Machine Learning-Based Disease Prediction Using Lifestyle and Health Data

Ms. Rutuja Nage[2], Ms. Renuka Kalikar[3], Ms. Maitreyee Patil[4], Miss. Manjiri Ghate[5], Dr. A.G. Kadu[1],
[2,3,4,5]*Student, Prof. Ram Meghe Institute of Technology & Research, Badnera.*
[1] *Professor, Prof. Ram Meghe Institute of Technology & Research, Badnera.*

Abstract *- The rising prevalence of lifestyle-related illnesses has encouraged much research attention on the artificial intelligence-based methods of early risk evaluation and individualized healthcare. Lately, there has been a prospective research on the incorporation of machine learning, generative artificial intelligence and recommender systems to examine lifestyle tendencies, forecast diabetic probability, and prescribe preventive measures. The given review is a systematic examination of available literature on the topic of lifestyle-based disease prediction models, with a specific focus on tree-based ensemble models like the Light Gradient Boosting Machine (LightGBM) that have become popular expectation frameworks when working with structured health data. More so, the paper will analyze the future potential of generative AI models to provide individualized preventive advice on nutrition, physical exercise, and behavior change. Symptom-aware doctor recommendation strategies such as keyword-based disease mapping, locationand rating-aware ranking strategies, are also revised. Combining the current research findings, this paper will compare the most applied datasets, algorithms, and system design options and define the main challenges associated with data reliability, interpretability, preservation of privacy, and clinical applicability. The review indicates open research opportunities and gives insights to assist in the creation of effective, ethical, user-friendlier AIdriven preventive health.*

Keywords: *prediction of diseases, generative artificial intelligence, health informatics, machine learning.*

I. INTRODUCTION

Dietary habits, exercise, sleep habits, stress exposure, and drug use are all lifestyle choices that have been well-accepted as leading to the development and deterioration of chronic diseases. Modifiable behavioral determinants are now a part of noncommunicable conditions such as cardiovascular disease, diabetes, hypertension, and obesity which are underlined by the need to identify risks early and prevent them in contemporary healthcare systems. The conventional methods of healthcare that mostly involve reactive care once the symptoms have been observed have drawbacks when considering the increasing number of such disorders worldwide. Consequently, the research interest in the data-driven and preventative healthcare paradigms has increased. The development of artificial intelligence (AI) and machine learning (ML) has made it possible to analyze large-scale lifestyle and health data sets to identify nonlinear and complex relationships that are hard to capture with traditional statistical approaches. Specifically, the supervised learning methods that use structured health records have shown promising prospects in the ability to predict the risk of diseases using lifestyle characteristics. Of these methods, the gradient-boosting-based models have received a lot of interest owing to their excellent results with tabular data, their capability to handle missing values, and their capability to capture feature interactions. Light Gradient Boosting Machine (LightGBM), specifically, has found extensive use in recent studies in the healthcare context due to its compute efficiency, scalability and applicability to high-dimensional lifestyle data. In addition to predicting the risk of disease, recent studies have examined the utilization of generative AI models in helping to prevent personalized medical care. These models are being used more frequently to produce personalized advice in terms of diet, exercise and behavior change and are intended to enhance interest and compliance rates with healthy lifestyles. Parallel to this, intelligent doctor recommendation systems have also appeared as a complementary line of research and has been based on the analysis of symptoms, disease mapping and other contextual factors like location and expertise of

clinicians to enable efficient patient-provider matching. This review paper is an intellectual synthesis of the recent literature in these interrelated research domains, as the lifestyle-based models of predicting diseases, AI-based preventive recommendations, and physician recommendation mechanisms. Instead of introducing a particular implementation, the paper evaluates the popular algorithms, datasets, and system design strategies according to the existing literature. Important issues in the fields of data quality, interpretability, privacy, and real-world applicability are touched upon and open research directions are outlined as a step to creating useful, ethical, and user-friendly AI-based preventive health models.

II. RELATED WORK

The recent development of research on the topic has been emphasizing the concept of merging machine learning, lifestyle analytics, and AI-driven recommendation systems to assist in preventive healthcare. Research like [1], [3] and [4] indicate that the advanced gradient-boosting models, including LightGBM, are effective in predicting cardiovascular and hypertension-related risks, based on lifestyle and medical factors. These papers indicate that the boosted decision trees have the ability to do better in working with heterogeneous health data than the traditional statistical methods. Wider studies on predicting chronic diseases with the help of machine learning have been introduced in [5], [6], and [7], where several algorithms were tested on very different data sets. These studies point out the significance of considering behavioral, dietary and demographic factors in order to maximize the predictive power. In addition, open source studies like [8], [9], and [10] present the evidence that Light GBM is one of the most competitive models to risk in heart-disease modeling because of its efficiency and high-dimensionality capabilities. Simultaneously, a number of works study lifestyle-based or multi-factor disease analytics. The importance of daily habits and long-term health trends in the development of chronic diseases is emphasized in such publications as [2], [11], and [12]. More recent research such as [13], [14], and [15] discusses advanced control models, IoT-based health monitoring, and predictive models which pay attention to stress, biomarkers, and genetic inputs, which are also enriching the insights into the relationship between lifestyle and health. When applied to preventive healthcare, the use of generative AI and its potential to serve individualized health advice, generate clinical notes, and offer preventive-care advice have been examined in [16], [17], and [19] and found to be supported by large language models (LLMs). Studies like [18] extend these functionalities to the automated public-health kiosks, which demonstrates how AI-based systems can be used to offer real-time recommendations. Lastly, doctor recommendation systems have developed based on hybrid model, symptom-disease mapping and personalized matching algorithms. As it is summed up in [20], healthcare recommendation systems are becoming more and more modern and involve both ML and specifics of the users like their location, rating, and chronic illnesses to make more precise physician recommendations. All in all, the current body of literature proves that combination of machine learning-enhanced disease prediction, preventive advice provided by generative AI, and systemic doctor recommendation is well-grounded. Nevertheless, unified platforms that combine all the three elements to improve the accessibility, personalization, and preventive care outcomes are still required- an area that this review will focus on.

III. RESEARCH GAP ANALYSIS

Despite the notable advances in AI-assisted healthcare systems and machine learning-based disease prediction, there are definitely a number of gaps in the current research. To start with, although numerous studies ([1], [3], [4], [8], [9]) provide high accuracy in predicting individual diseases, including cardiovascular or kidney disorders, most models are not generalized in many different lifestyle-related diseases. No research literature exists on the integrated multi-disease prediction models that combine various lifestyle, behavioral, and clinical variables into a single model. Second, even though lifestyle parameters are cited as the main predictors, most datasets applied to the past research ([2], [5], [6]) do not include the complete representation of lifestyle, including, though not limited to, the quality of sleep, stress, or long-term behavioral patterns. This limits the practical use of ML models which need more comprehensive, longitudinal data to conduct more

reliable health risk forecasting. Third, AI has been studied in the context of personalized health advice ([16], [17], [19]) but most of the existing applications do not rely on machine-learning prediction observers. The gap in the research is the development of integrated mechanisms in which preventive recommendations based on the LLM are dynamically informed by the real assessments of the disease risks provided by the ML and the history of the patients. Fourth, the doctor recommendation systems are not developed as other components. Some studies like [20] use majorly the recommendation algorithms or symptom mapping. Nonetheless, there are limited literature sources on extensive end-to-end suggestions of doctors encompassing the extraction of symptoms, inference of illnesses, locality of users, hospital presence, doctor specialties, and services quality in a single ranking framework. Lastly, majority of the existing systems do not offer an integrated platform of lifestyle-based disease prediction, personalized preventative measures, and doctor recommend in one user-friendly ecosystem. Such a gap in integration is a significant research gap, which necessitates strong, scalable, AI-based healthcare systems, which would facilitate the management of diseases proactively and allow patients to access relevant medical services in time.

IV. CONCEPTUAL FRAMEWORK FOR LIFESTYLE-BASED PREVENTIVE HEALTHCARE SYSTEMS

According to recent literature, three key streams of preventive healthcare research have come together: the predictive risks of disease based on lifestyle, the preventive recommendation based on AI, and the intelligent doctor recommendation. These components are not viewed as a separate solution, but a number of studies indicate their combination to create a holistic, user-based preventive healthcare paradigm. Judging by the analysis of the literature conducted in Section II, there is a generalized conceptual model that can be identified that would describe the interaction of these components in contemporary AI-based healthcare systems.

A. Lifestyle/ Health Data Abstraction.

Current literature always points to the lifestyle and behavior as the inputs to preventive healthcare analytics. Researchers like [1], [2], [11], and [12] found that diet-related variables, physical activity, sleep schedules, stress, substance use, and demographic factors have a significant effect on the development and progression of chronic diseases. The literature focuses on how these heterogeneous inputs are abstracted in structured feature representations which can be efficiently processed by machine learning models. The problem of the partially noisy or unbalanced lifestyle data is also mentioned in several works, which proves the necessity of strong featurehandling methods.

B. Disease Risk Modelling with machine learning.

There is a large amount of literature concerning machine learning-based disease risk modeling, especially of lifestyle-related chronic diseases. The available comparative studies [3]-[10] show that ensemble based algorithms, particularly gradient boosting models are always better than traditional statistical methods in the prediction of cardiovascular disease, hypertension, diabetes and other related diseases. The capability of LightGBM in handling high-dimensional tabular data and capturing nonlinear relationships among lifestyle variables and retaining computational efficiency is a common feature listed in the literature. In the conceptual model, the analytic centre is the disease risk modeling which translates the lifestyle aspects into probabilistic or categorical risk factors of the disease.

C. Prevention Recommendation/Behavior guidance Stratum.

In addition to estimating risks, recent research highlights the need to transform the predictive information into the preventive guidance that can be acted upon. The work on lifestyle-based healthcare system [16], [17], and [19] demonstrates that using generative AI and knowledge-based systems could result in personalized recommendations related to dieting, physical activities, stress management, and sleep hygiene. The literature fits this layer as an interpretive interface between predictive analytics and the end user with a view to enhancing health literacy, behavioral adherence and risk mitigation over the long term. Noteworthy, some of the studies indicate that preventive recommendations tend to be the most effective when applied and scaled with the context of individual lifestyle profiles and disease risk outputs as opposed to generic guidelines.

D. Symptom-aware Physician Recommendation and Clinical Access.

Another important dimension of the reviewed literature is the doctor recommendation and clinical decision support systems. According to the survey based and application-oriented studies [18], [20], the current focus is developing on symptom-disease mapping, matching specialization, and mechanism of personalized ranking. Such systems usually take into account the variables of predicted types of diseases, symptoms that are reported by the users, proximity, availability, and indicators of quality of services. This element fits in the conceptual framework as a relay point between preventive analytics and formal healthcare services, which positively influence access to medical professionals.

E. Integrated Preventive Healthcare Ecosystem.

Taken together, the analyzed articles indicate that the optimal clinical and social effect is obtained when predicting the disease, its prevention, and matching of doctors are combined into a single conceptual ecosystem. These elements are not isolated, but instead, preventive care is a closed-loop, since data of lifestyle is used to determine the risk of disease; risk estimation is used to provide personalized preventive measures; and predictions and prevention are facilitated by intelligent clinical access controls. Some of the reviews [5], [6], and [20] point out that this type of integration enhances early intervention, personalization, and healthcare accessibility and poses challenges of data privacy, model transparency, ethical AI implementation, and user trust.

F. Open Challenges and Design Considerations.

The literature also mentions cross-cutting issues that impact every level of the conceptual model. These are the requirement of high-quality longitudinal lifestyle data [2], interpretable and explainable machine learning applications [4], the ethical nature of generative AI in health advice and counseling [17], and fairness and consistency in the recommendation systems [20]. Consequently, these issues need to be met so that conceptual frameworks may be transformed into clinically meaningful, as well as socially responsible preventive healthcare solutions.

V. COMPARATIVE ANALYSIS OF EXISTING STUDIES.

[1]	Cardiovascular disease	LightGBM, Random Forest, Logistic Regression	LightGBM	LightGBM Achieved great accuracy through successfully modeling nonlinear lifestyle interactions.
[3]	Hypertension	SVM, KNN, Random Forest, XGBoost	XGBoost / LightGBM	LightGBM Boosted models performed better than the traditional classifiers.
[4]	Coronary heart disease	LightGBM variants	Optimized LightGBM	Hyperparameter tuning LightGBM was found to detect quite successfully at an earlier stage.
[7]	Cardiovascular disease	LR, DT, RF, XGBoost	XGBoost	XGBoost Ensemble models were more effective at dealing with the heterogeneity of lifestyles.
[8]	Heart disease	KNN, SVM, RF, LightGBM	LightGBM	LightGBM It showed high generalization when using lifestyle data.
[9]	Heart disease	LightGBM, CatBoost, RF	LightGBM	LightGBM Improved leaf-wise growth predictively.
[10]	Heart disease	GBDT, RF	GBDT	Gradient boosting was successful in capturing complicated interactions among features.

Table 1: Comparison of Machine Learning Algorithms to Lifestyle-Based Disease prediction .

Ref.	Dataset Source	Sample Type	Lifestyle Features Included	Limitations Identified
[1]	Public health surveys	Structured tabular	Diet, activity, smoking, BMI	BMI Limited stress and sleep measures.
[2]	Nutrition & CKD datasets	Longitudinal	Diet patterns, physical activity	Low demographic mix.
[5]	Multiple public datasets	Mixed	Behavioral and clinical features	Absence of a single dataset.
[6]	Cardiometabolic datasets	Clinical + lifestyle	Activity, BMI, glucose levels	Long-term behavior trends are missing.
[11]	Lifestyle survey data	Self-reported	Diet, exercise, alcohol use	Potential reporting bias.
[12]	Health & lifestyle records	Structured	Lifestyle + clinical history	Cancer-specific focus
[15]	Multimodal datasets	Lifestyle + biomarkers	Social, genetic, health behaviors	Multi-high data complexity.

Table 2: Comparison of Datasets and Lifestyle Features Used in Reviewed Studies.

Ref.	System Type	Techniques Used	Personalization Factors	Identified Gaps
[16]	Preventive health prediction	ML-based risk modeling	Life style attributes	Limited explain ability
[17]	Personalized treatment	Generative AI (LLMs)	Patient history, symptoms	Poor integration with ML predictors.
[18]	Public health kiosk	AI-driven automation	Real-time health inputs	Dependency infrastructure.
[19]	Clinical documentation	Generative AI	Clinical notes, patient context	Not lifestyle-centered.
[20]	Doctor recommendation	Hybrid ML models	Location, specialty, ratings	Poor disease-risk integration.

Table 3: Comparison of Preventive Recommendation and Doctor Recommendation Approaches.

VI. RESULTS AND DISCUSSION

This section critically reviews the conclusions made in the recent literature on the topics of the prediction of the disease carried by lifestyle-based methods, the preventive recommendation system, and the intelligent doctor recommendation framework. Instead of presenting experimental findings based on a particular implementation, the discussion is the synthesis of the findings, trends, and insights based on comparing the available studies.

A. Lifestyle-Based Disease Prediction Performance Trends.

In assessing the reviewed literature, machine learningbased solutions are always shown to be better predictive than traditional statistical tools when used

on lifestyle and health data. Gradient-boosting models sometimes under the term of ensemble learning are commonly cited as the best-performing algorithms when it comes to chronic disease risk prediction. According to the studies summarized in Table I, Light Gradient Boosting Machine (LightGBM) and similar boosting algorithms are useful to learn nonlinear interactions between the lifestyle factors (diet, physical activity, smoking status, body mass index). According to the literature, LightGBM leaf-wise growth strategy and effective feature management is advantageous to better accuracy and convergence rate, particularly in balancing mixed data type and missing values tabular healthcare data. The reported performance improvements are however usually done in disease-specific settings with most models being trained and tested on one disease. This is a weakness of the existing literature as there are not many studies that examine multi-disease prediction in a unified model.

B. Effect of the Choice of Lifestyle Features and Design of the Dataset.

The datasets reviewed differ greatly in scope, composition of features, and the quality of the data as in Table II. Research involving a more comprehensive set of lifestyle characteristics, such as behavioral, dietary, and demographic variables, tends to show better predictive statistics. However, a large number of datasets do not provide deep coverage of such essential dimensions of lifestyle as the quality of sleep, stress, and long-term behavior patterns. Also, the use of selfreported data may create bias and restrict the extrapolation of results. These findings highlight the importance of standardized, comprehensive and longitudinal datasets of lifestyles to improve the accuracy of disease prediction models

C. Preventive Recommendation System and Generative AI.

Recently, the generative AI has been used more broadly to replace risk prediction with preventive healthcare in the context of generative AI. As observed in Table III, it has been shown that large language models can provide a personalized health recommendation, which deals with diet, exercise, as well as lifestyle change, using user-specific context. Although such systems demonstrate a potential increase in personalization and user interactions, the majority of them do not rely on machine-learning based risk prediction models. Consequently, preventive guidelines are usually not dynamically matched to the quantified risk of disease, which constrains their clinical applicability. The convergence of generative AI and predictive analytics is a research problem that is yet to be tackled through a variety of studies.

D. Clinical Decision Support and Doctor Recommendation.

Doctor recommendation systems are a relatively new but rather under-researched field of AI-based healthcare. The current literature is mostly devoted to matching or general healthcare recommendation frameworks based on symptoms. As it is seen in Table III,there are few systems that combine the disease-risk estimates, lifestyle context and healthcare accessibility factors in a single ranking mechanism. This discontinuity diminishes the efficiency of doctor recommendation systems as components of a preventive care pipeline and indicates a need to develop more patient-focused and holistic recommendation approaches.

E. Towards Incorporated Preventive Healthcare Frameworks.

Altogether, the reviewed literature shows that there is a definite trend toward individualized AI-based preventive care. Although each of the three elements, namely the prediction of diseases, the recommendation of prevention, and the matching of doctors, have fully developed, their connection to each other as a unified ecosystem has not been established. The results imply that the concept of developing unified frameworks linking the lifestyle analytics, predictive modelling, personalised guidance and clinical access should be the primary focus of future investigations. Overcoming the issues on data interoperability, explainability, privacy, and ethical implementation will be necessary to transform these research breakthroughs into practical healthcare outcomes.

VII. CONCLUSION

This review article has studied the current developments in the domain of disease prediction

based on lifestyle, AI-assisted preventive healthcare, and intelligent physician recommendation systems. The reviewed literature proves that machine learning algorithms, especially, gradient-boosting models, like Light Gradient Boosting Machine (LightGBM), can be effectively applied to simple lifestyle and health data to identify risks of chronic diseases. Research results have always shown that the inclusion of behavioral, dietary, and demographic characteristics is better predictive than traditional statistics. In addition to disease prediction, the emerging studies show the increasing tendency of generative AI to provide personalized preventive care. The large language models open up new possibilities to convert the projected health risks into the comprehensible and applicable dietary, physical exercise, and lifestyle change guidance. Nevertheless, the clinical relevance and personalization of preventive strategies are often limited by the existing systems that tend to separate the prediction and recommendation as two distinct entities. As also identified in the review, doctor recommendation systems are relatively undeveloped and still poorly integrated in the integration of disease risk, interpretation of symptoms, contextual factors like location and quality of service. On the whole, the literature suggests that the necessity to have integrated, AI-intelligent healthcare systems that combine lifestyle-based risk forecasting, individualized preventive recommendations, and smart clinical access is evident. Development of interoperable, explainable, and privacy conscious platforms that can facilitate proactive management of diseases and provision of personalized healthcare is what future research should work on.

REFERENCES

[1] S. A. Kissi, M. G. M. Talukder, and M. Z. Iqbal, "Data-driven predictive modelling of lifestyle risk factors for cardiovascular health," Electronics, vol. 14, no. 14, p. 2906, 2025.

[2] W. Luo et al., "Lifestyle and chronic kidney disease: A machine learning analysis," Frontiers in Nutrition, vol. 9, 2022.

[3] R. Sandhiya et al., "Enhanced hypertension disease prediction using machine learning," in Proc. ICOFE 2024, 2024.

[4] T. O. Omotehinwa et al., "Optimizing the Light Gradient-Boosting Machine algorithm for early detection of coronary heart disease," Heliyon, vol. 10, no. 1, 2024.

[5] R. Islam et al., "A comprehensive review for chronic disease prediction using machine learning," Journal of Electrical Systems and Information Technology, vol. 11, 2024.

[6] D. Musleh et al., "Machine learning approaches for predicting risk of cardiometabolic diseases," AI, vol. 8, no. 3, p. 31, 2024.

[7] J. Miah et al., "Improving cardiovascular disease prediction through comparative analysis of machine learning models," arXiv preprint, arXiv:2311.00517, 2023.

[8] R. S. Ramesh et al., "Cardiovascular disease prediction using machine learning," arXiv preprint, arXiv:2507.21898, 2025.

[9] A. Mahajan et al., "Heart disease risk assessment using LightGBM and other models," AIP Conference Proceedings, vol. 020107, 2024.

[10] Y. Wu et al., "Heart disease prediction using gradient boosting decision trees," in Proc. 2024 Int. Conf. Artificial Intelligence and Applications, 2024.

[11] S. Tuly et al., "Predicting heart disease risk with lifestyle factors," in Proc. ACM Int. Conf., 2024.

[12] M. A. Ahmed et al., "Predicting cancer risk using machine learning on lifestyle and health data," BMC Medical Informatics and Decision Making, 2025.

[13] S. H. A. Faruqui et al., "Model predictive control continuous-time Bayesian network for selfmanagement of multiple chronic conditions," arXiv preprint, arXiv:2205.13639, 2022.

[14] S. Tiwari and S. Agarwal, "An optimized hybrid solution for IoT-based lifestyle disease classification using stress data," arXiv preprint, arXiv:2204.03573, 2022.

[15] M. Green, "Evaluating the performance of personal, social, health-related, biomarker, and genetic data for predicting an individual's future health," arXiv preprint, arXiv:2104.12516, 2021.

[16] A. Hameed et al., "Machine learning for lifestylecentered health prediction," ResearchGate Preprint, 2025.

[17] S. Garg, “The role of generative AI in personalized medicine and treatment recommendations,” Clinical Medicine and Health Research Journal, vol. 5, 2025.
[18] S. Falahati et al., “An AI-powered public health automated kiosk system for personalized care (HERMES Kiosk),” arXiv preprint, arXiv:2504.13880, 2025.
[19] A. Biswas and W. Talukdar, “Intelligent clinical documentation: Harnessing generative AI for patient-centric clinical note generation,” arXiv preprint, arXiv:2405.18346, 2024.
[20] Y. Zhu et al., “A survey on personalized healthcare recommendation systems,” IEEE Access, vol. 12, pp. 1–20, 2024.
[21] E. Nasarian, D. Sharifrazi, S. Mohsenirad, K. Tsui, and R. Alizadehsani, "AI Framework for Early Diagnosis of Coronary Artery Disease," arXiv preprint, Aug. 29, 2023. Available: https://arxiv.org/abs/2308.15339
[22] S. Subramani, N. Varshney, and M. V. Anand, "Cardiovascular Diseases Prediction by Machine Learning Incorporation with Deep Learning," Frontiers in Medicine, vol. 10, Apr. 16, 2023, Article 1150933. Available: https://doi.org/10.3389/fmed.2023.1150933
[23] S. Dongre, R. Chandra, and S. Agarwal, "MLtoGAI: Semantic Web with Machine Learning for Enhanced Disease Prediction and Personalized Recommendations," arXiv preprint, Jul. 2024. Available: https://arxiv.org/abs/2407.20284