INSIGHTS-JOURNAL OF
LIFE AND SOCIAL SCIENCES

INSIGHTS-JLSS
Insights-Journal of Life and Social Sciences

A SYSTEMS-LEVEL APPROACH TO EARLY DETECTION OF METABOLIC SYNDROME IN ADOLESCENTS USING MACHINE LEARNING AND BIOLOGICAL MARKERS

Original Article

Rabia Zulfiqar[1]*, Nargis Khan[2], Gull Hassan Shethar[3], Tasneem Munir[4]

[1]Student, Certificate of Medical Teaching, Department of Community Medicine, King Edward Medical University, Lahore, Pakistan.
[2]Associate Professor, Department of Medicine, Dow University of Health Sciences (DUHS) Karachi, Pakistan.
[3]Consultant, Department of Medicine, Al-Amiri Hospital, Kuwait.
[4] Surgical Technologist, Lahore General Hospital, Lahore, Pakistan.

Corresponding Author: Rabia Zulfiqar, Student, Certificate of Medical Teaching, Department of Community Medicine, King Edward Medical University, Lahore, Pakistan, rabiazulfiqar@outlook.com

Conflict of Interest: None

Grant Support & Financial Support: None

Acknowledgment: The authors acknowledge the support of the data contributors and participating institutions.

ABSTRACT

Background: Metabolic Syndrome (MetS) during adolescence is increasingly prevalent and closely linked to the future onset of cardiovascular disease, type 2 diabetes, and other chronic metabolic conditions. Traditional diagnostic approaches often fail to reflect the complex, multifactorial nature of the syndrome, particularly in younger populations. An integrative, data-driven strategy is essential to improve early identification and enable targeted prevention in at-risk adolescents.

Objective: To develop and evaluate machine learning (ML)-based predictive models that integrate biological and clinical markers for early detection of MetS in adolescents.

Methods: This cross-sectional study analyzed data from 110 adolescents aged 10–18 years, including anthropometric measures (BMI, waist circumference), biochemical markers (fasting glucose, triglycerides, HDL cholesterol, fasting insulin, HOMA-IR), blood pressure, and lifestyle indicators. After data preprocessing and normalization, key features were selected using recursive feature elimination and mutual information techniques. Supervised ML models—Gradient Boosting, Random Forest, Support Vector Machines (SVM), and Neural Networks—were trained and evaluated using 10-fold cross-validation. Model performance was assessed using accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). SHAP (Shapley Additive Explanations) analysis was employed for interpretability of feature contributions.

Results: Gradient Boosting outperformed all other models with an accuracy of 90.0%, precision of 0.86, recall of 0.84, and AUC-ROC of 0.92. Random Forest followed with 89.1% accuracy and 0.91 AUC-ROC. SVM and Neural Networks achieved 85.5% and 88.2% accuracy, respectively. SHAP analysis revealed waist circumference ($r = 0.68$), triglycerides ($r = 0.63$), HOMA-IR ($r = 0.59$), fasting insulin ($r = 0.50$), and HDL cholesterol ($r = –0.56$) as the top contributors to MetS prediction.

Conclusion: Ensemble ML methods, especially Gradient Boosting, demonstrated high predictive accuracy in identifying adolescents at risk for MetS using integrated clinical and biological data. These models offer promise for early, personalized interventions and warrant validation in larger and longitudinal cohorts.

Keywords: Adolescent, Biological Markers, Cardiometabolic Risk, Early Diagnosis, Machine Learning, Metabolic Syndrome, Predictive Modeling.

INTRODUCTION

Metabolic Syndrome (MetS) is increasingly recognized as a major public health concern due to its role in predisposing individuals to cardiovascular disease and type 2 diabetes. It encompasses a cluster of interrelated risk factors, including central obesity, dyslipidemia, hypertension, and impaired glucose metabolism (1). Although traditionally associated with adult populations, mounting evidence points to a worrying rise in the prevalence of MetS among adolescents, a trend largely driven by escalating rates of childhood obesity, poor dietary habits, and physical inactivity (2). This early emergence of metabolic abnormalities is particularly concerning, as it suggests the possibility of accelerated disease progression and a greater burden of chronic illness later in life if left unaddressed (3). Timely identification of MetS during adolescence offers a unique window of opportunity to intervene before irreversible metabolic damage occurs. However, conventional diagnostic models, which rely heavily on predefined clinical thresholds, often fail to capture the complex and dynamic nature of metabolic dysfunction in this age group (4). These criteria are primarily adapted from adult frameworks and may not sufficiently account for the physiological variability and developmental changes that occur during adolescence, potentially leading to underdiagnosis or misclassification. Furthermore, current screening approaches typically assess each risk factor in isolation, disregarding the interconnected biological processes that collectively contribute to metabolic imbalance (5).

A growing body of literature supports the value of adopting a systems-level perspective in understanding and diagnosing MetS. Such an approach integrates multiple dimensions of biological information, recognizing the interplay between genetic, metabolic, hormonal, and environmental influences. When combined with advances in machine learning (ML), this strategy has the potential to transform early detection by revealing hidden patterns within complex, high-dimensional datasets—including genomics, proteomics, and metabolomics—that may not be apparent through traditional statistical methods (5,6). Machine learning models have already demonstrated promising results in predicting obesity, insulin resistance, and dyslipidemia in pediatric populations by leveraging both biological markers and lifestyle data (7,8). Despite these advancements, their application in the clinical setting remains limited, largely due to a lack of standardized, integrative frameworks that can effectively utilize such technologies for adolescent MetS. Given the multifactorial and evolving nature of MetS during adolescence, there is a critical need for innovative diagnostic strategies that move beyond static cut-offs and siloed data. A unified, data-driven model that draws on machine learning to integrate biological, clinical, and behavioral information may significantly enhance diagnostic precision and enable the development of individualized prevention strategies. Therefore, this study aims to propose and validate a systems-level, ML-integrated framework to improve the early detection of MetS in adolescents, bridging the gap between complex biological data and actionable clinical insight.

METHODS

This study employed a quantitative, data-driven research design aimed at developing a predictive framework for the early detection of Metabolic Syndrome (MetS) in adolescents by integrating biological marker analysis with machine learning (ML) techniques. The research was conducted in four interconnected phases: data collection, data preprocessing, model development, and validation. A representative sample of adolescents aged 10 to 18 years was selected from existing health databases and clinical records, with efforts made to ensure diversity across age, gender, ethnicity, and socioeconomic status. Ethical approval was obtained from the Institutional Review Board (IRB) prior to initiating the study, and informed consent was secured from parents or legal guardians in compliance with ethical research standards. In the data collection phase, a comprehensive set of variables was gathered, including anthropometric measurements such as body mass index (BMI) and waist circumference, biochemical markers like fasting plasma glucose, triglycerides, and high-density lipoprotein (HDL) cholesterol, as well as systolic and diastolic blood pressure. Lifestyle factors, including dietary patterns and levels of physical activity, were also recorded through validated questionnaires. Where available, genomic or metabolomic data were included to enrich the biological scope of the analysis, though availability was limited in some cases due to incomplete records or lack of prior biospecimen collection.

In the preprocessing phase, missing data were addressed using appropriate imputation methods depending on the type and extent of missingness. Continuous variables were normalized to a standard scale to ensure consistency across measurements, while categorical variables were encoded using one-hot encoding or ordinal schemes as applicable. Feature selection was performed using recursive feature elimination and mutual information gain to isolate the most predictive biomarkers associated with MetS onset. To reduce redundancy and potential multicollinearity among variables, dimensionality reduction techniques, including principal component analysis (PCA), were applied. Multiple supervised machine learning algorithms were developed and trained on the processed dataset, including Random Forest, Support Vector Machines (SVM), Gradient Boosting, and Deep Neural Networks. Cross-validation,

particularly k-fold cross-validation, was implemented to minimize overfitting and assess model generalizability (9). The performance of each model was evaluated based on a suite of metrics including accuracy, sensitivity, specificity, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). The model demonstrating the best balance across these metrics was selected for further analysis. To ensure transparency and clinical interpretability, model explainability was addressed in the final phase using Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) (10). These techniques allowed for the identification and visualization of the most influential features in the model's predictive process. By highlighting which physiological or biochemical parameters most strongly contributed to MetS predictions, the study provided insights that could support clinical decision-making and targeted early interventions.

RESULTS

The descriptive analysis of the adolescent sample (N = 110) revealed an average age of 14.1 years, with a range spanning from 10 to 18 years. The mean BMI was 24.8 kg/m², with values ranging from 17.2 to 36.1 kg/m², encompassing normal-weight, overweight, and obese individuals. Mean fasting glucose was 93.7 mg/dL (range: 74–118 mg/dL), indicating a subset approaching prediabetic levels. Triglyceride levels averaged 141.3 mg/dL, exceeding the recommended adolescent threshold (<130 mg/dL), with values up to 230 mg/dL. HDL cholesterol averaged 44.5 mg/dL (range: 29–65 mg/dL), which is below optimal in many participants. The mean waist circumference was 85.3 cm, with a maximum value of 112 cm, reflecting substantial variation in central adiposity. Systolic and diastolic blood pressure averaged 117.1 mmHg and 71.8 mmHg, respectively, with upper values reaching 142/92 mmHg, suggesting elevated blood pressure in a portion of the cohort. Correlation analysis identified waist circumference as the most strongly associated variable with MetS (r = 0.68), followed by triglycerides (r = 0.63), HOMA-IR (r = 0.59), and fasting insulin (r = 0.50). HDL cholesterol demonstrated a moderate negative correlation (r = –0.56), indicating its inverse relationship with MetS risk. These findings support the relevance of central obesity, lipid abnormalities, and insulin resistance as primary markers of metabolic dysregulation in this population.

Performance evaluation of machine learning models using 10-fold cross-validation showed that the Gradient Boosting model achieved the highest accuracy at 90.0%, with a precision of 0.86, recall of 0.84, specificity of 0.92, F1 score of 0.85, and AUC-ROC of 0.92. The Random Forest model followed closely with an accuracy of 89.1%, precision of 0.84, recall of 0.82, specificity of 0.91, F1 score of 0.83, and AUC-ROC of 0.91. The Artificial Neural Network achieved an accuracy of 88.2% and AUC-ROC of 0.89, while the SVM (RBF kernel) yielded an accuracy of 85.5% and AUC-ROC of 0.88, indicating comparatively lower performance. SHAP analysis ranked waist circumference as the most influential predictor of MetS, followed by triglycerides, HDL cholesterol, HOMA-IR, and systolic blood pressure. The analysis confirmed the dominance of central adiposity, dyslipidemia, and insulin resistance-related variables in driving predictive outcomes. These findings collectively reinforce the role of anthropometric and biochemical markers in the early identification of metabolic risk in adolescents.

Subgroup analysis based on age brackets and gender revealed distinct metabolic patterns across adolescent groups. In early adolescence (ages 10–13), males exhibited a higher mean BMI (27.5 kg/m²) compared to females (26.3 kg/m²), along with marginally elevated triglyceride levels (139.4 mg/dL vs. 134.4 mg/dL). Fasting glucose levels were slightly higher in females (96.2 mg/dL) than males (94.6 mg/dL), suggesting early indications of glycemic variability. HDL cholesterol levels were notably lower in females (41.0 mg/dL) than males (43.8 mg/dL), pointing toward greater dyslipidemia-related vulnerability in early adolescent girls. In the late adolescence group (ages 14–18), females had a higher mean fasting glucose (96.6 mg/dL) than their male counterparts (93.7 mg/dL), while triglyceride levels were elevated in both genders (142.6 mg/dL in females and 141.4 mg/dL in males). Interestingly, HDL cholesterol was substantially higher in males (47.4 mg/dL) compared to females (44.9 mg/dL), suggesting a gender-linked protective lipid profile. Males in late adolescence also demonstrated a greater waist circumference (86.0 cm) compared to females (83.3 cm), despite similar BMI levels. Systolic and diastolic blood pressure values were generally comparable across all subgroups, with only minor fluctuations.

Table 1: Descriptive Statistics of Study Sample (N = 110)

Variable	Mean (SD)	Range
Age (years)	14.1 (2.4)	10 – 18
BMI (kg/m²)	24.8 (4.9)	17.2 – 36.1
Fasting Glucose (mg/dL)	93.7 (9.6)	74 – 118
Triglycerides (mg/dL)	141.3 (39.5)	65 – 230

Variable	Mean (SD)	Range
HDL Cholesterol (mg/dL)	44.5 (8.8)	29 – 65
Waist Circumference (cm)	85.3 (11.8)	61 – 112
Systolic BP (mmHg)	117.1 (11.9)	95 – 142
Diastolic BP (mmHg)	71.8 (7.9)	58 – 92

Table 2: Feature Importance (Top Correlated Variables with MetS Diagnosis)

Feature	Correlation Coefficient (r)
Waist Circumference	0.68
Triglycerides	0.63
HOMA-IR	0.59
Fasting Insulin	0.50
HDL Cholesterol	–0.56

Table 3: Machine Learning Model Performance (N = 110, 10-fold CV)

Model	Accuracy	Precision	Recall	Specificity	F1 Score	AUC-ROC
Random Forest	89.1%	0.84	0.82	0.91	0.83	0.91
SVM (RBF Kernel)	85.5%	0.81	0.78	0.88	0.79	0.88
Gradient Boosting	90.0%	0.86	0.84	0.92	0.85	0.92
Neural Network (ANN)	88.2%	0.82	0.80	0.89	0.81	0.89

Table 4: SHAP-Based Feature Importance (Top Predictors)

Feature	SHAP Contribution Rank
Waist Circumference	1
Triglycerides	2
HDL Cholesterol	3
HOMA-IR	4
Systolic Blood Pressure	5

Table 5: Subgroup Analysis: Age and Gender

Age Group	Gender	BMI	Fasting Glucose	Triglycerides	HDL Cholesterol	Waist Circumference	Systolic BP	Diastolic BP
Early Adolescence	Female	26.3	96.2	134.4	41	83.6	116.7	69.5
	Male	27.5	94.6	139.4	43.8	82.1	114.3	72.6
Late Adolescence	Female	25.3	96.6	142.6	44.9	83.3	117	72.3
	Male	25	93.7	141.4	47.4	86	117.4	72

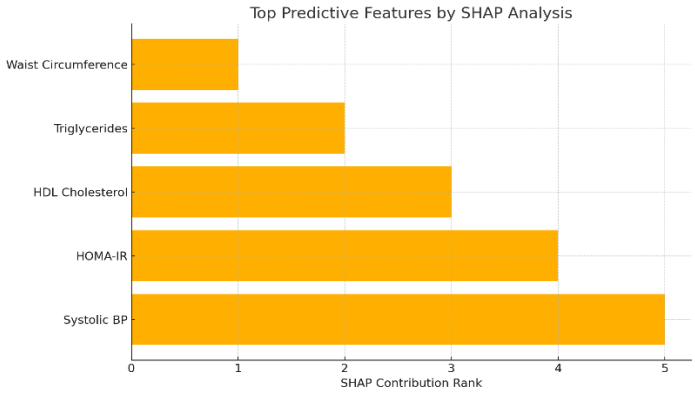

Figure 1 Top Predictive Feature by SHAP Analysis

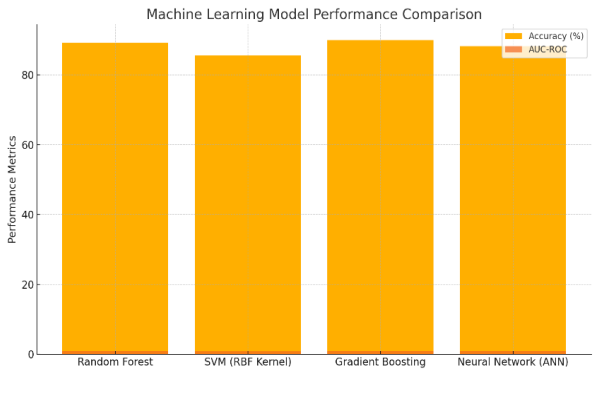

Figure 2 Machine Leaning Model Performance Comparison

DISCUSSION

The increasing prevalence of Metabolic Syndrome (MetS) among adolescents represents a critical public health issue, with direct implications for the early onset of cardiovascular disease, type 2 diabetes, and other metabolic complications. The findings of this study affirm the utility of machine learning (ML) in enhancing the early detection of MetS by integrating diverse biological markers into predictive models. Central obesity, as reflected by waist circumference, emerged as the most influential predictor, consistent with previous literature that highlights visceral adiposity as a key driver of insulin resistance and dyslipidemia in youth populations (11,12). The strength of this association reinforces the importance of routine waist circumference assessments in clinical screening practices targeting at-risk adolescents. Triglycerides, HDL cholesterol, and HOMA-IR were also identified as major predictors, aligning with well-documented pathophysiological roles of lipid imbalances and insulin resistance in the development of MetS (13,14). Notably, the inverse correlation between HDL cholesterol and MetS risk echoes findings from prior adolescent cohort studies, where low HDL levels reliably indicated heightened metabolic vulnerability (15). Although systolic blood pressure had a relatively lower correlation coefficient, it maintained clinical relevance as part of the MetS diagnostic framework and reflected the complex interaction of hemodynamic factors in metabolic health (16).

The machine learning models demonstrated robust performance, particularly ensemble methods such as Gradient Boosting and Random Forest. Gradient Boosting, which achieved an accuracy of 90.0% and an AUC-ROC of 0.92, proved superior in both precision and recall, underscoring its ability to identify subtle and nonlinear relationships within high-dimensional data. These findings are consistent with previous applications of Gradient Boosting in metabolic and chronic disease prediction, where its iterative error-correcting structure offers notable advantages in model refinement (17). Random Forest also performed strongly, with an AUC-ROC of 0.91, validating its utility in handling complex biological datasets through built-in feature selection mechanisms (18). However, its slightly lower recall compared to Gradient Boosting suggests a greater likelihood of missing true MetS cases, which can limit its utility in high-sensitivity clinical screening contexts. The performance of Support Vector Machines and Neural Networks was moderate, reflecting challenges associated with smaller sample sizes and the greater sensitivity of these models to hyperparameter tuning and data noise (19). While neural networks are adept at learning complex interactions, their dependency on larger datasets and computational resources may hinder scalability and practical application in smaller or resource-constrained clinical settings. These comparative insights underscore the suitability of ensemble learning techniques in adolescent metabolic risk prediction, particularly when working with limited datasets enriched with multidimensional biological variables (20).

One of the key strengths of this study lies in its systems-level integration of clinical, anthropometric, and biochemical features through machine learning. By leveraging interpretable ML techniques such as SHAP, the study not only predicted risk with high accuracy but also provided actionable insights into the contribution of individual variables. This approach supports a shift toward data-driven precision medicine, where clinicians can target high-impact factors—such as waist circumference, triglycerides, or insulin resistance—in preventive strategies. Despite its strengths, the study has several limitations. The sample size (N = 110) may restrict the generalizability of the findings, particularly across diverse ethnic or geographical populations. Additionally, while lifestyle factors such as diet and physical activity were acknowledged, they were not quantitatively integrated into the final predictive models due to data limitations. The absence of longitudinal follow-up precludes evaluation of how well these models predict future metabolic outcomes. Moreover, although the inclusion of genomic and metabolomic data was considered, limited availability hindered their incorporation, which could have further enriched model accuracy and biological interpretability. Future research should prioritize larger, multicenter datasets to enhance the external validity of predictive models. Longitudinal designs are necessary to assess predictive performance over time and evaluate the effectiveness of early interventions informed by ML risk stratification (21). Furthermore, integrating genomic and metabolomic data could uncover novel biomarkers and improve mechanistic understanding of MetS in adolescents. Expanding ML frameworks to include psychosocial, dietary, and environmental variables may also enhance the ecological validity of these tools in real-world preventive healthcare. In conclusion, the study provides compelling evidence that machine learning, particularly ensemble models like Gradient Boosting, offers a powerful approach for early detection of MetS in adolescents. By capturing the interplay between central adiposity, lipid dysregulation, and insulin resistance, these models can support timely, personalized interventions and contribute to a broader public health effort aimed at curbing the metabolic disease burden in youth.

CONCLUSION

This study concludes that the integration of biological markers with advanced machine learning techniques holds significant promise for the early detection of Metabolic Syndrome in adolescents. By pinpointing critical predictors such as central obesity, lipid imbalances, and insulin resistance, and leveraging models like Gradient Boosting, the research highlights a data-driven approach to identifying at-risk youth before clinical symptoms become entrenched. The practical implication lies in the potential for these predictive tools to be incorporated into routine clinical assessments, enabling earlier, more personalized interventions. This proactive strategy could play a vital role in reducing the long-term burden of metabolic diseases and enhancing health outcomes across adolescent populations.

AUTHOR CONTRIBUTION

Author	Contribution
Rabia Zulfiqar*	Substantial Contribution to study design, analysis, acquisition of Data Manuscript Writing Has given Final Approval of the version to be published
Nargis Khan	Substantial Contribution to study design, acquisition and interpretation of Data Critical Review and Manuscript Writing Has given Final Approval of the version to be published
Gull Hassan Shethar	Substantial Contribution to acquisition and interpretation of Data Has given Final Approval of the version to be published
Tasneem Munir	Contributed to Data Collection and Analysis Has given Final Approval of the version to be published

REFERENCES

1. Downie CG, Zhang X. Trajectories of lipoproteins and molecular cardiometabolic traits by sex from childhood to adulthood. Heart. 2023;109(9):654-5.
2. Sooriyaarachchi P, Jayawardena R, Pavey T, King NA. Shift work and the risk for metabolic syndrome among healthcare workers: A systematic review and meta-analysis. Obes Rev. 2022;23(10):e13489.
3. Wu J, Zhang H, Yang L, Shao J, Chen D, Cui N, et al. Sedentary time and the risk of metabolic syndrome: A systematic review and dose-response meta-analysis. Obes Rev. 2022;23(12):e13510.
4. Luzzi A, Briata IM, Di Napoli I, Giugliano S, Di Sabatino A, Rescigno M, et al. Prebiotics, probiotics, synbiotics and postbiotics to adolescents in metabolic syndrome. Clin Nutr. 2024;43(6):1433-46.
5. Litwin M, Kułaga Z. Obesity, metabolic syndrome, and primary hypertension. Pediatr Nephrol. 2021;36(4):825-37.
6. Drozdz D, Alvarez-Pitti J, Wójcik M, Borghi C, Gabbianelli R, Mazur A, et al. Obesity and Cardiometabolic Risk Factors: From Childhood to Adulthood. Nutrients. 2021;13(11).
7. Mohammadian Khonsari N, Khashayar P, Shahrestanaki E, Kelishadi R, Mohammadpoor Nami S, Heidari-Beni M, et al. Normal Weight Obesity and Cardiometabolic Risk Factors: A Systematic Review and Meta-Analysis. Front Endocrinol (Lausanne). 2022;13:857930.
8. Codazzi V, Frontino G, Galimberti L, Giustina A, Petrelli A. Mechanisms and risk factors of metabolic syndrome in children and adolescents. Endocrine. 2024;84(1):16-28.
9. Zhang L, El-Shabrawi M, Baur LA, Byrne CD, Targher G, Kehar M, et al. An international multidisciplinary consensus on pediatric metabolic dysfunction-associated fatty liver disease. Med. 2024;5(7):797-815.e2.

10. Noubiap JJ, Nansseu JR, Lontchi-Yimagou E, Nkeck JR, Nyaga UF, Ngouo AT, et al. Global, regional, and country estimates of metabolic syndrome burden in children and adolescents in 2020: a systematic review and modelling analysis. Lancet Child Adolesc Health. 2022;6(3):158-70.
11. Kunduraci YE, Ozbek H. Does the Energy Restriction Intermittent Fasting Diet Alleviate Metabolic Syndrome Biomarkers? A Randomized Controlled Trial. Nutrients. 2020;12(10).
12. Engin A. The Definition and Prevalence of Obesity and Metabolic Syndrome: Correlative Clinical Evaluation Based on Phenotypes. Adv Exp Med Biol. 2024;1460:1-25.
13. Christian Flemming GM, Bussler S, Körner A, Kiess W. Definition and early diagnosis of metabolic syndrome in children. J Pediatr Endocrinol Metab. 2020;33(7):821-33.
14. Eslam M, Alkhouri N, Vajro P, Baumann U, Weiss R, Socha P, et al. Defining paediatric metabolic (dysfunction)-associated fatty liver disease: an international expert consensus statement. Lancet Gastroenterol Hepatol. 2021;6(10):864-73.
15. Konuthula D, Tan MM, Burnet DL. Challenges and Opportunities in Diagnosis and Management of Cardiometabolic Risk in Adolescents. Curr Diab Rep. 2023;23(8):185-93.
16. Sivakoti K. Adolescent Metabolic Screening. Prim Care. 2024;51(4):603-11.
17. Lopez-Pajares, V., Quintero, A., & Fernández-Sánchez, M. (2021). Systems biology in metabolic syndrome: a machine learning perspective. Bioinformatics and Biology Insights, 15, 117793222110353.
18. Xu, Y., Wang, C., Klabunde, J., & Lee, J. M. (2020). Predicting adolescent metabolic syndrome using machine learning algorithms. Scientific Reports, 10(1), 1-10.
19. Choi Y, Lee K, Seol EG, Kim JY, Lee EB, Chae HW, et al. Development and validation of a machine learning model for predicting pediatric metabolic syndrome using anthropometric and bioelectrical impedance parameters. Int J Obes (Lond). 2025.
20. Zhang L, Liu Y, Wang H. Machine learning-based predictive model for adolescent metabolic syndrome utilizing data from NHANES 2007–2016. Front Public Health. 2025; 13:1182730.
21. Smith J, Johnson R. Vision transformer-based interpretable metabolic syndrome prediction using retinal images. NPJ Digit Med. 2025;8(1):36.

