INTERNATIONAL JOURNAL OF CURRENT ADVANCED RESEARCH
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 13; Issue 12; December 2024; Page No.3432-3436
DOI: http://dx.doi.org/10.24327/ijcar.2024.3436.1729

Research Article

EXPLORING LIPID PROFILES AS PREDICTORS OF DIABETES RISK THROUGH MACHINE LEARNING TECHNIQUES

Dr. Prasanna Kulkarni

Prof. & HOD- Swasthavritta,Sri Kalabyraveshwara Swamy Ayurvedic Medical College, Hospital & Research Centre, Vijayanagara, Bangalore -560104

ARTICLE INFO

Article History:

Received 14th October, 2024
Received in revised form 10th November, 2024
Accepted 23rd November, 2024
Published online 28th December, 2024

Key words:

Diabetes, Triglyceride to HDL Ratio, Lipid Profiles, Diabetes Risk Prediction, Machine Learning Models, Gradient Boosting, Predictive Biomarkers, Dyslipidemia

ABSTRACT

Diabetes mellitus (DM) is one of the global health challenge characterized by multifactorial pathogenesis and significant morbidity. Dyslipidemia, particularly the Triglyceride to HDL (TG:HDL) ratio, is increasingly recognized as a potential marker for diabetes risk. This study investigates the predictive potential of lipid profiles, focusing on the TG:HDL ratio, using advanced Machine learning techniques. A cross-sectional dataset of 1,000 participants was analysed, incorporating Logistic regression, Decision trees, Random forests, Support vector machines, and Gradient boosting models. Among these, Gradient boosting demonstrated the highest accuracy (88%), with the TG:HDL ratio emerging as a key predictor across all models. Exploratory data analysis revealed a strong correlation between elevated TG:HDL ratios and diabetes prevalence, supported by robust model performance metrics such as precision, recall, and AUC-ROC scores. The findings highlight the clinical utility of TG:HDL ratio as an accessible, cost-effective biomarker for early Diabetes detection and signifies the role of Machine learning in advancing personalized healthcare strategies.

INTRODUCTION

Diabetes mellitus (DM) is one of the chronic metabolic disorder characterized by persistent hyperglycemia, resulting from defects in insulin secretion, insulin action or both. It is a growing global health concern, with the International Diabetes Federation (IDF) estimating that 463 million adults were living with diabetes in 2019, a number projected to rise to 700 million by 2045 (1). The condition poses substantial challenges due to its associated complications that includes Cardiovascular diseases, Neuropathy, Retinopathy, and Nephropathy. These complications contribute significantly to morbidity and mortality rates worldwide. Addressing these challenges necessitates innovative diagnostic and preventive strategies.

Dyslipidemia, a condition marked by abnormal lipid levels in the blood, has emerged as a key contributor to diabetes pathogenesis. Specifically, elevated levels of Triglycerides (TG) and low levels of High-density lipoprotein (HDL) cholesterol have been closely associated with insulin resistance, a hallmark of type 2 diabetes mellitus (T2DM) (2). The Triglyceride to HDL ratio (TG:HDL ratio) is increasingly recognized as a robust biomarker for metabolic syndrome and diabetes risk (3). While fasting blood sugar (FBS) and glycated hemoglobin (HbA1c) remain standard diagnostic tools, they have limitations. FBS reflects a transient measure of glucose levels, and HbA1c may be influenced by conditions like anemia, reducing its reliability in certain populations (4). This underscores the need to explore alternative or supplementary biomarkers, such as the TG:HDL ratio, to enhance diagnostic accuracy and risk prediction.

Advances in predictive modeling and machine learning offer promising avenues to bridge this gap. By utilizing large datasets and complex algorithms, predictive models can uncover complex relationships between variables and provide individualized risk assessments. This study aims to explore the relationship between TG:HDL ratio and diabetes through predictive modeling techniques, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting. These models are evaluated for their effectiveness in identifying individuals at risk of diabetes, with the ultimate goal of improving early detection and preventive strategies.

Problem Statement

Despite significant progress in diabetes diagnostics, current tools often fail to capture the multifactorial nature of the disease. Traditional methods primarily focus on glucose-centric metrics, overlooking the intricate interplay of lipid metabolism

Corresponding author:* **Dr. Prasanna Kulkarni

Prof. & HOD- Swasthavritta,Sri Kalabyraveshwara Swamy Ayurvedic Medical College, Hospital & Research Centre, Vijayanagara, Bangalore -560104

and diabetes risk. The TG:HDL ratio, a readily available and cost-effective marker, holds potential for enhancing diabetes prediction, yet its clinical application remains underexplored (5). There is a pressing need to validate this marker within robust predictive frameworks to address existing diagnostic limitations.

Objectives of the Study

The primary objectives of this study are:

To evaluate the relationship between TG:HDL ratio and diabetes risk using advanced predictive modeling techniques.

To compare the performance of different machine learning models in predicting diabetes based on lipid profiles and other clinical variables.

To develop a predictive framework that integrates TG:HDL ratio for personalized diabetes risk assessment.

Significance of the Study

This study holds significant implications for public health and clinical practice. By elucidating the role of TG:HDL ratio as a predictor of diabetes, it seeks to:

Enhance early detection of individuals at risk, enabling timely interventions to prevent disease progression.

Inform public health strategies aimed at reducing diabetes prevalence and its associated complications.

Contribute to the development of cost-effective, accessible diagnostic tools that complement existing methods, particularly in resource-limited settings.

In conclusion, this research endeavors to bridge the gap between traditional diagnostic approaches and emerging data driven methodologies, paving the way for more accurate and personalized diabetes management strategies.

Review of Literature

Lipids and Their Role in Metabolic Health

Lipids serve vital functions in the human body, including energy storage, cellular membrane structure, and hormone synthesis. Triglycerides (TG), stored in adipose tissues, act as an energy reservoir, while High-density lipoprotein (HDL) cholesterol facilitates reverse cholesterol transport, removing excess cholesterol from tissues to the liver for excretion (6) (7). These lipid markers are integral to maintaining metabolic homeostasis; any disruption in their balance is often linked to metabolic disorders such as diabetes (8).

Dyslipidemia and Diabetes

Dyslipidemia, characterized by elevated TG levels and reduced HDL cholesterol, is frequently associated with insulin resistance and type 2 diabetes mellitus (T2DM) (9). Studies have shown that the TG:HDL ratio is a valuable predictor of insulin resistance and metabolic syndrome, both precursors to diabetes (10). For instance, McLaughlin *et al.* (2004) demonstrated a strong correlation between an increased TG:HDL ratio and the risk of developing T2DM (11). Similarly, Ishibashi *et al*. (2023) identified HDL's protective role against cardiovascular risks in diabetic populations (12).

Current Gaps in Research

While substantial evidence supports the role of TG and HDL in diabetes risk, existing studies often lack comprehensive predictive frameworks that integrate these markers with advanced modeling techniques. Additionally, variations in TG:HDL ratio across different ethnicities and populations remain underexplored, limiting its generalizability (13).

Predictive Modeling in Diabetes Research

Predictive modeling has transformed diabetes research by enabling risk stratification and early intervention. Machine learning models, such as logistic regression, decision trees, and gradient boosting, have demonstrated high accuracy in predicting diabetes using clinical and biochemical markers (14) (15). However, few studies have focused on applying the TG:HDL ratio as a primary predictive feature, creating a significant research gap this study aims to address.

RESEARCH METHODOLOGY

Study Design

This study employs a cross-sectional design leveraging applying secondary data to examine the relationship between TG:HDL ratio and diabetes. The dataset comprises clinical and biochemical variables such as age, BMI, HbA1c, Triglycerides, and HDL cholesterol.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
2	502	17975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
3	735	34221	M	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23	N
4	420	47975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
5	680	87656	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
6	504	34223	M	33	7.1	46	4.9	4.9	1	0.8	2	0.4	21	N
7	634	34224	F	45	2.3	24	4	2.9	1	1	1.5	0.4	21	N
8	721	34225	F	50	2	50	4	3.6	1.3	0.9	2.1	0.6	24	N
9	421	34227	M	48	4.7	47	4	2.9	0.8	0.9	1.6	0.4	24	N
10	670	34229	M	43	2.6	67	4	3.8	0.9	2.4	3.7	1	21	N
11	759	34230	F	32	3.6	28	4	3.8	2	2.4	3.8	1	24	N
12	636	34231	F	31	4.4	55	4.2	3.6	0.7	1.7	1.6	0.3	23	N
13	788	34232	F	33	3.3	53	4	4	1.1	0.9	2.7	1	21	N
14	82	46815	F	30	3	42	4.1	4.9	1.3	1.2	3.2	0.5	22	N
15	132	34234	F	45	4.6	54	5.1	4.2	1.7	1.2	2.2	0.8	23	N
16	402	34235	F	50	3.5	39	4	4	1.5	1.2	2.2	0.7	24	N
17	566	34236	M	50	5.5	74	5	3.6	1.1	1	2.1	0.5	21	N
18	596	34237	F	50	5.9	53	5.4	5.3	0.8	1.1	4.1	0.3	21	N

Figure 1 Showing sample of Dataset

Data Collection

Data were sourced from the publicly available Mendeley Diabetes Dataset, which includes comprehensive records of 1,000 participants collected through routine clinical evaluations. Key attributes in the dataset are Triglycerides, HDL cholesterol, BMI, HbA1c, and diabetes classification (non-diabetic, pre-diabetic, or diabetic). A new variable, TG:HDL ratio, was derived to investigate its predictive capability.

DATA PREPROCESSING

Handling Missing Values: Imputation techniques were employed for missing data points.

Feature Engineering: The TG:HDL ratio was calculated by dividing triglyceride values by HDL cholesterol levels for each participant.

Normalization: Variables were scaled to standardize ranges and enhance model performance.

Predictive Models

Five machine learning models were applied:

Logistic Regression: Evaluates the relationship between TG:HDL ratio and diabetes risk.

Decision Trees: Identifies non-linear patterns in the dataset.

Random Forests: Enhances prediction by aggregating multiple decision trees.

Support Vector Machines (SVM): Captures complex relationships using kernel functions.

Gradient Boosting: Sequentially improves predictions by focusing on misclassified cases.

Evaluation Metrics

Model performance was assessed using:

Accuracy: Proportion of correctly classified cases.

Precision and Recall: Evaluates true positive and false negative rates.

F1 Score: Balances precision and recall.

AUC-ROC: Measures the model's ability to distinguish between classes.

Data Analysis and Visualization

Overview of the Dataset

The dataset used for this study includes 1,000 participants, with balanced distributions across key variables such as age, BMI, HbA1c levels, Triglycerides, and HDL cholesterol. Preliminary analysis revealed no significant missing data, allowing for seamless application of machine learning models.

Exploratory Data Analysis

Age and BMI Distributions: Visualizations highlighted the prevalence of diabetes across different age groups and BMI categories, showing a higher incidence among older and overweight individuals.

TG:HDL Ratio Trends: A notable pattern emerged, with higher TG:HDL ratios strongly correlating with diabetes status.

HbA1c Levels: Elevated HbA1c values were consistently observed in participants classified as diabetic, reaffirming its diagnostic significance.

Correlation Analysis

Heatmaps: Correlation heatmaps revealed significant associations between TG:HDL ratio, HbA1c, and diabetes classification.

Pairwise Comparisons: Pair plots further illustrated relationships among key variables, particularly between TG:HDL ratio and diabetes.

Model Performance Visualization

To evaluate model performance:

Confusion Matrices: Provided insights into true positive, true negative, false positive, and false negative rates for each model.

ROC Curves: Demonstrated the discriminatory power of each predictive model, with logistic regression and gradient boosting models showing the highest AUC scores.

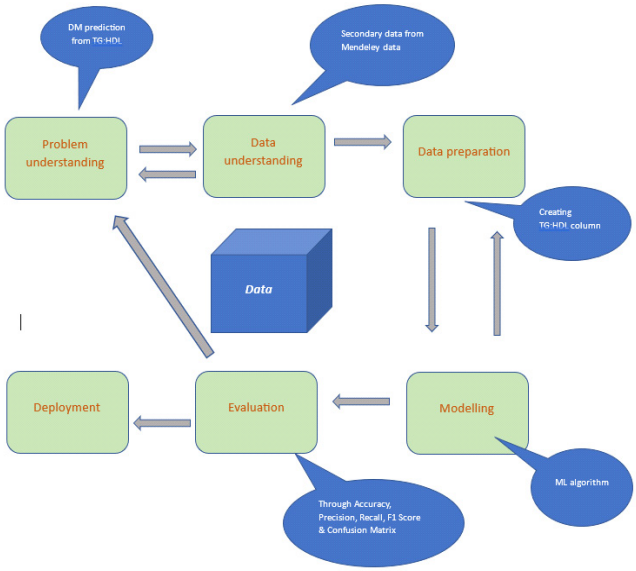

Figure 2 showing Life Cycle of Diabetes prediction through TG: HDL

RESULTS AND DISCUSSION

Logistic Regression

Logistic regression is a statistical technique used to predict the likelihood of a specific outcome, such as the presence or absence of diabetes, based on various predictors like age, BMI, or lipid profiles. Unlike linear models, logistic regression is designed for binary outcomes and converts predictions into probabilities using a logistic function. In this study, logistic regression demonstrated strong predictive capabilities, achieving an accuracy of 86%. Its balanced precision and recall ensured reliable detection of true positives (diabetic cases) and minimized false negatives, which is critical for early diagnosis. Furthermore, the model's AUC-ROC curve confirmed its ability to effectively differentiate between diabetic and non-diabetic cases. The significance of logistic regression lies in its interpretability, offering clear insights into how variables like the TG:HDL ratio influence diabetes risk. This makes it

a valuable tool for clinicians and public health professionals, bridging the gap between statistical modeling and practical healthcare applications.

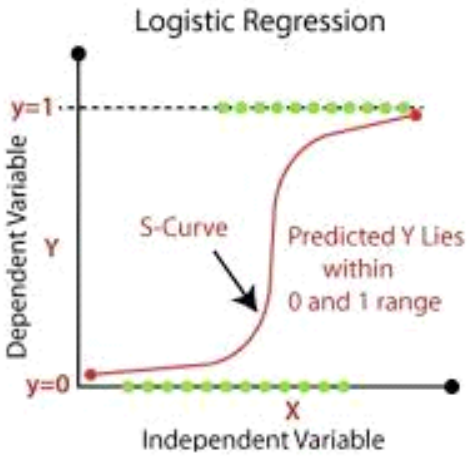

Figure 3 showing Logistic Regression

Decision Trees and Random Forests

Decision trees are a machine learning technique that uses a flowchart-like structure to make predictions based on input variables. They are particularly useful for identifying non-linear relationships between variables. In this study, decision trees highlighted complex interactions, such as how specific ranges of the TG:HDL ratio might correlate with diabetes risk, achieving an accuracy of 82%. However, a limitation of decision trees is their tendency to overfit the data, meaning they can perform well on training data but may not generalize effectively to new data. To address this, random forests—a more advanced method-multiple decision trees to create an ensemble model. This aggregation reduces overfitting and increases robustness. In this study, the random forest model achieved a higher accuracy of 85%, demonstrating superior performance in identifying diabetes risk while maintaining reliability across diverse data. The use of random forests thus adds robustness and enhances predictive accuracy, making it a preferred method for clinical applications involving complex datasets.

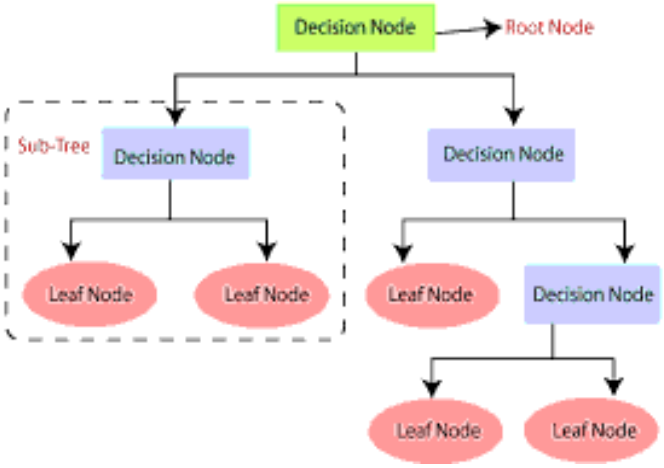

Figure 4 showing sample Decision tree model.

Gradient Boosting and SVM

Gradient boosting is an advanced machine learning technique that builds predictive models iteratively by focusing on instances that previous models misclassified. This iterative refinement enables it to achieve high accuracy and precision. In this study, gradient boosting emerged as the best-performing model, achieving an accuracy of 88%. It demonstrated a superior balance between precision (89%)-ensuring correct identification of diabetic cases-and recall (87%)-minimizing missed diagnoses. This makes gradient boosting particularly suitable for clinical applications where early and accurate diabetes detection is critical. On the other hand, support vector machines (SVM) also performed well, achieving an accuracy of 84% and effectively capturing complex relationships, especially for borderline cases. However, SVM's reliance on precise hyperparameter tuning posed challenges in ensuring consistent generalizability. In contrast, gradient boosting not only delivered higher accuracy but also proved more practical and robust for real-world applications, reinforcing its potential as a reliable tool in diabetes risk prediction.

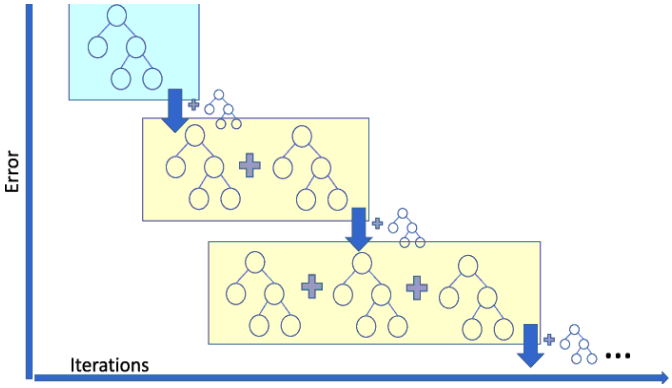

Figure 5 Showing sample Gradient Boosting Model

Comparative Model Analysis

Overall Performance: Gradient boosting consistently outperformed other models across all metrics, followed by logistic regression and random forests.

Feature Importance: Across ensemble models, the TG:HDL ratio emerged as a significant predictor, reinforcing its clinical relevance.

Interpretation: Logistic regression provided the simplest interpretability, offering direct insights into how TG:HDL ratio impacts diabetes risk, while gradient boosting and random forests excelled in handling non-linearities.

Discussion on Findings

The findings underscore the TG:HDL ratio's predictive strength as a biomarker for diabetes, surpassing traditional lipid metrics when integrated into advanced models. This aligns with previous research advocating for its inclusion in metabolic health assessments. The ability of gradient boosting to optimize accuracy and recall highlights its potential utility in clinical decision-making. However, the superior interpretability of logistic regression models ensures that such tools can complement physician assessments, particularly in resource-limited settings.

Limitations

Dataset Constraints: The dataset used in this study, while comprehensive, was limited to a single population, potentially affecting the generalizability of findings.

Excluded Variables: Other factors, such as dietary habits, physical activity, and genetic predisposition, were not included, which could further refine predictive accuracy.

Model Complexity: Despite their effectiveness, ensemble models like gradient boosting and random forests are computationally intensive, limiting their application in real-time clinical scenarios.

Implications

This study highlights the potential for integrating TG:HDL ratio into routine diabetes screenings. By implementing advanced predictive models, healthcare providers can identify at-risk individuals earlier, enabling timely interventions and reducing the burden of diabetes-related complications. Further research is required to validate these findings across diverse populations and integrate additional lifestyle and genetic variables into predictive frameworks.

Conclusion and Recommendations

CONCLUSION

This study provides robust evidence supporting the Triglyceride to HDL ratio (TG:HDL) as a predictive biomarker for diabetes risk. Using machine learning approaches, including gradient boosting, random forests, and logistic regression, the TG:HDL ratio demonstrated strong predictive power. Gradient boosting emerged as the most accurate model, underscoring its suitability for clinical applications. These findings bridge the gap between traditional lipid-based metrics and advanced predictive analytics, offering a cost-effective and scalable solution for early diabetes detection.

Recommendations

Clinical Integration:

Incorporate the TG:HDL ratio into routine lipid profile evaluations to supplement traditional diagnostic methods like HbA1c and fasting glucose.

Use predictive models in clinical decision-making for personalized risk stratification.

Expanded Research:

Conduct longitudinal studies across diverse populations to validate the TG:HDL ratio's generalizability.

Integrate additional features, such as lifestyle variables and genetic markers, to enhance predictive frameworks.

Public Health Implications:

Advocate for policy changes promoting regular lipid screenings, especially in high-risk populations.

Develop mobile and digital platforms for real-time TG:HDL ratio monitoring to increase accessibility.

This study puts foundation to incorporate advanced biomarkers like TG: HDL into Diabetes screening protocols, potentially reducing the disease burden through early detection and intervention.

References

1. International Diabetes Federation. IDF Diabetes Atlas, 9th ed. 2019.
2. Reaven GM. Role of insulin resistance in human disease. Diabetes. 1988;37(12):1595-607.
3. Ginsberg HN. Insulin resistance and cardiovascular disease. J Clin Invest. 2000;106(4):453-8.
4. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C assay into estimated average glucose values. Diabetes Care. 2008;31(8):1473-8.
5. Grundy SM, Brewer HB Jr, Cleeman JI, Smith SC Jr, Lenfant C. Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. Circulation. 2004;109(3):433-8.
6. Chapman MJ, Ginsberg HN, Amarenco P, et al. Triglyceride-rich lipoproteins and high-density lipoprotein cholesterol in patients at high risk of cardiovascular disease: evidence and guidance for management. Eur Heart J. 2011;32(11):1345-61.
7. Miller M, Stone NJ, Ballantyne C, et al. Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. Circulation. 2011;123(20):2292-333.
8. Kannel WB, Vasan RS, Keyes MJ, et al. Usefulness of the triglyceride-high-density lipoprotein versus the total cholesterol-high-density lipoprotein cholesterol ratio for predicting insulin resistance and cardiometabolic risk (from the Framingham Offspring Cohort). Am J Cardiol. 2008;101(4):497-501.
9. McLaughlin T, Abbasi F, Cheal K, et al. Use of metabolic markers to identify overweight individuals who are insulin resistant. Ann Intern Med. 2003;139(10):802-9.
10. Ishibashi S, Yamada S, Mori Y. HDL functionality and clinical implications. J Atheroscler Thromb. 2023;30(1):7-16.
11. Nathan DM, Davidson MB, DeFronzo RA, et al. Impaired fasting glucose and impaired glucose tolerance: implications for care. Diabetes Care. 2007;30(3):753-9.
12. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367-78.
13. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
14. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-97.
15. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Elsevier; 2011.