructure.
4) Integration Potential: Can be combined with wearable devices, mobile health apps, and electronic health records for continuous monitoring.

However, challenges such as dataset bias, model interpretability, and the need for clinical validation must be addressed to ensure reliability and widespread adoption. Future advancements in explainable AI (XAI), federated learning, and multi-modal data integration (combining anthropometrics with genetic and biochemical markers) will further enhance predictive performance.

In conclusion, Machine Learning driven diabetes detection using anthropometric parameters holds significant potential to revolutionize preventive healthcare, reduce global diabetes burden, and improve patient outcomes through early, affordable, and non-invasive screening. Continued research, real-world validation, and collaboration between AI experts and medical professionals will be crucial for its successful implementation.

REFERENCES

[1] Maryam Saberi-Karimian, Amin Mansoori and Maryam Mohammadi Bajgiran et al., Data mining approaches for type 2 diabetes mellitus prediction using anthropometric measurements, Journal of clinical laboratory analysis, 12 December 2022 https://doi.org/10.1002/jcla.24798.

[2] Wei J, Liu X, Xue H, Wang Y, Shi Z. Comparisons of Visceral Adiposity Index, Body Shape Index, Body Mass Index and Waist Circumference and Their Associations with Diabetes Mellitus in Adults. Nutrients. 2019; 11(7):1580. https://doi.org/10.3390/nu11071580

[3] Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE, Bright D, Williams R (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 Res Clin Pract 157:107843. https://doi.org/10.1016/j.diabres.2019.107843.

[4] Chawla R, Madhu S, Makkar B, Ghosh S, Saboo B, Kalra S et al (2020) Rssdi-esi clinical practice recommendations for the management of type 2 diabetes mellitus 2020. Indian J. Endocrinol. Metab 24(1):1

[5] Sacks DB, Arnold M, Bakris GL, Bruns DE, Horvath AR, Kirkman MS, Lernmark A, Metzger BE, Nathan DM (2011) Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. Diabetes Care 34(6):61–99. https://doi.org/10.2337/dc11-9998https://diabetesjournals.org/care/article-pdf/34/6/e61 /609322/e61.pdf.

[6] Kazmi NHS, Gillani S, Afzal S, Hussain S (2013) Correlation between glycated haemoglobin levels and random blood glucose. J Ayub Med Coll 25(1–2):86–88Return to ref 25 in article

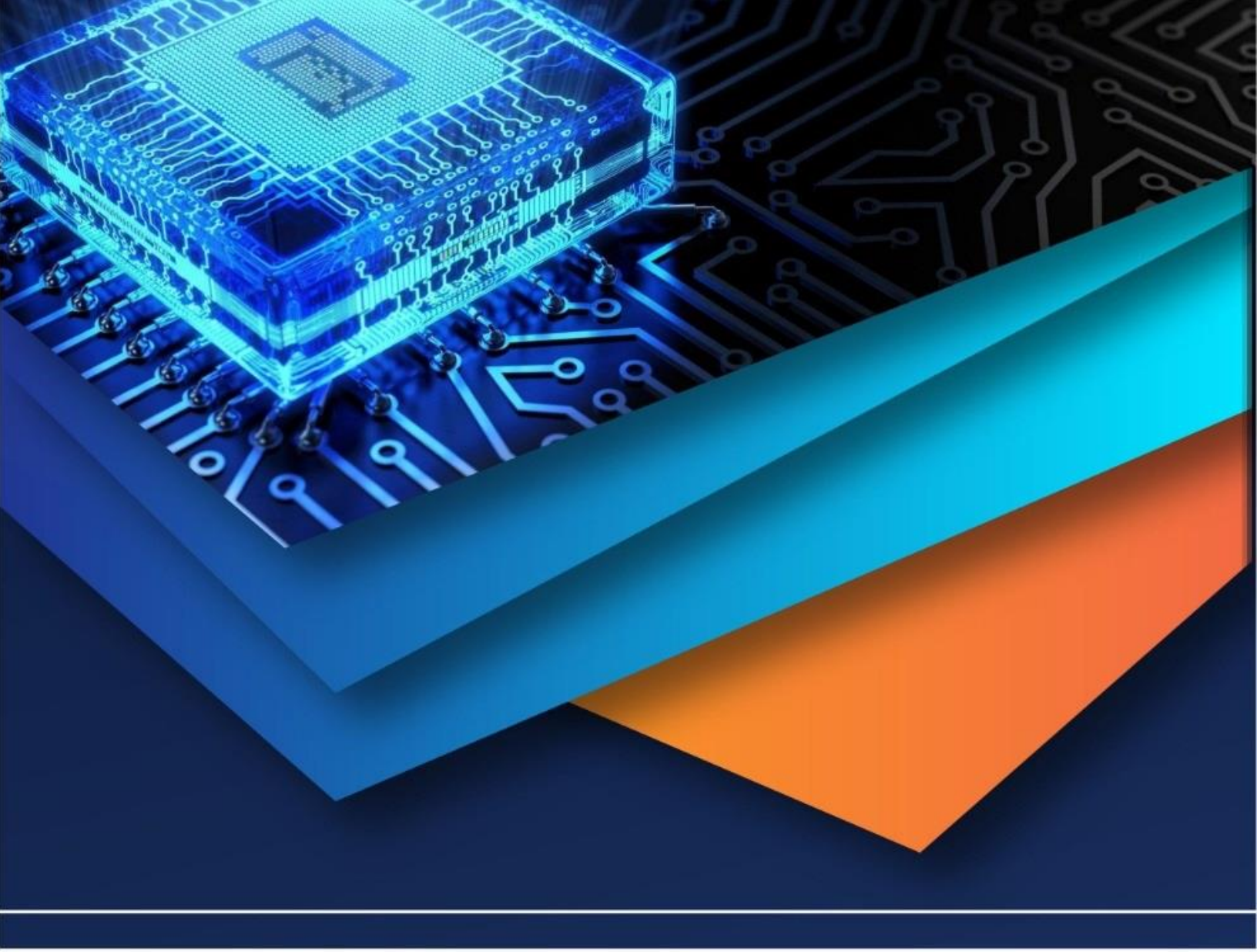
[7] Zaccardi F, Dhalwani NN, Papamargaritis D, Webb DR, Murphy GJ, Davies MJ, Khunti K (2017) Nonlinear association of bmi with all-cause and cardiovascular mortality in type 2 diabetes mellitus: a systematic review and meta-analysis of 414,587 participants in prospective studies. Diabetologia 60(2):240–248.

[8] Wee, B.F., Sivakumar, S., Lim, K.H. et al. Diabetes detection based on machine learning and deep learning approaches. Multimed Tools Appl 83, 24153–24185 (2024). https://doi.org/10.1007/s11042-023-16407-5.

[9] Lugner, M., Rawshani, A., Helleryd, E. et al. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. Sci Rep 14, 2102 (2024). https://doi.org/10.1038/s41598-024-52023-5

[10] Hosseini N, Tanzadehpanah H, Mansoori A, Sabzekar M, Ferns GA, Esmaily H, Ghayour-Mobarhan M. Using a robust model to detect the association between anthropometric factors and T2DM: machine learning approaches. BMC Med Inform Decis Mak. 2025 Jan 31;25(1):49. doi: 10.1186/s12911-025-02887-y. PMID: 39891090; PMCID: PMC11786328.

[11] Liu G, Li Y, Hu Y, Zong G, Li S, Rimm EB, Hu FB, Manson JE, Rexrode KM, Shin HJ et al (2018) Influence of lifestyle on incident cardiovascular disease and mortality in patients with diabetes mellitus. J Am Coll Cardiol 71(25):2867–2876.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY