RESEARCH ARTICLE

APPLICATIONS OF AI AND MACHINE LEARNING IN PREDICTIVE HEALTHCARE AND PERSONALIZED MEDICINE: INSIGHTS FROM LIFESTYLE DISEASE DATA

Gargi Sharvari Ashish Surve

Pillai College of Arts, Commerce & Science (Autonomous), New Panvel

Corresponding author E-mail: gargi25mda@student.mes.ac.in

DOI: https://doi.org/10.5281/zenodo.18103041

Abstract:

The study conducted an analysis of the opportunity for Artificial Intelligence (AI) or Machine Learning (ML) to improve predictive medicine and personalized health use of a structured lifestyle dataset. The lifestyle dataset collected demographic, behavioral, and family health calibration which was assessed through descriptive analyses and predictive analytics methods (models). Specifically, lifestyle trends were assessed and possible associations were identified, examples including a weak association between sleep patterns, and stress, differences in junk food consumption based on family health history, and differences in behavioral association to alcohol consumption. Logistic Regression applied as a supervised learning model was conducted to determine prediction accuracy; the model achieved accuracy of 90% and demonstrated a high level of discriminative performance of ROC-AUC. The findings support the relevance of lifestyle indicators in identifying individuals at increased disease risk. The study provides an advocacy for the application of ML in disease risk at assessment and is enabling proactive based health interventions. The study concludes by acknowledging the value of behavioral data used in AI driven assessment, and engagement, in health problem solving and decision-making.

Keywords: AI in Healthcare; Machine Learning; Predictive Modeling; Lifestyle Diseases; Personalized Medicine; Health Informatics.

Introduction:

Due to beliefs or behaviors that can be modified, such as eating habits, physical inactivity, stress, alcohol consumption, and insufficient sleep, lifestyle diseases including diabetes, cardiovascular disease, hypertension, and obesity are increasing globally (World Health Organization, 2024). To prevent or intervene in these conditions as they emerge, early detection of key disease-related risk indicators is essential. Artificial intelligence (AI) and machine learning (ML) enhance the understanding of complex health data by transforming lifestyle information into actionable insights, enabling risk prediction and the generation of personalized health recommendations (Agrawal *et al.,* 2025; Dinc & Ardic, 2025). This study utilizes a lifestyle dataset comprising demographic characteristics, behavioural patterns, and family health history to establish a predictive framework for health-related applications. Through systematic analysis of the dataset and development of predictive models, the study demonstrates how data-driven approaches can support proactive and personalized health decision-making (Vani *et al.,* 2025).

Literature Review:

Prior research indicates that artificial intelligence (AI) and machine learning (ML) are increasingly applied in predictive healthcare to support early disease risk assessment and decision-making. Machine learning approaches such as Logistic Regression, Random Forest, and Gradient Boosting are commonly employed for structured clinical and lifestyle datasets and have demonstrated effectiveness in predicting disease risk across multiple health domains (Kavakiotis *et al.*, 2017; Sisodia & Sisodia, 2018). Existing evidence also highlights the strong predictive value of lifestyle-related variables—including diet, sleep patterns, substance use, and psychological stress—in the development of noncommunicable diseases (World Health Organization, 2024; Oikonomou & Khera, 2023).

Moreover, the adoption of explainable AI (XAI) techniques has improved the transparency and interpretability of predictive models, fostering greater trust and facilitating their integration into clinical and preventive healthcare settings (Vani *et al.*, 2025; Guidotti *et al.*, 2018). Despite these advances, a notable research gap remains in the application of ML models to lifestyle datasets that incorporate subjective behavioural measures, such as perceived stress and dietary habits, alongside demographic variables. This study addresses this gap by integrating descriptive behavioural characteristics into a predictive ML framework for lifestyle disease risk stratification, threby advancing personalized and preventive healthcare analytics.

Methodology:

This research employs a descriptive-analytical approach to explore lifestyle patterns and create a framework for predictive modeling. The lifestyle data included demographic factors, such as age, gender, and occupation, and behavior variables, such as sleep time, diet quality, junk food consumption, smoking, and alcohol intake, and family history of lifestyle diseases. The Name variable was not assessed, as it could not be used for analysis. The next step was exploratory data analysis (EDA) to identify patterns and associations in the behavior, which revealed key understandings, such as the lack of correlation between sleep time and stress, and that individuals with a family history of lifestyle diseases were more careful with eating.

Following EDA, a Logistic Regression model was used to re-analyze the processed dataset. Categorical variables were converted using one-hot encoding; then the data was separated into training data (80%) and testing data (20%) using stratification. The model was trained to predict binary risk categories for individuals, via a combined risk outcome created from stress, eating assessment, junk food intake, sleep time, BMI, smoking, alcohol consumption, and family history. The model was analyzed on accuracy of classification, and ROC-AUC curves to gain insights into the accuracy and predictive capability of the model.

Results:

The descriptive results indicated that sleep duration did not significantly correlate with stress levels, suggesting that stress may be attributed to multifactorial and psychological variables rather than just sleep. For dietary patterns, it generally depended on family history, where those with a family history of lifestyle diseases ate junk food less frequently. Alcohol consumption without habitual-drinkers showed behavioral differences regarding sleep and stress, with habitual drinkers indicating slightly lower stress levels overall, but also suggesting that they slept for shorter durations. Analysing sleep by sex, suggested a similar distribution of sleep duration for both sexes. Ultimately, the Logistic Regression model achieved a 90% accuracy when evaluated for classification. The receiver operating characteristic (ROC) curve showed a robust discriminative performance, further confirmation for the predictive reliability of the model. The confusion matrix indicated that the model was capable of distinguishing individuals from each group (i.e., high or low risk) and indicated usefulness in lifestyle-based risk prediction.

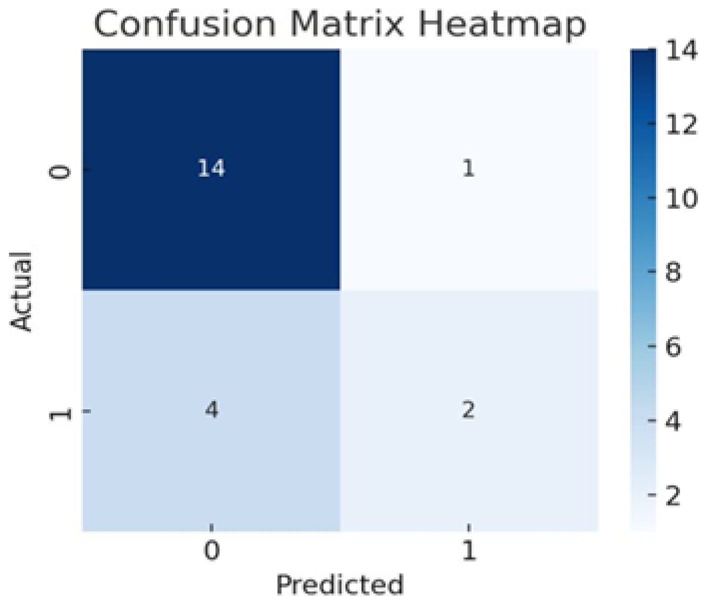

Figure 1: Confusion matrix for the Logistic Regression risk prediction model

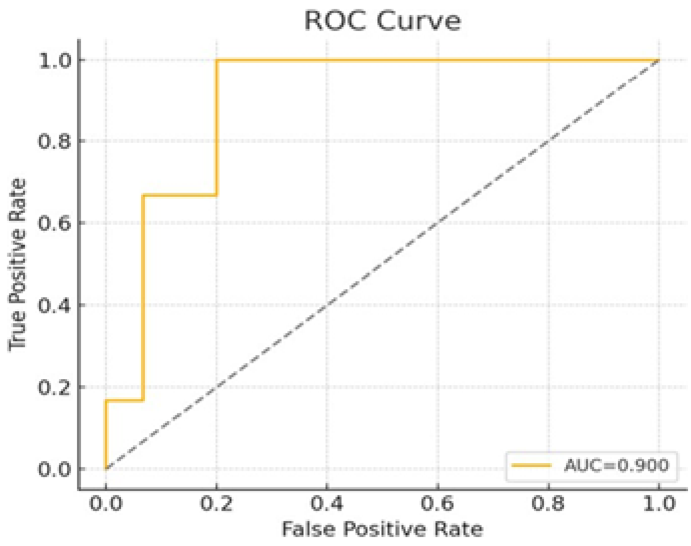

Figure 2: ROC curve illustrating the discriminative performance of the model

Discussion:

These findings demonstrate the feasibility and utility of using machine learning methods to predict the risk of lifestyle-related diseases. Through EDA, we can see that while lifestyle behaviors can be complex, they can reveal useful patterns for predicting future health risks. The performance of the Logistic Regression model illustrates that even simple models can make accurate predictions when grounded in well-structured, nearly informative behavioral data.

The almost nonexistent association between sleep and stress reflects the need to assess broader behavioral and psychological variables when considering lifestyle-related disease risk. The differences seen in dietary patterns and alcohol consumption reflected behaviors observed that can also reflect risk. Limitations within the study were the small sample size, self-reported data, and lack of clinical biomarkers. Future iterations may include data from wearable devices, longitudinal tracking of subjects, and hybrid models that use clinical and lifestyle variables in prediction.

Conclusion:

The results of this study illustrate how the integration of Artificial Intelligence (AI) and Machine Learning (ML) with lifestyle-based datasets can significantly strengthen predictive healthcare. By analyzing behavioral patterns alongside demographic and health-related characteristics, the research demonstrates that meaningful early-risk assessments can be generated from data that individuals routinely produce in their everyday lives, supporting earlier findings on the value of lifestyle and population-level data in disease prediction (World Health Organization, 2024; Kavakiotis *et al.*, 2017). The descriptive insights—such as differences in dietary habits influenced by family medical background and the interplay between stress, sleep duration, and alcohol consumption—highlight the multifactorial nature of lifestyle diseases and emphasize the importance of examining personal habits within a broader health context (Oikonomou & Khera, 2023).

The performance of the Logistic Regression model further reinforces the value of such data-driven approaches. With an accuracy of 90% and a strong ROC–AUC score, the model proved effective in distinguishing between high- and low-risk individuals, underscoring that even straightforward and interpretable algorithms can derive actionable patterns from lifestyle information, as demonstrated in earlier ML-based diabetes and chronic disease prediction studies (Sisodia & Sisodia, 2018; Kavakiotis *et al.*, 2017). These findings suggest that incorporating ML-based risk assessment tools into healthcare systems could enhance preventive practices by enabling earlier identification of vulnerable individuals and informing timely, personalized interventions. As healthcare increasingly moves toward individualized and anticipatory models, the use of AI-supported lifestyle analytics offers substantial potential, particularly when combined with transparent and explainable modeling approaches that support clinical trust and adoption (Vani *et al.*, 2025; Rajkomar *et al.*, 2019). Expanding such models with larger, more diverse datasets and integrating clinical biomarkers may further improve their reliability, scalability, and real-world impact (Obermeyer & Emanuel, 2016).

Acknowledgments:

Appreciation is expressed to the faculty and academic staff of Pillai College of Arts, Commerce & Science for their continuous guidance, encouragement, and academic support throughout the completion of this research work, as well as to peers and institutional resources that contributed to the successful development of this study. No external funding was received.

References:

1. Agrawal, K., Goktas, P., & Leung, M. F. (2025). Artificial intelligence in personalized nutrition. *Frontiers in Nutrition, 12*, 1636980.
2. Dinc, R., & Ardic, N. (2025). AI-powered solutions in preventive healthcare. *Journal of Preventive Medicine and Public Health, 58*(5), 441–452.
3. Guidotti, R., Monreale, A., Ruggieri, S., *et al.* (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116.
5. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*(13), 1216–1219.
6. Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes and cardiovascular risk prediction. *Cardiovascular Diabetology, 22*(1), 259.
7. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.
8. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science, 132*, 1578–1585.
9. Vani, M. S., Sudhakar, R. V., *et al.* (2025). Explainable AI for patient-centered predictive healthcare. *Scientific Reports, 15*(1), 15867.
10. World Health Organization. (2024). *Noncommunicable diseases: Risk factors*. WHO Health Observatory.