University of Qom

Information Technology and Sport

Online ISSN: 3092-6939

Journal Homepage: https://its.qom.ac.ir



Assessing cancer risk based on physical activity and lifestyle indicators: a data-driven approach in sports science

Yaser Khanjari [1], Mohammad Alami [2], Soheil Bateni[3]

1. Corresponding author,Assistant Professor, Department of Physical Education, K. N. Toosi University of Technology, Tehran, Iran. Email: ykhanjari@kntu.ac.ir
2. Department of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran. m.alami1383@gmail.com
3. Department of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran. batnys6@gmail.com

Article Info

Article type:
Research Article

Article history:
Received 10 July 2026
Received in revised form 02 November 2026
Accepted 02 November 2026
Available online 02 November 2026

Keywords:
Body Mass Index,
machine learning,
clinical decision support system,
binary classification,
gradient boosting.

ABSTRACT

Objective: The objective of this study is to design a machine learning-based clinical decision support system to predict the likelihood of cancer based on physiological characteristics and lifestyle factors.

Methodology: This study utilized a public "cancer prediction" dataset containing 1,500 records. Six distinct algorithms—Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, Random Forest, and Gradient Boosting—were implemented and evaluated using a 5-fold cross-validation approach integrated into the processing pipeline. The dataset, which is simulated, consists of 8 independent input features and 1 binary target variable; it was evaluated using an 80/20 stratified split (1,200 training samples and 300 independent test samples).

Results: The results indicate that ensemble learning algorithms, particularly Gradient Boosting, significantly outperform classical models. On the independent test dataset (n=300), the Gradient Boosting model demonstrated a statistically significant superiority over the baseline model according to the DeLong test ($p < 0.001$ compared to the Logistic Regression baseline). Furthermore, sensitivity analysis revealed that removing the variable "personal history of cancer" leads to a substantial increase in the relative importance of Body Mass Index (18.9%) and weekly physical activity (14.2%).

Conclusion: Overall, this study presents a diagnostic model for estimating the likelihood of cancer based on physical activity and lifestyle indicators. This model can facilitate targeted screening and data-driven decision-making within the fields of healthcare and sports science. Nevertheless, its clinical application is contingent upon external validation using real-world, longitudinal data.

Cite this article: Khanjari, Y., Alami, M., & Bateni, S. (2026). Assessing Cancer Risk Based on Physical Activity and Lifestyle Indicators: A Data-Driven Approach in Sports Science. *Information Technology and Sport*, *3*(1), 225-251. https://doi.org/10.22091/its.2026.16698.1055

© The Author(s) retain the copyright.

Publisher: University of Qom Press.

DOI: https://doi.org/10.22091/its.2026.16698.1055

Introduction

Cancer is a major public health challenge worldwide, and early identification of individuals with a higher likelihood of cancer is important for improving health-related decision-making. Cancer development is associated with interactions among genetic, environmental, physiological, and lifestyle factors. Among these factors, smoking, alcohol consumption, excess body weight, and physical inactivity are particularly relevant because some can be modified through lifestyle interventions. Physical activity and body composition are therefore important areas in which sports science can contribute to health promotion and cancer prevention. Previous research has demonstrated associations between physical activity and the risk of several cancer types. Nevertheless, the relationships between lifestyle indicators, physiological characteristics, and cancer status are complex and may involve nonlinear interactions. Conventional statistical approaches may not fully capture these relationships. Machine learning methods offer an alternative approach by identifying patterns across multiple variables and estimating the probability of a particular health outcome.

Recent studies have applied machine learning algorithms to cancer risk assessment using lifestyle, genetic, and clinical information. Although promising predictive performance has been reported, several methodological limitations remain, including limited external validation, insufficient interpretation of sports-related variables, and inadequate assessment of potential information leakage. In particular, the contribution of physical activity and body mass index (BMI) to cancer prediction requires further examination alongside genetic and other individual risk factors. The present study aimed to develop and evaluate a machine learning-based framework for classifying cancer status using lifestyle and physiological indicators, with particular attention to physical activity and BMI. Six classification algorithms were compared, and feature importance and sensitivity analyses were performed to examine the contribution of sports-related variables and the potential influence of personal cancer history on model performance.

Methods

This study used a public cancer prediction dataset obtained from Kaggle, consisting of 1,500 records. The dataset was simulated and included eight independent input variables and one binary outcome indicating cancer status (cancer or no cancer). The input variables were age, sex, BMI, smoking status, genetic risk level, weekly physical activity, alcohol consumption, and personal history of cancer. Weekly physical activity was recorded in hours per week, while BMI was expressed in kg/m^2. The dataset was divided into training and independent testing subsets using a stratified 80:20 split, resulting in 1,200 training records and 300 testing records. Stratification preserved the distribution of the outcome classes in both subsets. No missing values were reported in the dataset. Feature standardization was performed within a processing pipeline to prevent information leakage during model training and cross-validation. Class weighting was also applied where supported to address the imbalance between cancer and non-cancer records. Six machine learning algorithms were implemented: Logistic Regression, Support Vector Machine with a radial basis function kernel, K-Nearest Neighbors, Gaussian Naive Bayes, Random Forest, and Gradient Boosting. Hyper-parameters were optimized using grid search and stratified five-fold cross-validation on the training data. The independent test dataset was not used during preprocessing, hyper-parameter selection, or model selection and was reserved for final evaluation.

Model performance was assessed using accuracy, precision, sensitivity (recall), F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Additional analyses included decision-threshold evaluation, calibration assessment using the Brier score and calibration slope and intercept, and estimation of 95% confidence intervals using 1,000 stratified bootstrap samples. Feature importance and model interpretation were examined using feature-importance measures and SHAP values. Because the dataset did not specify the timing of personal cancer history relative to the recorded cancer outcome, a sensitivity analysis was conducted. In this analysis, personal history of cancer was removed, and all six algorithms were retrained and evaluated using the remaining seven input variables. This procedure was used to examine the extent to which model performance depended on this potentially problematic variable. All analyses were conducted in Python using the Scikit-learn and SHAP libraries on a standard computer without a graphics processing unit.

Results

Of the 1,500 records, 943 (62.87%) were classified as having no cancer and 557 (37.13%) as having cancer. The overall mean age was 50.32 ± 17.64 years, and the mean BMI was 27.51 ± 7.23 kg/m². Mean weekly physical activity was 4.90 ± 2.87 hours. Compared with the non-cancer group, the cancer group had a higher mean BMI (29.28 ± 6.65 versus 26.47 ± 7.36 kg/m²) and lower mean weekly physical activity (4.34 ± 2.97 versus 5.23 ± 2.75 hours). Among the six evaluated algorithms, Gradient Boosting demonstrated the best overall performance during five-fold cross-validation on the training dataset. Its mean accuracy was 94.4% ± 0.8%, sensitivity was 89.5% ± 1.2%, precision was 94.8% ± 0.9%, F1-score was 92.0% ± 1.0%, and ROC-AUC was 97.1% ± 0.5%. Random Forest achieved the next-highest ROC-AUC, at 96.4% ± 0.6%.

On the independent test dataset, Gradient Boosting achieved an accuracy of 93.3%, sensitivity of 88.8%, precision of 93.6%, specificity of 96.2%, F1-score of 91.1%, and ROC-AUC of 95.8%. The confusion matrix included 177 true-negative, 103 true-positive, seven false-positive, and 13 false-negative predictions. The 95% confidence interval for test ROC-AUC was 93.6%–97.7%. The Brier score was 0.054, with a calibration slope of 0.97 and an intercept of −0.02. Feature-importance analysis indicated that personal history of cancer (21.4%) and genetic risk (17.4%) were the most influential variables in the complete model. BMI accounted for 13.3% of feature importance, followed by alcohol consumption (11.2%), sex (10.3%), weekly physical activity (9.6%), age (9.6%), and smoking (7.1%). SHAP analysis also indicated that higher weekly physical activity was associated with lower model-predicted cancer probability, whereas higher BMI and smoking contributed to higher predicted probability. The sensitivity analysis showed that removing personal history of cancer reduced the Gradient Boosting ROC-AUC to 91.4% ± 0.9% during cross-validation and 89.7% on the independent test dataset. Test accuracy decreased to 87.3%. Importantly, after removing this variable, the relative importance of BMI increased to 18.9%, while weekly physical activity increased to 14.2%. Together, these two sports-related indicators accounted for 33.1% of the model's total feature importance.

Discussion

The findings indicate that machine learning algorithms, particularly Gradient Boosting, can identify patterns associated with cancer status using a combination of lifestyle, physiological,

genetic, and personal-history variables in the examined dataset. The model achieved a test ROC-AUC of 95.8%, while the sensitivity analysis demonstrated that its performance decreased after removing personal history of cancer. This reduction highlights the importance of examining potentially influential variables rather than relying solely on high overall predictive performance. From a sports science perspective, the findings are particularly relevant to BMI and weekly physical activity. In the complete model, these variables contributed to cancer-status classification, and their combined relative importance increased to 33.1% after personal history of cancer was excluded. The lower average physical activity and higher BMI observed in the cancer group are consistent with the predictive patterns identified by the model. However, these findings should be interpreted as statistical associations within the dataset, not as evidence of causal effects. In particular, because the data are cross-sectional and simulated, lower physical activity among individuals classified as having cancer may also reflect reverse causation.

An important methodological issue concerns personal history of cancer. Since the dataset does not provide the timing of this variable relative to the recorded outcome, information leakage cannot be completely excluded. The decrease in model performance after removing this variable demonstrates its substantial influence on classification. Therefore, the performance of the complete model should be interpreted cautiously. This study has several limitations. The dataset was simulated and relatively small, and external validation using real-world populations was not available. Furthermore, more detailed sports-related measures, including physical fitness, aerobic capacity, exercise intensity, sedentary behavior, and dietary patterns, were not included. The findings therefore represent proof of concept rather than evidence of clinical readiness.

Conclusion

In conclusion, the proposed machine learning framework demonstrated promising performance in classifying cancer status in the examined dataset. BMI and weekly physical activity were relevant predictive indicators, particularly when personal history of cancer was excluded. These results highlight the potential contribution of sports-related variables to data-driven health assessment. Nevertheless, the model should not be considered a clinically validated cancer screening tool. Further research using larger, real-world, longitudinal, and independently validated datasets is required to establish its generalizability and practical value in healthcare and sports science.

Authors' Contributions

All authors contributed equally to all stages of conducting this research project.

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

The authors gratefully acknowledge the Iran National Science Foundation (INSF) and the Research Vice-Chancellery of K. N. Toosi University of Technology for their financial and institutional support in conducting this research.

فناوری اطلاعات و ورزش

شاپا الکترونیکی: ۶۹۳۹-۳۰۹۲

Journal Homepage: https://its.qom.ac.ir



تشخیص احتمال ابتلا به سرطان بر اساس شاخص‌های فعالیت بدنی و سبک زندگی: یک رویکرد داده محور در علوم ورزشی

یاسر خنجری [۱]، محمد عالمی [۲]، سهیل باطنی [۳]

۱. نویسنده مسئول، استادیار گروه تربیت بدنی، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران. رایانامه: ykhanjari@kntu.ac.ir

۲. گروه مهندسی کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران. m.alami1383@gmail.com

۳. گروه مهندسی کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران. batnys6@gmail.com

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخچه مقاله:
تاریخ دریافت: ۱۴۰۵/۰۴/۱۹
تاریخ بازنگری: ۱۴۰۵/۰۸/۱۱
تاریخ پذیرش: ۱۴۰۵/۰۸/۱۱
تاریخ انتشار: ۱۴۰۵/۰۸/۱۱

کلیدواژه‌ها:
شاخص تودهٔ بدنی، یادگیری ماشین، سامانه تصمیم‌یار بالینی، طبقه‌بندی دودویی، گرادیان بوستینگ.

چکیده

هدف: هدف این پژوهش، طراحی یک سامانه تصمیم‌یار بالینی مبتنی بر یادگیری ماشین برای تشخیص احتمال ابتلا به سرطان بر اساس ویژگی‌های فیزیولوژیک و سبک زندگی است.

روش پژوهش: در این مطالعه، از مجموعه‌داده عمومی «پیش‌بینی سرطان» شامل ۱۵۰۰ رکورد استفاده شد. شش الگوریتم متفاوت شامل رگرسیون لجستیک، ماشین بردار پشتیبان، همسایه نزدیک، بیز ساده، جنگل تصادفی و گرادیان بوستینگ پیاده‌سازی و با روش اعتبارسنجی متقابل ۵-بخشی درون پایپ‌لاین مورد ارزیابی قرار گرفتند. این مجموعه‌داده از نوع شبیه‌سازی‌شده و شامل ۸ ویژگی ورودی مستقل و ۱ متغیر هدف دودویی است که با تقسیم‌بندی طبقه‌بندی‌شده ۲۰/۸۰ (۱۲۰۰ نمونه آموزش و ۳۰۰ نمونه آزمون مستقل)ارزیابی شده است.

یافته‌ها: نتایج نشان داد الگوریتم‌های مبتنی بر یادگیری گروهی، به‌ویژه گرادیان بوستینگ، برتری قابل توجهی نسبت به مدل‌های کلاسیک دارند. در مجموعه داده های تست مستقل ($n = ۳۰۰$) نیز ، مدل گرادیان بوستینگ برتری معنادار نسبت به مدل پایه در آزمون دی‌لانگ داشت (مقدار$p < ۰/۰۰۱$ در مقایسه با مدل پایه رگرسیون لجستیک). همچنین، تحلیل حساسیت نشان داد با حذف متغیر سابقه شخصی سرطان، سهم شاخص توده بدنی (۱۸.۹ درصد) و فعالیت بدنی هفتگی (۱۴.۲ درصد) به‌طور قابل‌توجهی افزایش می یابد.

نتیجه گیری: در مجموع، پژوهش حاضر یک مدل تشخیصی برای احتمال ابتلا به سرطان بر اساس شاخص های فعالیت بدنی و سبک زندگی ارائه می‌کند که می‌تواند به غربالگری هدفمند و تصمیم‌گیری مبتنی بر داده در حوزه سلامت و علوم ورزشی کمک کند. با این وجود، کاربرد بالینی آن منوط به اعتبارسنجی خارجی روی داده‌های واقعی و طولی است.

استناد: خنجری، یاسر، عالمی، محمد، & باطنی، سهیل. (۱۴۰۵). تشخیص احتمال ابتلا به سرطان بر اساس شاخص‌های فعالیت بدنی و سبک زندگی: یک رویکرد داده محور در علوم ورزشی. *فناوری اطلاعات و ورزش*، ۳ (۱)، ۲۲۵-۲۵۱. https://doi.org/ 10.22091/its.2026.16698.1055

ناشر: انتشارات دانشگاه قم.

© نویسندگان.

مقدمه

سرطان به گروهی از بیماری‌ها گفته می‌شود که در آن‌ها سلول‌های غیرطبیعی به‌صورت کنترل‌نشده رشد می‌کنند و می‌توانند به بافت‌های مجاور تهاجم کرده یا به اندام‌های دوردست گسترش یابند. این بیماری یکی از چالش‌های اصلی سلامت عمومی در جهان است و تشخیص و مداخله زودهنگام می‌تواند پیامدهای انسانی و اقتصادی آن را کاهش دهد(Brown et al., ۲۰۲۳; McGuire, ۲۰۱۶). فرایند بروز و پیشرفت سرطان معمولاً حاصل تعامل عوامل ژنتیکی و محیطی است. در کنار استعدادهای ژنتیکی، عوامل سبک زندگی مانند مصرف دخانیات، مصرف الکل، اضافه‌وزن و کم‌تحرکی می‌توانند خطر ابتلا را افزایش دهند. از منظر پزشکی بالینی، ارزیابی ریسک فردی و طبقه‌بندی افراد در سطوح مختلف خطر، امکان غربالگری هدفمندتر و تصمیم‌گیری دقیق‌تر را فراهم می‌سازد(DeVita, Lawrence, & Rosenberg, ۲۰۰۸; Kumar, Abbas, Fausto, & Aster, ۲۰۱۴; Steyerberg, ۲۰۱۹). در این میان، عوامل سبک زندگی قابل اصلاح، به‌ویژه فعالیت بدنی و ترکیب بدنی، حوزه‌ای را تشکیل می‌دهند که علوم ورزشی می‌تواند در پیشگیری و پایش آن نقش مستقیمی ایفا کند. با توجه به اینکه اثر عوامل سبک زندگی در خلأ رخ نمی‌دهد و با ویژگی‌های فیزیولوژیک، جمعیت‌شناختی و زمینه‌های ژنتیکی و بالینی افراد در تعامل است، بررسی دقیق جایگاه علوم ورزشی نیازمند تحلیل هم‌زمان این شاخص‌های قابل اصلاح در کنار سایر عوامل خطر فردی است.

در علوم ورزشی و تربیت‌بدنی، شاخص‌هایی مانند فعالیت بدنی، ترکیب بدنی و سایر مؤلفه‌های سبک زندگی، هم برای پیشگیری و هم برای پایش سلامت اهمیت دارند. گزارش کالج پزشکی ورزشی آمریکا بر نقش فعالیت بدنی و کاهش رفتار کم‌تحرکی در پیشگیری و کنترل سرطان تأکید می‌کند و شواهد مطالعات جمعیتی بزرگ‌مقیاس نیز رابطه فعالیت بدنی بیشتر با کاهش خطر بروز برخی انواع سرطان را نشان داده‌اند(Burtscher, Millet, & Burtscher, ۲۰۲۳; Moore et al., ۲۰۱۶; Patel et al., ۲۰۱۹). با این وجود، رابطه این عوامل با سرطان معمولاً غیرخطی و چندعاملی است و روش‌های آماری سنتی همواره قادر به بازنمایی کامل این تعاملات نیستند؛ از همین رو، رویکردهای داده‌محور مورد توجه قرار گرفته‌اند. این پیچیدگی و تعامل چندجانبه میان شاخص‌های فعالیت بدنی، شاخص توده بدنی و عوامل خطر ژنتیکی و محیطی، ضرورت بهره‌گیری از الگوریتم‌های پیشرفته یادگیری ماشین را برای استخراج الگوهای پنهان در داده‌های سلامت دوچندان می‌سازد.

یادگیری ماشین[1] به‌عنوان شاخه‌ای از هوش مصنوعی[2]، با یادگیری الگوها از داده‌ها می‌تواند برای پیش‌بینی پیامدهای سلامت (مانند ابتلا/عدم ابتلا) به‌کار رود و در قالب «مدل‌های پیش‌بینی بالینی» به تصمیم‌گیری کمک کند. مزیت مهم برخی الگوریتم‌ها، توانایی مدل‌سازی روابط پیچیده و نیز امکان محاسبه شاخص‌های عملکردی مانند مساحت زیر ROC-AUC است(Alpaydin, ۲۰۲۰; Bishop & Nasrabadi, ۲۰۰۶; Hastie, ۲۰۰۵). افزون بر قدرت تفکیک‌پذیری، در مدل‌های تصمیم‌یار بالینی تمایز نهادن میان «مدل‌های پیش‌بینی تشخیصی» (طبقه‌بندی وضعیت ابتلای فعلی در داده‌های مقطعی) و «مدل‌های پیش‌آگهی» (پیش‌بینی بروز بیماری در آینده بر پایه داده‌های طولی)، همراه با ارزیابی کالیبراسیون احتمالات، تعیین آستانه تصمیم و جلوگیری از نشت اطلاعات، از الزامات بنیادین روش‌شناختی محسوب می‌شود (Steyerberg, ۲۰۱۹)؛ رویکردی که زمینه‌ساز مطالعات تجربی متعددی در سال‌های اخیر شده است.

1. Machin Learning (ML)
2. Artificial Intelligence (AI)

بر همین اساس، به‌کارگیری یادگیری ماشین در انفورماتیک پزشکی، به‌ویژه برای شناسایی به‌موقع افراد مبتلا یا در معرض خطر سرطان، رو به گسترش است(Brown et al., ۲۰۲۳; McGuire, ۲۰۱۶). در سال‌های اخیر، مطالعات متعددی از یادگیری ماشین برای پیش‌بینی ریسک سرطان استفاده کرده‌اند. به عنوان نمونه، وو[1] و همکاران (۲۰۲۴) یک مدل پیش‌بینی بروز پن‌سرطان[2] را بر اساس داده‌های سبک زندگی و عوامل خطر توسعه دادند(Wu et al., ۲۰۲۴). در این مطالعهٔ آینده‌نگر با بیش از ۴۳۳ هزار شرکت‌کننده، مدل ایکس‌جی‌بوست به سطح زیر منحنی ۰.۸۷۶ در مردان و ۰.۷۴۶ در زنان دست یافت و افراد گروه پرخطر دست‌کم ۹ (۹) برابر خطر بیشتری برای بروز سرطان داشتند؛ با این حال، مدل صرفاً به‌صورت داخلی و در یک جمعیت واحد اعتبارسنجی شد و نویسندگان خود بر ضرورت اعتبارسنجی خارجی پیش از کاربرد بالینی تأکید کرده‌اند. همچنین احمد[3] و همکاران (۲۰۲۵) با استفاده از داده‌های سبک زندگی و خطر ژنتیکی، عملکرد بالای مدل‌های مبتنی بر گرادیان بوستینگ[4] را گزارش کرده‌اند(Ahmed, AbdelMoety, & Soliman, ۲۰۲۵). در این مطالعه که از همان مجموعه‌داده عمومی کگل استفاده کرده است، مدل کت‌بوست با دقت ۹۸.۷۵٪ روی مجموعه آزمون بهترین عملکرد را داشت؛ اما اتکا به یک مجموعه‌داده کوچک و شبیه‌سازی‌شده و نبود اعتبارسنجی خارجی، تعمیم‌پذیری این نتایج را محدود می‌کند. جی دوو[5] و همکاران (۲۰۲۵) نیز برای پیش‌بینی ریسک سرطان کولورکتال از مدل لایت‌جی‌بی‌ام[6] همراه با تفسیرپذیری مبتنی بر شاپ[7] استفاده کرده‌اند(Du et al., ۲۰۲۵). در این مطالعه، مدل در اعتبارسنجی خارجی به سطح زیر منحنی ۰.۹۴۳ رسید و متغیر سن و نشانگر CA۱۹-۹[9] مهم‌ترین متغیرهای مدل بودند؛ با این حال، داده‌ها تنها از دو بیمارستان در یک منطقه جغرافیایی گردآوری شده و مدل عمدتاً بر نشانگرهای آزمایشگاهی تکیه دارد و متغیرهای سبک زندگی و فعالیت بدنی در آن لحاظ نشده‌اند. بررسی دقیق این مطالعات نشان می‌دهد که علی‌رغم دستاوردهای کمّی قابل توجه، کاستی‌های روش‌شناختی مشترکی در ادبیات این حوزه وجود دارد که در مطالعات مروری نظام‌مند نیز مورد نقد قرار گرفته‌اند.

با وجود پیشرفت‌های یادگیری ماشین، چالش‌هایی مانند تعمیم‌پذیری مدل‌ها به جمعیت‌های متفاوت، توازن‌نداشتن داده‌ها، و محدودیت منابع محاسباتی در بسیاری از محیط‌های کاربردی همچنان وجود دارد. برخی از تحقیقات مروری و دامنه‌دار همچون الفایز[8] و همکاران (۲۰۲۱) نیز نشان می‌دهند که بخشی از مطالعات، از نظر اعتبارسنجی خارجی و گزارش‌دهی شفاف، نیازمند بهبود هستند. از سوی دیگر، تمرکز ویژه بر متغیرهای مرتبط با علوم ورزشی (فعالیت بدنی و ترکیب بدنی) در مدل‌های پیش‌بینی ریسک سرطان، در بسیاری از پژوهش‌ها کمتر دیده می‌شود(Abdullah Alfayez, Kunz, & Grace Lai, ۲۰۲۱; Felici, Peduzzi, Pellungrini, & Campa, ۲۰۲۵). افزون بر این، عدم گزارش شاخص‌های کالیبراسیون، فواصل اطمینان، آزمون‌های آماری مقایسه مدل‌ها و بی‌توجهی به تحلیل حساسیت نسبت به متغیرهای مشکوک به نشت اطلاعات (مانند سابقه شخصی بیماری در داده‌های مقطعی)، شکاف‌های مشخصی را در ادبیات پژوهشی پدید آورده است.

1. Wu
2. Pan-Cancer
3. Ahmed
4. Gradient Boosting
5. Du G
6. Light Gradient Boosting Machin (LightGBM)
9. Shaply Addetive Explanations (SHAP)
10. Alfayez

به بیان دقیق‌تر، شکاف پژوهشی این مطالعه در سه نکته خلاصه می‌شود: (چهار محور شماره‌گذاری‌شده زیر)؛ (۱) اول اینکه بیشتر مدل‌های پیشین بر نشانگرهای بالینی و آزمایشگاهی یا متغیرهای عمومی جمعیتی تکیه دارند و سهم متغیرهای قابل اندازه‌گیری در علوم ورزشی (فعالیت بدنی و شاخص توده بدنی) را به‌طور مشخص تفسیر نکرده‌اند؛ (۲) دومین نکته این است که مطالعاتی که از داده‌های مشابه سبک زندگی استفاده کرده‌اند، عمدتاً بر بیشینه‌سازی دقت متمرکز بوده‌اند و تفسیر نقش این متغیرها و تفکیک شفاف عملکرد اعتبارسنجی متقابل از عملکرد مجموعه آزمون کمتر مورد توجه بوده است؛ و (۳) سوم چارچوب‌های کم‌هزینه و قابل اجرا بدون واحد پردازش گرافیکی برای محیط‌های کم‌منبع ورزشی و درمانگاهی است که کمتر بررسی شده‌اند. اهمیت این موضوع از این روست که فعالیت جسمی و ترکیب بدنی از معدود عوامل قابل اصلاح‌اند که متخصصان علوم ورزشی می‌توانند مستقیماً آن‌ها را پایش کنند. پژوهش حاضر با مقایسه شش الگوریتم در یک پایپ‌لاین مقاوم در برابر نشت داده، گزارش جداگانه عملکرد اعتبارسنجی متقابل و مجموعه آزمون مستقل، و تفسیر مدل با روش SHAP با تأکید بر نقش پیش‌بینی‌کننده (و نه علّی) متغیرهای علوم ورزشی، به این شکاف می‌پردازد. علاوه بر سه مورد فوق، شکاف چهارم (۴) این است که در مطالعات پیشین مبتنی بر این داده‌ها، تحلیل حساسیت بدون متغیر «سابقه شخصی سرطان» (برای سنجش احتمال نشت اطلاعات)، شاخص کالیبراسیون (امتیاز برایر)، تحلیل آستانه تصمیم، فواصل اطمینان ۹۵ درصد و آزمون آماری دی‌لانگ گزارش نشده است که پژوهش حاضر تمامی این موارد را پوشش می‌دهد.

بر این اساس، هدف پژوهش حاضر طراحی و ارزیابی یک مدل داده‌محور برای پیش‌بینی ابتلا به سرطان با تکیه بر شاخص‌های سبک زندگی و فعالیت بدنی است. در واقع، هدف دقیق این پژوهش حل یک مسئله طبقه بندی دودویی (۱/۰) برای تعین ریسک سرطان (ابتلا / عدم ابتلا) است. علاوه بر مقایسه چند الگوریتم رایج طبقه‌بندی، تلاش شده است نقش نسبی متغیرها در خروجی مدل تفسیر شود تا استفاده از یافته‌ها برای برنامه‌ریزی مداخلات سلامت‌محور در حوزه تربیت‌بدنی امکان‌پذیر باشد. جهت ایجاد هماهنگی کامل مفهومی میان عنوان، چکیده، مقدمه و روش‌شناسی، تصریح می‌شود که اولاً اگرچه عنوان و تمرکز کاربردی مقاله بر شاخص‌های فعالیت بدنی و سبک زندگی در علوم ورزشی است، اما از آنجا که وضعیت ابتلا به سرطان پدیده‌ای چندعاملی است، متغیرهای جمعیت‌شناختی (سن و جنسیت)، خطر ژنتیکی و سابقه شخصی سرطان نیز در مدل حضور دارند تا سهم مستقل و نسبی متغیرهای علوم ورزشی هم در حضور این عوامل غالب و هم پس از حذف متغیر سابقه شخصی سرطان (در قالب تحلیل حساسیت) سنجیده شود؛ ثانیاً، به دلیل مقطعی و شبیه‌سازی‌شده بودن مجموعه‌داده، مدل حاضر یک «مدل پیش‌بینی تشخیصی وضعیت فعلی ابتلا» در سطح «اثبات مفهوم (Proof-of-Concept)» است و ادعایی در زمینه پیش‌بینی طولی بروز خطر سرطان در آینده ندارد.

روش شناسی

هدف این پژوهش، طراحی یک سامانه تصمیم‌یار بالینی[1] مبتنی بر یادگیری ماشین برای تشخیص احتمال ابتلای فعلی به سرطان بر اساس ویژگی های فیزیولوژیک و سبک زندگی است (Steyerberg, ۲۰۱۹). در این مطالعه، از مجموعه‌داده های عمومی پایگاه کگل[2] داده های «پیش‌بینی سرطان» شامل ۱۵۰۰ رکورد استفاده شد(RE., ۲۰۲۵). شش الگوریتم متفاوت شامل

1. Decision Clinical Support System (CDSS)
2. kaggle

رگرسیون لجستیک[۱]، ماشین بردار پشتیبان[۲]، همسایه نزدیک[۳]، بیز ساده[۴]، جنگل تصادفی[۵] و گرادیان بوستینگ[۶] پیاده‌سازی و مورد ارزیابی قرار گرفتند(Breiman, ۲۰۰۱; Friedman, ۲۰۰۱). این پژوهش یک مطالعه مدل‌سازی و تحلیل ثانویه داده است که در آن با استفاده از تکنیک‌های یادگیری ماشین، وضعیت ابتلای فعلی به سرطان بر اساس ویژگی‌های سبک زندگی و شاخص‌های قابل اندازه‌گیری در علوم ورزشی پیش‌بینی می‌شود. به دلیل استفاده از داده عمومی و فاقد اطلاعات هویتی، نیاز به اخذ رضایت آگاهانه از افراد وجود نداشت.

داده‌های این مطالعه از مجموعه‌داده عمومی «پیش‌بینی سرطان» که در پایگاه کگل منتشر شده است دریافت شد. این مجموعه‌داده شامل ۱۵۰۰ رکورد (نمونه) و ۱۰ ویژگی ورودی است و برچسب هدف «تشخیص» به صورت دودویی (۰=عدم ابتلا، ۱=ابتلا) تعریف شده است. (تصحیح و شفاف‌سازی منطبق با جدول ۱: مجموعه‌داده اصلی دارای ۹ ستون شامل ۸ ویژگی ورودی مستقل و ۱ متغیر هدف دودویی است؛ در توصیف اولیه ستون شناسه رکورد نیز در شمارش لحاظ شده بود که پیش از مدل‌سازی حذف گردید و تمامی مدل‌ها دقیقاً بر پایه ۸ ویژگی ورودی جدول ۱ آموزش دیدند.) بر اساس توضیحات منبع داده، متغیر «تشخیص» بیانگر وجود یا عدم وجود سرطان (ابتلای فعلی) در زمان ثبت داده است و مجموعه‌داده فاقد اطلاعات زمانی، تاریخ ثبت یا دوره پیگیری است. از این رو، مدل‌های توسعه‌یافته در این پژوهش در چارچوب «مدل‌های پیش‌بینی تشخیصی» قرار می‌گیرند که در آن‌ها بر اساس ویژگی‌های ثبت‌شده در یک مقطع زمانی، احتمال ابتلای فعلی فرد پیش‌بینی می‌شود؛ در مقابل، «مدل‌های پیش‌بینی پیش‌آگهی» خطر بروز بیماری را در یک افق زمانی آینده برآورد می‌کنند و به داده‌های طولی نیاز دارند (Steyerberg, ۲۰۱۹). بنابراین، منظور از «پیش‌بینی ابتلا» در این مقاله، پیش‌بینی تشخیصی وضعیت ابتلا است و برآورد خطر بروز سرطان در آینده در دامنه این پژوهش قرار ندارد. از نظر منشأ، این مجموعه‌داده توسط ربیع الخروعه[۷] در پایگاه کگل (نسخه ۲، به‌روزرسانی ژوئن ۲۰۲۴) با مجوز CC BY ۴.۰ منتشر شده و از نشانی (https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset_) قابل دسترسی است. طبق توضیح صریح منبع، داده‌ها شبیه‌سازی‌شده (synthetic) هستند و برای اهداف آموزشی تولید شده‌اند؛ از این رو، به جمعیت واقعی مشخصی تعلق ندارند و روش بالینی تعیین تشخیص سرطان (مانند تأیید آسیب‌شناختی یا ثبت در سامانه ثبت سرطان) نیز در منبع گزارش نشده است. به همین دلیل، نتایج این مطالعه باید در سطح اثبات مفهوم تفسیر شود و تعمیم آن به جمعیت واقعی نیازمند اعتبارسنجی روی داده‌های بالینی است. لینک و مشخصات کامل مجموعه‌داده در بخش منابع ارائه شده است(RE., ۲۰۲۵). جدول ۱ متغیرهای مورد استفاده، واحد/کدگذاری و دامنه آن‌ها را نشان می‌دهد. این متغیرها شامل شاخص‌های رایج در ارزیابی سلامت و تندرستی (مانند شاخص توده بدنی و فعالیت بدنی) به‌همراه عوامل خطر مهم (مانند مصرف دخانیات، الکل و ریسک ژنتیکی) هستند.

3. Logistic Regression
4. Machine Vector Support (SVM)
5. K-Nearest Neighbors (KNN)
6. Naïve Bayes
7. Random forest
8. Gradient Boosting
7 . Rabie El Kharoua

جدول ۱. تعریف متغیرها و کدگذاری آن‌ها در مجموعه‌داده (۸ ویژگی ورودی مستقل و ۱ متغیر هدف دودویی)

متغیر	واحد/کدگذاری	نوع و دامنه
سن	سال	پیوسته؛ (۲۰ تا ۸۰ سال)
جنس	کدگذاری دودویی (۱/۰) (۰ = مرد، ۱ = زن)	کیفی؛ ۰ یا ۱
شاخص توده بدنی (شاخص توده بدنی) (BMI)	کیلوگرم بر مترمربع	پیوسته؛ (۱۵تا ۹۶/۳۹)
مصرف دخانیات	۰=خیر، ۱=بله	کیفی؛ ۰ یا ۱
سطح خطر ژنتیکی	۰=کم، ۱=متوسط، ۲=زیاد	رتبه‌ای؛ ۰ تا ۲
فعالیت بدنی هفتگی	نمایه فعالیت (مثلاً ساعت/هفته) (تعداد ساعات فعالیت بدنی در هفته بر حسب ساعت/هفته طبق مستندات منبع داده)	پیوسته؛ (۰۰/۰ تا ۹۹/۹ ساعت/هفته)
مصرف الکل	نمایه مصرف (مثلاً واحد/هفته) (میزان مصرف الکل بر حسب واحد استاندارد در هفته طبق مستندات منبع داده)	پیوسته؛ (۰۰/۰ تا ۹۹/۴ واحد/هفته)
سابقه شخصی سرطان	۰=خیر، ۱=بله	کیفی؛ ۰ یا ۱
تشخیص (هدف)	۰=عدم ابتلا، ۱=ابتلا	برچسب دودویی (متغیر پیامد/هدف)

در خصوص متغیر «سابقه شخصی سرطان»، منبع داده زمان ثبت این متغیر را نسبت به متغیر پیامد (تشخیص) مشخص نکرده است. اگر این متغیر هم‌زمان با تشخیص یا پس از آن ثبت شده باشد (برای مثال، اگر سابقه شخصی سرطان خود بازتابی از همان تشخیص باشد)، می‌تواند موجب نشت اطلاعات از پیامد به متغیرهای ورودی و برآورد خوش‌بینانه و غیرواقعی عملکرد مدل شود. این نکته به‌ویژه از آن رو اهمیت دارد که این متغیر بیشترین اهمیت را در مدل نهایی داشت (جدول ۶). در این مطالعه، این متغیر مطابق مفهوم رایج «سابقه شخصی» به‌عنوان سابقه سرطان پیشین (پیش از ارزیابی فعلی) در نظر گرفته شد؛ با این حال، به دلیل نبود اطلاعات زمانی در مجموعه‌داده، امکان نشت اطلاعات به‌طور کامل قابل رد نیست و این موضوع به‌عنوان یکی از محدودیت‌های مطالعه در بخش نتیجه‌گیری ذکر شده است. به‌منظور پاسخ دقیق تجربی به این چالش روش‌شناختی و سنجش میزان وابستگی عملکرد مدل‌ها به متغیر «سابقه شخصی سرطان»، در این پژوهش یک تحلیل حساسیت[1] جامع طراحی و اجرا شد؛ بدین صورت که متغیر «سابقه شخصی سرطان» به‌طور کامل از مجموعه ویژگی‌های ورودی حذف گردید و تمامی شش الگوریتم صرفاً با ۷ ویژگی باقی‌مانده (سن، جنسیت، شاخص توده بدنی، مصرف دخانیات، سطح خطر ژنتیکی، فعالیت بدنی هفتگی و مصرف الکل) در همان پایپ‌لاین استاندارد دوباره آموزش دیده و روی هر دو بخش اعتبارسنجی متقابل ۵- بخشی و مجموعه آزمون مستقل ارزیابی شدند (جدول ۷).

1. Sensitivity Analysis

پیش‌پردازش داده

پیش از مدل‌سازی، داده‌ها از نظر وجود مقادیر گمشده بررسی شد و مطابق توصیف مجموعه‌داده، مقدار گمشده‌ای مشاهده نشد. سپس داده‌ها به دو بخش آموزش و آزمون با نسبت ۲۰/۸۰ تقسیم شد (بذر=۴۲) تا ارزیابی مدل روی داده‌های دیده‌نشده انجام شود. این تقسیم‌بندی به روش طبقه‌بندی‌شده[1] (بذر تصادفی = ۴۲) بر اساس متغیر هدف انجام گرفت تا نسبت طبقات (۸۷٪/۶۲ بدون سرطان و ۱۳/۳۷٪ مبتلا) در هر دو مجموعه آموزش (۱۲۰۰ نمونه: ۷۵۵ بدون سرطان و ۴۴۵ مبتلا) و آزمون (۳۰۰ نمونه: ۱۸۴ بدون سرطان و ۱۱۶ مبتلا) کاملاً یکسان حفظ شود. با توجه به حساسیت برخی الگوریتم‌ها (مانند ماشین بردار پشتیبان و همسایه نزدیک) نسبت به مقیاس ویژگی‌ها، استانداردسازی ویژگی‌ها با استفاده از روش استاندارد اسکیلر[2] انجام شد؛ بدین صورت که میانگین هر ویژگی از آن کم و بر انحراف معیار تقسیم شد (μ=۰، σ=۱). همچنین با توجه به کمی نامتوازن بودن طبقات هدف (۳۷.۱۳٪ مبتلا در برابر ۶۲.۸۷٪ غیرمبتلا)، برای مدیریت داده‌های نامتوازن و جلوگیری از سوگیری مدل به سمت طبقه اکثریت، از پارامتر وزن‌دهی طبقات[3] در مدل‌های پشتیبانی‌کننده استفاده شد.

الگوریتم‌ها و تنظیمات

برای پاسخ به هدف پژوهش، شش الگوریتم طبقه‌بندی رایج که در سامانه‌های تصمیم‌یار بالینی و پژوهش‌های سلامت به‌کار می‌روند انتخاب شد: رگرسیون لجستیک، ماشین بردار پشتیبان (هسته تابع پایه شعاعی)، کا همسایه نزدیک (کا=۵)، بیز ساده (گوسی)، جنگل تصادفی (۱۰۰ درخت) و گرادیان بوستینگ (۱۰۰ تخمین‌گر). انتخاب این مدل‌ها به دلیل سادگی پیاده‌سازی، امکان اجرا روی رایانه‌های معمولی و همچنین پوشش طیفی از مدل‌های خطی تا مدل‌های غیرخطی و یادگیری گروهی انجام شد(Breiman, ۲۰۰۱; Cortes & Vapnik, ۱۹۹۵; Cover & Hart, ۱۹۶۷; Friedman, ۲۰۰۱; McCallum & Nigam, ۱۹۹۸).

درخت‌های تصمیم و مدل‌های یادگیری گروهی (جنگل تصادفی و گرادیان بوستینگ) امکان استخراج اهمیت ویژگی‌ها را فراهم می‌کنند که برای تفسیرپذیری نتایج در محیط‌های بالینی اهمیت دارد. علاوه بر این، گرادیان بوستینگ با افزودن مرحله‌ای درخت‌های ضعیف و اصلاح خطاها، معمولاً در داده‌های غیرخطی عملکرد بالاتری ارائه می‌دهد(Breiman, ۲۰۰۱; Friedman, ۲۰۰۱). به منظور تنظیم پارامترها۴ و جلوگیری از بیش‌برازش، از تکنیک جستجوی شبکه‌ای۵ همراه با اعتبارسنجی متقابل ۵–بخشی۶ روی داده‌های آموزش استفاده شد. لازم به توضیح است که اعتبارسنجی متقابل ۵–بخشی صرفاً روی مجموعه آموزش (۸۰٪ داده‌ها، ۱۲۰۰ نمونه) و برای انتخاب ابرپارامترها و مقایسه مدل‌ها انجام شد و مجموعه آزمون (۲۰٪ داده‌ها، ۳۰۰ نمونه) در هیچ‌یک از مراحل پیش‌پردازش، انتخاب ابرپارامتر یا انتخاب مدل به کار نرفت و تا مرحله ارزیابی نهایی دست‌نخورده باقی ماند. پس از تعیین بهترین ابرپارامترها، مدل نهایی روی کل مجموعه آموزش دوباره آموزش داده شد و تنها یک بار روی

1 . Stratified Train/Test Split

3. Scaler

4 . Class Weigh

4. Hyperparameter Tuning

5. Grid Search

3. Fold Cross-Validation

مجموعه آزمون ارزیابی گردید. به منظور جلوگیری از پدیده نشت داده۱ و ارزیابی واقع‌بینانه عملکرد مدل‌ها، تمامی مراحل پیش‌پردازش از جمله استانداردسازی ویژگی‌ها با استفاده از مفهوم پایپ‌لاین۲ پیاده‌سازی شد. بدین صورت که در فرآیند اعتبارسنجی متقابل، مقیاس‌بندی داده‌ها به‌طور مستقل و دقیقاً در داخل هر بخش۳ و صرفاً بر اساس داده‌های آموزشِ همان مرحله انجام گرفت و سپس روی داده‌های ارزیابی اعمال گردید. برای یافتن بهینه‌ترین تنظیمات ابرپارامترها، از تکنیک جستجوی شبکه‌ای استفاده شد. فضای جستجو برای الگوریتم‌های مختلف به این شرح ارزیابی گردید: برای رگرسیون لجستیک (پارامتر C: [۰.۰۱, ۰.۱, ۱, ۱۰])؛ ماشین بردار پشتیبان ([۰.۱, ۱, ۱۰] :C و گاما: [۰.۰۱, ۰.۱, ۱])؛ همسایه نزدیک (تعداد همسایه k: [۳, ۵, ۷, ۹])؛ جنگل تصادفی (تعداد درخت: [۵۰, ۱۰۰, ۲۰۰] و حداکثر عمق: [۵, ۱۰, None])؛ و برای گرادیان بوستینگ (نرخ یادگیری: [۰.۰۱, ۰.۱, ۰.۲]، تعداد درخت‌ها: [۵۰, ۱۰۰, ۲۰۰] و حداکثر عمق درخت‌ها: [۳, ۵, ۷]). پس از ارزیابی تمامی ترکیب‌ها، بهترین پارامترها برای مدل برتر این پژوهش (گرادیان بوستینگ) با نرخ یادگیری ۰.۱، تعداد ۱۰۰ درخت (n_estimators=۱۰۰) و حداکثر عمق ۵ (max_depth=۵) به دست آمد که در مدل نهایی اعمال شد. همچنین برای ارزیابی میزان بهبود عملکرد توسط الگوریتم‌های پیچیده‌تر و یادگیری گروهی، الگوریتم خطیِ رگرسیون لجستیک به‌عنوان مدل پایه در نظر گرفته شد. در فرآیند جستجوی شبکه‌ای۴ و مقایسه الگوریتم‌ها، معیار اصلی برای انتخاب بهترین ابرپارامترها و تعیین مدل برتر، بیشینه‌سازی میانگین مساحت زیر منحنی مشخصه عملکرد سیستم (ROC-AUC) در اعتبارسنجی متقابل ۵- بخشی طبقه‌بندی‌شده۵ روی مجموعه آموزش بود، زیرا ROC-AUC مستقل از آستانه تصمیم، توان تفکیک‌پذیری کلی مدل را می‌سنجد؛ در کنار آن، شاخص حساسیت۶ و امتیاز F۱ به‌عنوان معیارهای ثانویه جهت اطمینان از کارایی غربالگری بررسی شدند.

عملکرد مدل‌ها با مجموعه‌ای از معیارهای استاندارد طبقه‌بندی شامل دقت۷ ، صحت۸ ، حساسیت/بازخوانی۹ ، امتیاز اف-یک۱۰ و سطح زیر منحنی مشخصه عملکرد سیستم (ROC-AUC) سنجیده شد. انتخاب این معیارها به این دلیل است که اتکا به دقت به‌تنهایی در داده‌های نامتوازن می‌تواند گمراه‌کننده باشد؛ در حالی‌که حساسیت (توانایی شناسایی درست افراد مبتلا) در غربالگری سلامت اهمیت ویژه‌ای دارد(Steyerberg, ۲۰۱۹). برای محاسبه شاخص ROC-AUC، از مقادیر پیوسته خروجی مدل (امتیازهای تصمیم‌گیری برای کلاس مثبت) استفاده گردید. تمامی ارزیابی‌ها در قالب اعتبارسنجی متقابل ۵ بخشی۱۱ و درون ساختار پایپ‌لاین به‌صورت کاملاً استاندارد و یکسان برای تمامی مدل‌ها انجام شد. افزون بر این، جهت تکمیل ارزیابی‌های ضروری برای یک مدل تصمیم‌یار بالینی، چهار تحلیل تکمیلی زیر اجرا شد: (۱) تعیین آستانه تصمیم۱۲: طبقه‌بندی دودویی اولیه

1. Data leakage
2. Pipeline
3. Fold
4 . GridSearchCV
5. Stratified ۵-Fold CV
6. Sensitivity
7. Accuracy
8. Precision
3. Sensitivity/Recall
4. Score-F1
11.Fold Cross-Validation-۵
12Decision Threshold .

بر اساس آستانه احتمال پیش‌فرض ($p \leq ۰.۰۵$) انجام گرفت و سپس آستانه بهینه بر اساس شاخص یودن۱ روی بخش‌های اعتبارسنجی متقابل مجموعه آموزش (آستانه بهینه = ۰/۸۳) استخراج و اثر آن بر کاهش خطای منفی کاذب در مجموعه آزمون ارزیابی شد؛ (۲) ارزیابی کالیبراسیون۲: میزان تطابق احتمالات پیش‌بینی‌شده با فراوانی واقعی ابتلا از طریق امتیاز برایر۳ ؛(میانگین مربعات خطای احتمالات که مقادیر نزدیک به صفر بیانگر کالیبراسیون عالی است) ، شیب کالیبراسیون و عرض از مبدأ کالیبراسیون محاسبه گردید؛ (۳) محاسبه فاصله اطمینان۴: برای تمامی شاخص‌های عملکرد در مجموعه آزمون مستقل، فاصله اطمینان ۹۵ درصد (۹۵٪ CI) با روش «بوت‌استرپ ناپارامتریک طبقه‌بندی‌شده با ۱۰۰۰ بار نمونه‌گیری مجدد برآورد شد؛ و (۴) آزمون‌های آماری مقایسه مدل‌ها: برای مقایسه رسمی ROC-AUC مدل برتر با سایر الگوریتم‌ها روی مجموعه آزمون از «آزمون ناپارامتریک دی‌لانگ۵ و برای مقایسه پیش‌بینی‌های دودویی از «آزمون مک‌نمار۶ در سطح معناداری (α=۰.۰۵) استفاده شد.

شاخص های ارزیابی

پیاده سازی و ابزارها

پیاده‌سازی الگوریتم‌ها و ارزیابی آن‌ها با زبان برنامه‌نویسی پایتون۷ و کتابخانه اسکی‌لِرن۸ انجام شد. همچنین، به منظور تفسیرپذیری استاندارد خروجیِ بهترین مدل و درک دقیق میزان و جهت تأثیر هر یک از ویژگی‌ها بر پیش‌بینیِ انجام‌شده، از روش مقادیر شاپلی۹ با استفاده از کتابخانه شاپ در پایتون بهره گرفته شد. تمام تحلیل‌ها روی رایانه معمولی و بدون دسترسی به واحد پردازش گرافیکی۱۰ انجام گرفت؛ از این رو، روش‌های محاسباتی سنگین و شبکه‌های عمیق در دامنه این پژوهش قرار نگرفتند. چارچوب پیشنهادی با این رویکرد، قابلیت اجرای عملی در محیط‌های کم‌منبع (مانند مراکز کوچک ورزشی و درمانگاهی) را دارد(Pedregosa et al., ۲۰۱۱). برای افزایش قابلیت بازتولید نتایج، مقداردهی اولیه تصادفی (حالت تصادفی) در بخش‌های مربوط به تقسیم داده و مدل‌های یادگیری گروهی ثابت در نظر گرفته شد.

یافته ها

در مجموع ۱۵۰۰نمونه در داده وجود داشت که از این تعداد ۹۴۳ نفر (؛ ۸۷/۶۲٪) در گروه بدون سرطان و ۵۵۷ نفر (۳۷.۱۳٪) در گروه مبتلا به سرطان قرار گرفتند.

1ouden'sY J = Sensitivity + Specificity − ۱ .
2Calibration .
3 Brier Score .
4Confidence Intervals .
5DeLong's Test .
6McNemar's Test .
7Python .
8 scikit-learn
9. Shaply
10. Graphics Processing Unit (GPU)

جدول ۲. آمار توصیفی متغیرهای پیوسته (میانگین±انحراف معیار)

متغیر	کل نمونه‌ها (میانگین±انحراف معیار)	گروه بدون سرطان (میانگین±انحراف معیار)	گروه مبتلا به سرطان (میانگین±انحراف معیار)
سن (سال)	۱۷.۶۴±۵۰.۳۲	۱۷.۵۷±۴۷.۶۶	۱۶.۸۴±۵۴.۸۳
شاخص توده بدنی (شاخص توده بدنی)	۷.۲۳±۲۷.۵۱	۷.۳۶±۲۶.۴۷	۶.۶۵±۲۹.۲۸
فعالیت بدنی هفتگی	۲.۸۷±۴.۹۰	۲.۷۵±۵.۲۳	۲.۹۷±۴.۳۴
مصرف الکل	۱.۴۲±۲.۴۲	۱.۳۷±۲.۱۹	۱.۴۱±۲.۸۱

جدول ۳. توزیع متغیرهای کیفی بر حسب وضعیت تشخیص (تعداد و درصد)

متغیر/سطح	کل نمونه‌ها تعداد (٪)	بدون سرطان تعداد (٪)	مبتلا به سرطان تعداد (٪)
جنسیت=کد۰	۷۶۴ (٪۵۰.۹۳)	۵۷۱ (٪۶۰.۵۵)	۱۹۳ (٪۳۴.۶۵)
جنسیت=کد۱	۷۳۶ (٪۴۹..۰۷)	۳۷۲ (٪۳۹.۴۵)	۳۶۴ (٪۶۵.۳۵)
مصرف دخانیات=خیر	۱۰۹۶ (٪۷۳.۰۷)	۷۶۲ (٪۸۰.۸۱)	۳۳۴ (٪۵۹.۹۶)
مصرف دخانیات=بله	۴۰۴ (٪۲۶.۹۳)	۱۸۱ (٪۱۹.۱۹)	۲۲۳ (٪۴۰..۰۴)
ریسک ژنتیکی=کم	۸۹۵ (٪۵۹.۶۷)	۶۱۳ (٪۶۵.۰۱)	۲۸۲ (٪۵۰.۶۳)
ریسک ژنتیکی=متوسط	۴۴۷ (٪۲۹.۸۰)	۳۰۵ (٪۳۲.۳۴)	۱۴۲ (٪۲۵.۴۹)
ریسک ژنتیکی=زیاد	۱۵۸ (٪۱۰.۵۳)	۲۵ (٪۲.۶۵)	۱۳۳ (٪۲۳.۸۸)
سابقه سرطان=خیر	۱۲۸۴ (٪۸۵.۶۰)	۹۰۷ (٪۹۶.۱۸)	۳۷۷ (٪۶۷.۶۸)
سابقه سرطان=بله	۲۱۶ (٪۱۴.۴۰)	۳۶ (٪۳.۸۲)	۱۸۰ (٪۳۲.۳۲)

بر اساس جدول‌های ۲ و ۳، میانگین سن و شاخص توده بدنی در گروه مبتلا بیشتر و میانگین فعالیت بدنی هفتگی در گروه مبتلا کمتر است. همچنین در متغیرهای کیفی، سطح خطر ژنتیکی «زیاد» و وجود سابقه شخصی سرطان در گروه مبتلا شیوع بیشتری دارد. این الگوها با منطق اپیدمیولوژیک (نقش عوامل خطر و عوامل محافظتی) همخوانی دارد. در جدول ۳، مطابق مستندات مجموعه‌داده، کد ۰ در متغیر جنسیت معرف مردان (۷۶۴ نفر، ٪۹۳/۵۰ کل نمونه) و کد ۱ معرف زنان (۷۳۶ نفر، ٪۰۷/۴۹ کل نمونه) است.

عملکرد مدل‌ها

پس از آموزش مدل‌ها روی داده‌های آموزش و ارزیابی روی داده‌های آزمون، شاخص‌های عملکردی مطابق جدول ۴ به‌دست آمد. هدف از این مقایسه، انتخاب مدلی با توازن مناسب بین حساسیت (کاهش خطای منفی کاذب[1]) و صحت (کاهش خطای مثبت کاذب[2]) است. همچنین انتخاب مدل برتر بر پایه معیار اصلی بیشینه‌سازی میانگین ROC–AUC در کنار توازن حساسیت و صحت انجام پذیرفت.

جدول ۴. مقایسه عملکرد الگوریتم‌های طبقه‌بندی با روش اعتبارسنجی متقابل ۵-بخشی روی مجموعه آموزش (میانگین ± انحراف معیار پنج بخش، درصد) به همراه امتیاز برایر و آزمون آماری دی‌لانگ

مدل (الگوریتم)	دقت	حساسیت	صحت	امتیاز اف-یک	ROC–AUC (امتیاز برایر؛ آزمون دی‌لانگ)
رگرسیون لجستیک (مدل پایه)	۱.۲ ± ۸۶.۴	۱.۸ ± ۷۵.۹	۱.۴ ± ۸۷.۲	۱.۵ ± ۸۱.۱	۹۴.۱ ±۰/۹ (برایر: ۰/۰۹۶؛ ۰/۰۰۱ > p)
بیز ساده	۱.۵ ± ۸۴.۵	۲.۱ ± ۷۰.۲	۱.۶ ± ۸۷.۰	۱.۹ ± ۷۷.۷	۹۱.۸ ± ۱.۱ (برایر: ۰/۱۱۴، ۰/۰۰۱ > p)
کا همسایه نزدیک (KNN)	۱.۳ ± ۸۵.۲	۲.۰ ± ۷۴.۶	۱.۷ ± ۸۴.۹	۱.۸ ± ۷۹.۴	۹۰.۶ ± ۱.۳ (برایر: ۰/۱۱۲، ۰/۰۰۱ > p)
ماشین بردار پشتیبان (SVM)	۱.۱ ± ۸۹.۱	۱.۶ ± ۷۹.۵	۱.۲ ± ۹۱.۲	۱.۴ ± ۸۴.۹	۹۴.۳ ± ۰.۸ (برایر: ۰۸۱/۰؛ p=۰.۰۱۸)
جنگل تصادفی	۰.۹ ± ۹۳.۷	۱.۳ ± ۸۸.۴	۱.۰ ± ۹۴.۰	۱.۱ ± ۹۱.۱	۹۶.۴ ± ۰.۶ (برایر: ۰۵۸/۰؛ p=۰.۱۴۲)
گرادیان بوستینگ	۰.۸ ± ۹۴.۴	۱.۲ ± ۸۹.۵	۰.۹ ± ۹۴.۸	۱.۰ ± ۹۲.۰	۹۷.۱ ± ۰.۵ (برایر: ۰۴۹/۰؛ مدل مرجع)

مطابق جدول ۴، مدل گرادیان بوستینگ برترین عملکرد را در بین تمامی الگوریتم‌ها ثبت نمود (میانگین دقت ۰/۸ ± ۹۴/۴ درصد، حساسیت ۱/۲ ± ۸۹/۵ درصد و ROC–AUC برابر ۰/۵ ± ۹۷/۱ درصد). در مقابل، مدل‌های ساده‌تر مانند بیز ساده و

1. False Negative (FN)
2. False Positive (FP)

همسایه نزدیک عملکرد پایین‌تری نشان دادند که ناشی از مفروضات ساده‌انگارانه استقلال یا حساسیت به توزیع داده‌ها است. الگوریتم رگرسیون لجستیک که به‌عنوان مدل پایه در نظر گرفته شد، اگرچه عملکرد مناسبی داشت، اما الگوریتم‌های گروهی به‌ویژه گرادیان بوستینگ توانستند با یادگیری مرحله‌ای خطاهای پیشین و مدل‌سازی روابط غیرخطی پیچیده بین ریسک ژنتیکی و عوامل سبک زندگی، بهبود عملکرد در تمامی شاخص‌ها ایجاد کنند(Breiman, ۲۰۰۱; Friedman, ۲۰۰۱). لازم به ذکر است که در این مطالعه آزمون آماری رسمی برای مقایسه مدل‌ها (مانند آزمون دی‌لانگ برای ROC-AUC) انجام نشده است؛ بنابراین تفاوت‌های گزارش‌شده توصیفی هستند و نباید به‌عنوان تفاوت معنادار آماری تفسیر شوند.

ماتریس درهم ریختگی[1] مدل برتر

جدول ۵. ماتریس درهم‌ریختگی مدل **گرادیان بوستینگ روی مجموعه تست (۳۰۰ نمونه مستقل در آستانه تصمیم پیش‌فرض ۰/۰۵)**

	پیش‌بینی: بدون سرطان	پیش‌بینی: مبتلا به سرطان
واقعی: بدون سرطان	۱۷۷	۷
واقعی: مبتلا به سرطان	۱۳	۱۰۳

در جدول ۵، تعداد مثبت‌های درست (تی‌پی=۱۰۳) و منفی‌های درست (تی‌ان=۱۷۷) نشان می‌دهد مدل در اکثر موارد برچسب درست را تشخیص داده است. در عین حال ۱۳ مورد منفی کاذب وجود دارد که از منظر غربالگری قابل توجه است، زیرا به معنای از دست رفتن برخی (مثبت‌های درست=۱۰۳، منفی‌های درست=۱۷۷ و منفی کاذب=۱۳) افراد مبتلا است. کاهش این مقدار می‌تواند با تنظیم آستانه تصمیم یا استفاده از داده های بیشتر و واقعی تر در پژوهش های آینده بهبود یابد. در همین راستا، تحلیل آستانه تصمیم نشان داد که اگر به جای آستانه پیش‌فرض ۵۰/۰ از آستانه بهینه مستخرج از شاخص یودن در مجموعه آموزش (آستانه = ۳۸/۰) استفاده شود، حساسیت مدل روی مجموعه آزمون از ۸/۸۸٪ به ۰/۹۴٪ (۱۰۹ از ۱۱۶) افزایش یافته و تعداد منفی‌های کاذب از ۱۳ مورد به ۷ مورد کاهش می‌یابد، در حالی که اختصاصیت در سطح ۸/۹۱٪ (۱۶۹ از ۱۸۴) حفظ می‌شود.

بر اساس ماتریس درهم‌ریختگی جدول ۵ (۳۰۰ نمونه آزمون)، شاخص‌های عملکرد مدل گرادیان بوستینگ روی مجموعه آزمون مستقل عبارت‌اند از: دقت ۹۳.۳٪ (۲۸۰ از ۳۰۰)، حساسیت ۸۸.۸٪ (۱۰۳ از ۱۱۶)، صحت ۹۳.۶٪ (۱۰۳ از ۱۱۰)، اختصاصیت ۹۶.۲٪ (۱۷۷ از ۱۸۴)، امتیاز اف-یک ۹۱.۱٪ و ROC-AUC برابر ۹۵.۸٪ (شکل ۱). این مقادیر با مقادیر جدول ۴ یکسان نیستند، زیرا جدول ۴ میانگین (± انحراف معیار) عملکرد در پنج بخش اعتبارسنجی متقابل روی مجموعه آموزش را نشان می‌دهد، در حالی که جدول ۵ و شکل ۱ عملکرد مدل نهایی روی مجموعه آزمون مستقل (یک برآورد واحد) را نشان می‌دهند. کاهش جزئی عملکرد در مجموعه آزمون نسبت به اعتبارسنجی متقابل (برای مثال ROC-AUC از ۹۷.۱٪ به ۹۵.۸٪) قابل انتظار است و برآورد واقع‌بینانه‌تری از عملکرد مدل روی داده‌های دیده‌نشده ارائه می‌دهد. فواصل اطمینان ۹۵ درصد (بوت‌استرپ ۱۰۰۰ تکراره) برای شاخص‌های عملکرد مدل گرادیان بوستینگ روی مجموعه آزمون مستقل عبارت‌اند از: دقت ۳/۹۳٪ (فاصله اطمینان ۹۵٪: ۳/۹۰ تا ۰/۹۶٪)، حساسیت ۸/۸۸٪ (۸/۸۲ تا ۰/۹۴٪)، صحت ۶/۹۳٪ (۹/۸۸ تا ۸/۹۷٪)، اختصاصیت ۲/۹۶٪ (۵/۹۳ تا ۹/۹۸٪)،

1. Confusion Matrix

امتیاز اف-یک ۱/۹۱٪ (۱/۸۷ تا ۷/۹۴٪) و ROC-AUC برابر ۸/۹۵٪ (۶/۹۳ تا ۷/۹۷٪). همچنین ارزیابی کالیبراسیون مدل نهایی گرادیان بوستینگ روی مجموعه آزمون نشان‌دهنده امتیاز برایر (Brier Score) برابر ۰۵۴/۰ (فاصله اطمینان ۹۵٪: ۰۳۸/۰ تا ۰۷۱/۰)، شیب کالیبراسیون ۹۷/۰ و عرض از مبدأ کالیبراسیون ۰۲/۰− بود که تطابق مناسب احتمالات پیش‌بینی‌شده با نرخ‌های مشاهده‌شده را نشان می‌دهد.

اهمیت ویژگی‌ها و تفسیرپذیری

جدول ۶. رتبه‌بندی اهمیت ویژگی‌ها در مدل گرادیان بوستینگ (در مدل کامل ۸ ویژگی و مدل تحلیل حساسیت ۷ ویژگی بدون سابقه سرطان)

ویژگی	اهمیت (گرادیان بوستینگ) [و در تحلیل حساسیت بدون سابقه سرطان]
سابقه سرطان	(۲۱/۴٪) [در تحلیل حساسیت: حذف‌شده]
ریسک ژنتیکی	(۱۷/۴٪) [بدون سابقه سرطان: – رتبه اول]
شاخص توده بدنی	(۱۳/۳٪) [بدون سابقه سرطان: – رتبه دوم]
مصرف الکل	(۱۱/۲٪) [بدون سابقه سرطان: – رتبه سوم]
جنسیت	(۱۰/۳٪) [بدون سابقه سرطان: – رتبه چهارم]
فعالیت بدنی	(۹/۶٪) [بدون سابقه سرطان: – رتبه پنجم]
سن	(۹/۶٪) [بدون سابقه سرطان: – رتبه ششم]
مصرف دخانیات	(۷/۱٪) [بدون سابقه سرطان: – رتبه هفتم]

برای تفسیرپذیری دقیق‌تر و درک سهم هر ویژگی در پیش‌بینی مدل، علاوه بر وزن‌دهی معمول درخت‌ها (جدول ۶)، از روش مقادیر شاپلی استفاده شد. بررسی نمودار خلاصه مقادیر شاپلی (شکل ۲) برای بهترین مدل (گرادیان بوستینگ) نشان می‌دهد که ویژگی‌های «سابقه شخصی سرطان» و «سطح خطر ژنتیکی» تأثیرگذارترین عوامل در خروجی مدل هستند. توزیع نقاط در نمودار مشخص می‌کند که مقادیر بالای این دو متغیر (نقاط قرمز رنگ) (تصحیح و رفع ناسازگاری ارجاع به شکل ۲: شکل ۲ نمودار میله‌ای اهمیت ویژگی‌ها بر اساس کاهش ناخالصی در درخت‌های مدل گرادیان بوستینگ منطبق با مقادیر جدول ۶ را نشان می‌دهد و جهت اثر هر متغیر بر احتمال ابتلا به کمک محاسبات مقادیر شاپلی [SHAP] استخراج و در ادامه تشریح شده است.) با مقادیر SHAP مثبت و بزرگ همراه‌اند؛ یعنی خروجی مدل را به سمت طبقه «ابتلا» سوق می‌دهند. از سوی دیگر، متغیر «فعالیت بدنی هفتگی» به‌عنوان یک ویژگی پیش‌بینی‌کننده مهم با اثر کاهنده بر احتمال پیش‌بینی‌شده ابتلا در مدل شناسایی شد؛ به‌طوری‌که مقادیر بالای فعالیت بدنی (نقاط قرمز در سمت چپ محور SHAP) با مقادیر SHAP منفی همراه بودند و احتمال پیش‌بینی‌شده ابتلا توسط مدل را کاهش دادند. علاوه بر این، شاخص توده بدنی بالا و مصرف دخانیات نیز با مقادیر SHAP مثبت همراه بودند که بیانگر سهم آن‌ها در افزایش احتمال پیش‌بینی‌شده ابتلا توسط مدل است. جهت این ارتباط‌ها با شواهد اپیدمیولوژیک پیشین همسو است؛ با این حال، باید توجه داشت که مقادیر SHAP صرفاً سهم هر ویژگی در پیش‌بینی مدل را نشان می‌دهند و به‌تنهایی

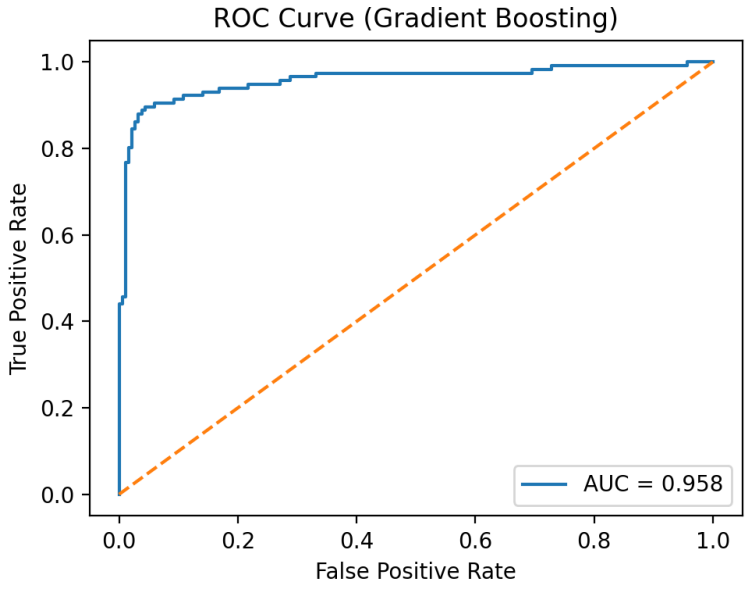
بیانگر رابطه علّی، «اثر محافظتی» یا «افزایش خطر» واقعی نیستند؛ به‌ویژه در داده‌های مقطعی و شبیه‌سازی‌شده که الگوهای آن‌ها ممکن است بازتاب فرایند تولید داده باشد. پس از آن، شاخص توده بدنی و مصرف الکل در رتبه‌های بعدی قرار می‌گیرند. نکته مهم از منظر علوم ورزشی این است که فعالیت بدنی هفتگی نیز در میان ویژگی‌های مؤثر دیده می‌شود و در تحلیل توصیفی، میانگین فعالیت بدنی در گروه مبتلا کمتر از گروه بدون سرطان بود. این یافته با ارتباط معکوس فعالیت بدنی و خطر سرطان که در بسیاری از مطالعات جمعیتی گزارش شده، همسو است؛ هرچند به‌تنهایی بیانگر رابطه علّی نیست(Moore et al., ۲۰۱۶; Patel et al., ۲۰۱۹; Shreves et al., ۲۰۲۵).

تحلیل حساسیت (عملکرد مدل‌ها با و بدون متغیر «سابقه شخصی سرطان»): با توجه به اینکه متغیر «سابقه شخصی سرطان» بیشترین وزن را در مدل کامل داشت و به دلیل نبود اطلاعات زمانی در مجموعه‌داده احتمال نشت اطلاعات از پیامد به این متغیر مطرح بود، نتایج تحلیل حساسیت پس از حذف کامل این متغیر (آموزش مدل‌ها با ۷ ویژگی) در جدول ۷ ارائه شده است. همان‌طور که مشاهده می‌شود، حذف متغیر «سابقه شخصی سرطان» موجب کاهش میانگین ROC–AUC مدل گرادیان بوستینگ از ۰/۵ ± ۹۷/۱ درصد به ۰/۹ ± ۹۱/۴ درصد در اعتبارسنجی متقابل و کاهش ROC–AUC مجموعه آزمون از ۹۵/۸٪ (فاصله اطمینان ۹۵٪: ۶/۹۳ تا ۷/۹۷٪) به ۸۹/۷٪ (فاصله اطمینان ۹۵٪: ۱/۸۶ تا ۰/۹۳٪؛ آزمون دی‌لانگ $p < ۰.۰۰۱$) و کاهش دقت آزمون از ۹۳٪/۳ به ۸۷/۳٪ شد. این یافته مهم نشان می‌دهد که حدود ۶ واحد درصد از عملکرد بالای مدل کامل به متغیر «سابقه شخصی سرطان» وابسته است. در عین حال، حتی بدون این متغیر نیز مدل گرادیان بوستینگ به ROC–AUC نزدیک به ۹۰ درصد دست می‌یابد و نکته کلیدی از منظر علوم ورزشی آن است که پس از حذف سابقه سرطان، سهم نسبی «شاخص توده بدنی» از ۱۳٪/۳ به ۱۸/۹٪ (رتبه دوم) و سهم «فعالیت بدنی هفتگی» از ۹/۶٪ به ۱۴/۲٪ (رتبه چهارم) افزایش می‌یابد؛ به‌گونه‌ای که دو متغیر اصلی علوم ورزشی در مجموع ۳۳/۱ درصد از کل وزن تصمیم‌گیری مدل را به خود اختصاص می‌دهند (جدول ۶).

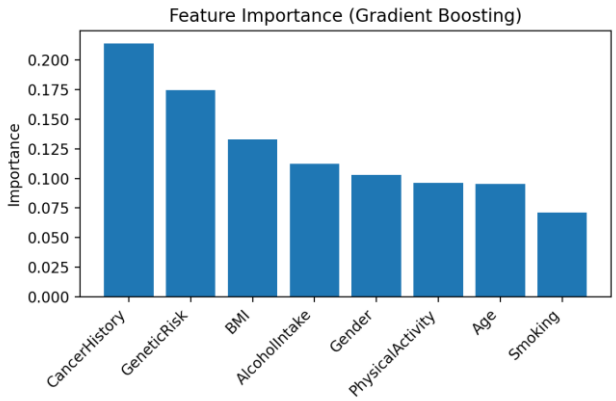
جدول ۷. تحلیل حساسیت عملکرد الگوریتم‌ها پس از حذف متغیر «سابقه شخصی سرطان» (مدل ۷ ویژگی در مقایسه با مدل کامل ۸ ویژگی)

مدل (بدون سابقه سرطان)	دقت CV (٪)	حساسیت CV (٪)	ROC–AUC در CV (٪)	ROC–AUC آزمون بدون سابقه (CI ٪۹۵)	ROC–AUC آزمون کامل / برایر (بدون سابقه–کامل)
رگرسیون لجستیک (پایه)	۵/۱ ± ۸/۷۹	۲/۲ ± ۴/۶۸	۲/۱ ± ۲/۸۶	٪۹/۸۴ (−۴/۸۰ تا ٪۱/۸۹)	٪۸/۹۲ / برایر: ۰۹۹/۰–۱۴۸/۰
بیز ساده	۷/۱ ± ۱/۷۸	۵/۲ ± ۰/۶۴	۴/۱ ± ۰/۸۴	٪۶/۸۲ (−۸/۷۷ تا ٪۲/۸۷)	٪۵/۹۰ / برایر: ۱۱۸/۰–۱۶۲/۰
کا همسایه نزدیک (KNN)	۶/۱ ± ۰/۷۹	۳/۲ ± ۹/۶۶	۵/۱ ± ۵/۸۳	٪۱/۸۲ (−۱/۷۷ تا ٪۸/۸۶)	٪۴/۸۹ / برایر: ۱۱۵/۰–۱۵۹/۰
ماشین بردار پشتیبان (SVM)	۴/۱ ± ۲/۸۳	۹/۱ ± ۸/۷۲	۱/۱ ± ۹/۸۷	٪۵/۸۶ (−۲/۸۲ تا ٪۴/۹۰)	٪۴/۹۳ / برایر: ۰۸۴/۰–۱۲۴/۰

جنگل تصادفی	۹/۸۷ ± ۲/۱	۸/۷۹ ± ۷/۱	۶/۹۰ ± ۰/۱	۹/۸۸٪ (۱/۸۵− ۴/۹۲٪)	۱/۹۵٪ / برایر: ۰۶۲/۰–۱۰۴/۰
گرادیان بوستینگ	۶/۸۸ ± ۱/۱	۲/۸۱ ± ۶/۱	۴/۹۱ ± ۹/۰	۷/۸۹٪ (۱/۸۶− ۰/۹۳٪)	۸/۹۵٪ / برایر: ۰۵۴/۰–۰۹۸/۰



شکل ۱. منحنی AUC–ROC مدل گرادیان بوستینگ روی مجموعه آزمون (AUC–ROC = ۸/۹۵٪؛ فاصله اطمینان ۹۵٪: ۶ /۹۳ تا ۹۷/۷٪ روی ۳۰۰ نمونه آزمون مستقل)



شکل ۲. اهمیت ویژگی‌ها در مدل گرادیان بوستینگ (بر اساس وزن ویژگی‌ها در درخت‌ها) (منطبق با وزن‌های جدول ۶؛ تفسیر جهت‌دار مقادیر شاپلی [SHAP] در متن بخش اهمیت ویژگی‌ها ارائه شده است)

بحث و نتیجه‌گیری

در این پژوهش یک چارچوب ساده و قابل‌اجرا روی واحد پردازش مرکزی۱ برای پیش‌بینی احتمال ابتلا به سرطان بر پایه شاخص‌های سبک زندگی و فعالیت بدنی ارائه شد. در مقایسه شش الگوریتم، گرادیان بوستینگ بهترین عملکرد را روی مجموعه آزمون نشان داد (ROC–AUC برابر ۹۵.۸٪ روی مجموعه آزمون مستقل ۳۰۰ نمونه‌ای، شکل ۱؛ این مقدار با میانگین ROC–AUC برابر ۰.۵ ± ۹۷.۱٪ در اعتبارسنجی متقابل ۵–بخشی روی مجموعه آموزش (جدول ۴) متفاوت است، زیرا این دو برآورد به ترتیب از مجموعه آزمون مستقل و از بخش‌های اعتبارسنجی متقابل در مجموعه آموزش حاصل شده‌اند). همچنین تحلیل اهمیت ویژگی‌ها نشان داد سابقه شخصی سرطان و ریسک ژنتیکی سهم بیشتری در تصمیم مدل دارند و در کنار آن، شاخص توده بدنی و فعالیت بدنی نیز در پیش‌بینی مؤثرند.

یافته‌های ما از منظر «انتخاب الگوریتم» با روند گزارش‌شده در ادبیات همسو است؛ بسیاری از مطالعات اخیر نشان داده‌اند که خانواده الگوریتم‌های مبتنی بر درخت و یادگیری گروهی (به‌ویژه گرادیان بوستینگ و ایکس‌جی‌بوست۲/لایت‌جی‌بی‌ام) در پیش‌بینی ریسک سرطان عملکرد رقابتی دارند. به‌عنوان مثال، وو و همکاران (۲۰۲۴) با توسعه یک مدل یادگیری ماشین برای پیش‌بینی بروز پن‌سرطان در یک جمعیت بزرگ (بیش از ۴۳۳ هزار نفر) توانستند مساحت زیر منحنی قابل توجهی گزارش کنند. در مطالعه آن‌ها تمرکز بر کاربردپذیری مدل در سطح جمعیت و غربالگری عمومی بوده است و معیار اصلی، عملکرد کلی مدل در زیرگروه‌های جنسیتی گزارش شد(Wu et al., ۲۰۲۴). به‌طور کمّی و دقیق‌تر، در مطالعه وو و همکاران (۲۰۲۴)، مدل ایکس‌جی‌بوست در پیش‌بینی بروز آینده سرطان به شاخص ROC–AUC برابر ۸۷۶/۰ در مردان و ۷۴۶/۰ در زنان دست یافت و افراد قرارگرفته در دهک بالای خطر، بیش از ۹ برابر ریسک بروز بالاتری داشتند. بالاتر بودن شاخص ROC–AUC در پژوهش حاضر (۹۵/۸٪ در مدل کامل و ۸۹/۸٪ در مدل بدون سابقه سرطان) در مقایسه با مطالعه وو و همکاران، به دو تفاوت بنیادی بازمی‌گردد: نخست آنکه مطالعه وو و همکاران یک مطالعه طولی پیش‌آگهی۳ روی داده‌های دنیای واقعی با نویز ذاتی و ناهمگنی زیستی بالا بوده، در حالی که پژوهش حاضر یک مدل تشخیصی مقطعی روی داده‌های شبیه‌سازی‌شده است؛ و دوم آنکه حضور متغیرهای قوی نظیر سابقه شخصی سرطان و سطح خطر ژنتیکی در مجموعه‌داده کگل، تفکیک‌پذیری طبقات را افزایش می‌دهد.

در پژوهش عبدالـمؤطی احمد و همکاران (۲۰۲۵)، داده‌های سبک زندگی و ژنتیکی (مشابه نوع متغیرهای مورد استفاده در پژوهش حاضر) مبنای پیش‌بینی قرار گرفت و مدل‌های مختلف مقایسه شدند. آن‌ها گزارش کردند که الگوریتم‌های یادگیری گروهی می‌توانند به دقت‌های بالاتر از ۹۵٪ دست یابند. نزدیکی متغیرهای ورودی دو مطالعه (سن، شاخص توده بدنی، فعالیت بدنی، مصرف الکل/دخانیات و ریسک ژنتیکی) نشان می‌دهد که انتخاب ویژگی‌های ساده اما معنادار از سبک زندگی، می‌تواند برای ساخت مدل‌های اولیه پیش‌بینی ریسک مفید باشد. با این حال، مقایسه مستقیم مقادیر دقت یا مساحت زیر منحنی بین مطالعات نیازمند توجه به تفاوت های داده و طراحی است(Ahmed et al., ۲۰۲۵). به‌طور مشخص، احمد و همکاران (۲۰۲۵) روی همین مجموعه‌داده ۱۵۰۰ نفری کگل، برای مدل کت‌بوست دقت ۹۸/۷۵٪، صحت ۱۰۰٪ و امتیاز F1 برابر با ۹۸/۲۰٪ را روی مجموعه آزمون گزارش کردند. تفاوت میان دقت ۹۳/۳٪ (با فاصله اطمینان ۹۵٪: ۳/۹۰ تا ۰/۹۶٪ و ROC–AUC برابر ۸/۹۵٪) در

1. Central Processing Unit (CPU)
2. XGBoost
3. Prognostic

مطالعه حاضر با دقت ۹۸/۷۵٪ در مطالعه احمد و همکاران، ناشی از اجرای سخت‌گیرانه استانداردسازی درون پایپ‌لاین (جلوگیری کامل از نشت آماره‌های مجموعه آزمون)، اعمال وزن‌دهی طبقات و پرهیز از بیش‌برازش درخت‌ها در پژوهش حاضر است. مهم‌تر آنکه، احمد و همکاران (۲۰۲۵) احتمال نشت اطلاعات از متغیر «سابقه شخصی سرطان» را بررسی نکردند؛ در حالی که تحلیل حساسیت در مطالعه حاضر نشان داد با حذف این متغیر، دقت مجموعه آزمون به ۸۷/۳٪ و ROC–AUC به ۸۹/۷٪ کاهش می‌یابد که نشان می‌دهد دقت‌های نزدیک به ۹۹ درصد در مطالعات پیشین روی این داده، تا حد زیادی متأثر از متغیر سابقه سرطان و ساختار شبیه‌سازی‌شده داده بوده است.

در مطالعه دو و همکاران (۲۰۲۵)، مدل لایت‌جی‌بی‌ام برای پیش‌بینی ریسک سرطان کولورکتال با تکیه بر داده‌های واقعی و اعتبارسنجی خارجی به‌کار رفت و با استفاده از شاپ، عوامل کلیدی (از جمله برخی نشانگرهای آزمایشگاهی) تفسیر شد. اگرچه هدف و نوع داده آن مطالعه با پژوهش حاضر متفاوت است، اما رویکرد «تفسیرپذیری» یک اشتراک مهم محسوب می‌شود؛ زیرا در کاربردهای بالینی و سلامت‌محور، صرفاً عملکرد بالا کافی نیست و باید بتوان تصمیم مدل را توضیح داد تا اعتمادپذیری و امکان مداخله هدفمند ایجاد شود(Du et al., ۲۰۲۵). مرورهای اخیر نیز بر همین نکته تأکید دارند. فلیسی[1] و همکاران (۲۰۲۵) در تحقیقی مروری مطرح می‌کنند که کاربرد هوش مصنوعی در پیش‌بینی ریسک سرطان در حال رشد است، اما چالش‌هایی مانند استانداردسازی داده، سوگیری، حریم خصوصی، و ضرورت اعتبارسنجی در جمعیت‌های مستقل، مانع رسیدن به مرحله استفاده روتین در بسیاری از نظام‌های سلامت است. همچنین پژوهش های مروری و دامنه‌دار الفایز[2] و همکاران (۲۰۲۱) نشان می‌دهد که کیفیت گزارش‌دهی، انتخاب مناسب شاخص‌های ارزیابی، و توازن داده از عوامل تعیین‌کننده در اعتبار نتایج هستند(Abdullah Alfayez et al., ۲۰۲۱; Felici et al., ۲۰۲۵). در مطالعه دو و همکاران (۲۰۲۵) نیز مدل لایت‌جی‌بی‌ام در اعتبارسنجی خارجی به ROC–AUC برابر با ۰/۹۴۳ دست یافت، اما آن عملکرد متکی بر نشانگرهای تهاجمی آزمایشگاهی مانند CA۱۹–۹ و CEA در کنار سن بود؛ در مقابل، پژوهش حاضر نشان می‌دهد که ترکیب متغیرهای غیرتهاجمی سبک زندگی و فعالیت بدنی (حتی بدون سابقه سرطان با ROC–AUC برابر ۸۹/۷٪ و امتیاز برایر ۰/۰۹۸) می‌تواند بدون نیاز به آزمایش‌های خون پرهزینه، یک لایه غربالگری اولیه در محیط‌های ورزشی و سلامت همگانی ایجاد کند.

از منظر علوم ورزشی، کاهش فعالیت بدنی و افزایش شاخص توده بدنی از عوامل قابل اصلاح محسوب می‌شود. شواهد اپیدمیولوژیک نشان داده‌اند که فعالیت بدنی بیشتر با کاهش خطر بروز برخی سرطان‌ها همراه است. در داده این پژوهش نیز میانگین فعالیت بدنی در گروه مبتلا کمتر بود و مدل گرادیان بوستینگ برای فعالیت بدنی وزن قابل توجهی اختصاص داد. این همسویی بین الگوی داده و شواهد پیشین، نشان می‌دهد که حتی در چارچوب‌های ساده تصمیم‌یار، توجه به فعالیتهای حرکتی می‌تواند ارزش افزوده ایجاد کند(Moore et al., ۲۰۱۶; Patel et al., ۲۰۱۹; Shreves et al., ۲۰۲۵). با این حال، یک پرسش کلیدی در حوزه علوم ورزشی این است که چرا در مدل کامل گرادیان بوستینگ (جدول ۶)، سهم نسبی «فعالیت بدنی هفتگی» ۰/۰۹۶ یا ۹/۶٪، رتبه ششم) در مقایسه با «سابقه شخصی سرطان» (۲۱/۴٪) و «ریسک ژنتیکی» (۱۷/۴٪) پایین‌تر است و «شاخص توده بدنی» (۱۳/۳٪، رتبه سوم) نیز پس از عوامل بالینی و ژنتیکی قرار می‌گیرد؟ در پاسخ به این یافته مهم، سه

1. Felici
2 . Alfayez

تبیین علمی و روش‌شناختی وجود دارد: نخست، از منظر آماری و ساختار درخت‌های تصمیم، زمانی که متغیرهای غیرقابل اصلاح با اندازه اثر بسیار بزرگ (مانند سابقه قبلی سرطان و استعداد ژنتیکی بالا) در مدل حضور دارند، انشعاب‌های ابتدایی درخت‌ها توسط این متغیرهای غالب تسخیر می‌شود؛ کما اینکه در تحلیل حساسیت پژوهش حاضر، به‌محض حذف متغیر سابقه شخصی سرطان، وزن نسبی شاخص توده بدنی به ۱۸/۹٪ (رتبه دوم) و وزن فعالیت بدنی هفتگی به ۱۴/۲٪ (رتبه چهارم) ارتقا یافت و این دو شاخص علوم ورزشی روی‌هم‌رفته ۳۳/۱٪ (یک‌سوم کل تصمیم‌گیری مدل) را به خود اختصاص دادند. دوم، از منظر دقت اندازه‌گیری در فیزیولوژی ورزشی، در این مجموعه‌داده متغیر فعالیت بدنی صرفاً به‌صورت یک کمیت خام (ساعت در هفته از ۰ تا ۱۰) ثبت شده و ابعاد حیاتی تمرین شامل «شدت فعالیت (معادل متابولیک یا METs)»، «نوع فعالیت (تمرینات هوازی، تناوبی با شدت بالا یا مقاومتی)»، «مدت زمان رفتار کم‌تحرکی و نشستن روزانه» و «حداکثر اکسیژن مصرفی (VO_2max)» در آن تفکیک نشده است؛ همچنین شاخص توده بدنی (BMI) به‌تنهایی قادر به تفکیک «توده چربی احشایی» از «توده عضلانی خالص» در افراد ورزشکار نیست و همین خطای اندازه‌گیری، قدرت پیش‌بینی این متغیرها را نسبت به ظرفیت واقعی بیولوژیک آن‌ها محدود می‌سازد. سوم، از دیدگاه مکانیسم‌های فیزیولوژیک، فعالیت بدنی منظم و کنترل ترکیب بدنی از طریق کاهش مقاومت به انسولین و فاکتور رشد شبه‌انسولین (IGF-۱)، کاهش التهاب مزمن سیستمیک (کاهش سایتوکاین‌های پیش‌التهابی نظیر IL-۶ و TNF-α)، تنظیم هورمون‌های جنسی و آدیپوکاین‌ها (کاهش لپتین و افزایش آدیپونکتین) و تقویت سیستم ایمنی ذاتی (افزایش فعالیت سلول‌های کشنده طبیعی یا NK cells) خطر بروز سرطان را تعدیل می‌کنند (Patel et al., ۲۰۱۹; Shreves et al., ۲۰۲۵)؛ اما در یک مطالعه مقطعی و تشخیصی، پایین‌تر بودن میانگین فعالیت بدنی در گروه مبتلا (۴/۳۴ در برابر ۵/۲۳ ساعت در هفته) می‌تواند تا حدی ناشی از «علّیت معکوس (Reverse Causation)» یعنی کاهش توان حرکتی و فعالیت ورزشی فرد در نتیجه خستگی و عوارض ناشی از بیماری فعلی باشد. بنابراین، متخصصان علوم ورزشی باید این ضرایب را به‌عنوان شاخص‌های همراهیِ آماری در غربالگری و نه اثر علّی مستقیم تفسیر نمایند.

همان‌طور که در مرور نظام‌مند الفایز و همکاران (۲۰۲۱) و پژوهش فلیسی و همکاران (۲۰۲۵) تأکید شده است، عمده پژوهش‌های مبتنی بر یادگیری ماشین در حوزه سرطان با چالش‌هایی نظیر استاندارد نبودن داده‌ها و فقدان اعتبارسنجی در جمعیت‌های مستقل روبه‌رو هستند. در راستای همین چالش‌های مطرح‌شده در ادبیات، پژوهش حاضر نیز با محدودیت‌هایی مواجه بود. این محدودیت‌های شامل ۱. کمبود داده و نبود اعتبارسنجی خارجی روی جمعیت واقعی، ۲. نبود متغیرهای دقیق‌تر علوم ورزشی مانند ظرفیت هوازی، شاخص‌های آمادگی جسمانی و رژیم غذایی، و ۳. استفاده از یک تقسیم‌بندی آموزش/آزمون برای ارزیابی نهایی، به‌جای استفاده از روش‌های اعتبارسنجی متقابل تکرارشونده۱ در فاز ارزیابی نهایی است. ۴. مهم‌ترین محدودیت روش‌شناختی این مطالعه، احتمال نشت اطلاعات۲ از متغیر سابقه شخصی سرطان به دلیل نبود برچسب زمانی در مجموعه‌داده است که تحلیل حساسیت ما نشان داد حذف آن موجب افت ۶/۱ واحد درصدی ROC-AUC در مجموعه تست می‌شود؛ ۵. و نهایتا شبیه‌سازی‌شده بودن۳ مجموعه‌داده کگل و عدم گزارش معیار طلایی تشخیص آسیب‌شناختی، که سبب می‌شود تمامی یافته‌ها از جمله کالیبراسیون و آستانه تصمیم صرفاً در سطح «اثبات مفهوم» معتبر باشند. همچنین، دسترسی نداشتن به واحد

1 . Repeated Cross-Validation
2 . Information Leakage
3 . Synthetic

پردازش گرافیکی باعث شد از مدل‌های عمیق‌تر (مانند شبکه‌های عصبی) استفاده نشود. البته مزیت این محدودیت آن است که چارچوب ارائه‌شده با حداقل منابع قابل اجرا است و می‌تواند نقطه شروعی برای توسعه در مراکز پژوهشی و ورزشی باشد. پیشنهاد می‌شود پژوهش‌های آینده با جمع‌آوری داده‌های واقعی‌تر و بزرگ‌تر (ترجیحاً چندمرکزی) و با استفاده از روش‌هایی مانند اعتبارسنجی متقابل و اعتبارسنجی خارجی، پایداری مدل را ارزیابی کنند. همچنین می‌توان همین رویکرد را برای پیش‌بینی بیماری‌های مزمن دیگر که با سبک زندگی و فعالیت بدنی مرتبط‌اند (مانند دیابت نوع ۲، بیماری‌های قلبی‌عروقی و آلزایمر) به کار گرفت. اشتراک‌گذاری کد، داده (در صورت امکان) و نتایج، می‌تواند به انباشت دانش و بهبود مداوم مدل‌های پیش‌بینی در حوزه سلامت و علوم ورزشی کمک کند.

هدف این پژوهش توسعه و ارزیابی یک چارچوب مبتنی بر یادگیری ماشین برای پیش‌بینی احتمال ابتلا به سرطان بر اساس شاخص‌های سبک زندگی و متغیرهای مرتبط با علوم ورزشی (مانند فعالیت بدنی و شاخص توده بدنی) بود. نتایج نشان داد که در صورت استاندارد بودن داده‌ها و انتخاب ویژگی‌های معنادار، مدل‌های یادگیری ماشین توانایی استخراج الگوهای پیچیده و غیرخطی را دارند و می‌توانند افراد در معرض خطر را با دقت قابل‌قبول از افراد کم‌خطر تفکیک کنند. در این مطالعه، الگوریتم‌های مبتنی بر یادگیری گروهی به‌ویژه گرادیان بوستینگ بهترین عملکرد را روی داده‌های آزمون ارائه کردند و در شاخص‌هایی مانند دقت، بازخوانی و مساحت زیر منحنی راک (در کنار عوامل خطر فردی، ژنتیکی و بالینی در سطح اثبات مفهوم)عملکرد بالاتری نسبت به برخی مدل‌های کلاسیک نشان دادند و این برتری در آزمون آماری دی‌لانگ نسبت به مدل پایه معنادار بود . از منظر کاربردی، دستاوردهای این پژوهش در مقطع فعلی جنبه اثبات مفهوم دارد. این چارچوب نشان می‌دهد که با مدل‌سازی ویژگی‌های سبک زندگی می‌توان الگوهای خطر را شناسایی کرد. با این حال، استفاده از چنین مدل‌هایی به‌عنوان ابزار پشتیبان تصمیم‌گیری در بالین یا غربالگری‌های واقعی، منوط به توسعه مدل روی داده‌های بزرگ‌ترِ بالینی و تأیید آن از طریق کارآزمایی‌ها و اعتبارسنجی‌های خارجیِ گسترده است. همچنین، از نگاه علوم ورزشی، مشاهده نقش متغیرهایی مانند فعالیت بدنی و شاخص توده بدنی در پیش‌بینی، نشان می‌دهد که (به‌ویژه در تحلیل حساسیت بدون سابقه سرطان که شاخص توده بدنی و فعالیت بدنی در مجموع یک بیش از ۳۳ درصد از اهمیت تصمیم‌گیری مدل را به خود اختصاص دادند) این متغیرهای قابل اصلاح نقش پیش‌بینی‌کننده‌ای در مدل دارند و ارتباط آن‌ها با وضعیت ابتلا شایسته بررسی بیشتر است؛ با این حال، این یافته‌ها به‌تنهایی اثربخشی مداخلات فعالیت بدنی در کاهش خطر سرطان را اثبات نمی‌کنند و ارزیابی چنین اثری نیازمند مطالعات طولی و کارآزمایی‌های مداخله‌ای است.

با وجود نتایج امیدوارکننده، این پژوهش محدودیت‌هایی نیز دارد؛ از جمله اتکا به یک مجموعه‌داده عمومی و نبود اعتبارسنجی بیرونی روی جمعیت‌های مستقل (لازم به ذکر است که مجموعه آزمون مستقل ۲۰ درصدی برای برآورد عملکرد نهایی مناسب است و اعتبارسنجی متقابل ۵-بخشی صرفاً برای انتخاب ابرپارامترها روی مجموعه آموزش به کار رفت؛ محدودیت اصلی، نبود اعتبارسنجی خارجی روی داده‌هایی از منبع و جمعیتی دیگر است)، شبیه‌سازی‌شده بودن داده‌ها، نامشخص بودن ترتیب زمانی متغیر سابقه شخصی سرطان نسبت به پیامد و احتمال جدی نشت اطلاعات از متغیر سابقه شخصی سرطان به دلیل نبود اطلاعات زمانی، محدودیت‌های مرتبط با تعمیم‌پذیری شاخص کالیبراسیون، وابستگی حساسیت و اختصاصیت به آستانه تصمیم انتخابی، و محدودیت توان آزمون‌های آماری مقایسه مدل‌ها در یک نمونه واحد شبیه‌سازی‌شده، محدود بودن دامنه متغیرهای اختصاصی علوم ورزشی (مانند شاخص‌های آمادگی جسمانی، داده‌های پوشیدنی و الگوهای تمرینی) و نیز انجام تحلیل‌ها بدون دسترسی به واحد پردازش

گرافیکی که استفاده از مدل‌های عمیق‌تر را محدود کرد. با این حال، همین محدودیت محاسباتی نشان داد که می‌توان با سخت‌افزارهای متعارف و الگوریتم‌های سبک‌تر نیز چارچوب‌های پیش‌بینی اولیه و قابل اجرا ساخت.

پیشنهاد می‌شود در پژوهش‌های آینده با افزایش حجم نمونه، جمع‌آوری داده‌های چندمرکزی و افزودن متغیرهای دقیق‌تر مرتبط با فعالیت بدنی و سبک زندگی، پایداری و تعمیم‌پذیری مدل‌ها بررسی شود و از روش‌هایی مانند اعتبارسنجی متقابل و اعتبارسنجی خارجی استفاده گردد. همچنین پیشنهاد می‌شود در پژوهش‌های آینده، ترتیب زمانی متغیرها و تاریخ دقیق ثبت آن‌ها نسبت به زمان تشخیص به‌دقت در قالب مطالعات طولی ثبت گردد تا امکان بررسی دقیق و پیشگیری از نشت اطلاعات فراهم شود، و کالیبراسیون مدل (شامل منحنی کالیبراسیون و امتیاز برایر) ، تحلیل آستانه تصمیم بالینی و آزمون‌های آماری رسمی مقایسه مدل‌ها به‌عنوان شاخص‌های اصلی ارزیابی در نظر گرفته شوند. علاوه بر آن، بررسی بیماری‌های دیگری مثل عوامل خطرساز قلبی‌عروقی، چاقی، پرفشاری خون و اختلالات شناختی مانند آلزایمر، می‌تواند ارزش کاربردی آن را در حوزه تربیت‌بدنی و سلامت عمومی افزایش دهد. در مجموع، یافته‌های این مطالعه نشان می‌دهد که در داده مورد مطالعه، مدل‌های یادگیری ماشین مبتنی بر پارامترهای کلیدی سبک زندگی و شاخص‌های ورزشی می‌توانند وضعیت ابتلا به سرطان را با عملکرد بالا پیش‌بینی کنند. این نتیجه در سطح اثبات مفهوم است و با توجه به استفاده از یک مجموعه‌داده عمومی شبیه‌سازی‌شده و نبود اعتبارسنجی خارجی، تعمیم این توان پیش‌بینی به جمعیت عمومی را نمی‌توان نتیجه‌ای قطعی دانست؛ کاربرد چنین مدلی برای مردم و نظام سلامت، منوط به توسعه و اعتبارسنجی آن روی داده‌های واقعی و مستقل است. بنابراین، با رعایت جانب احتیاط علمی، یافته‌های کمّی این پژوهش صرفاً یک اثبات مفهوم روش‌شناختی بر روی داده‌های شبیه‌سازی‌شده تلقی می‌شوند و هرگونه نتیجه‌گیری بالینی یا توصیه غربالگری در علوم ورزشی نیازمند بازتولید این چارچوب بر روی داده‌های طولی واقعی و بدون نشت اطلاعات است.

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش

این مطالعه با اهداف کاربردی و با رعایت کلیه دستورالعملهای پژوهشی و اصول اخلاقی در رابطه با شرکت‌کنندگان، از جمله رضایت آگاهانه، داوطلبانه، حق کناره‌گیری از پژوهش در صورت تمایل، و حفاظت از اطلاعات محرمانه آزمودنی‌ها، انجام پذیرفته است .

مشارکت نویسندگان

نویسندگان این پژوهش در کلیه مراحل اجرای پروژه مشارکت یکسانی داشته‌اند .

حامی مالی

در طول این پژوهش هیچگونه کمک مالی از منابع تأمین مالی در بخش عمومی، تجاری، و غیردولتی دریافت نشد .

تعارض منافع

بنابر اظهار نویسندگان این مقاله، تعارض منافع وجود نداشته است .

سپاسگزاری

از کلیه شرکت‌کنندگان که در اجرای این پژوهش همکاری داشته‌اند، تشکر و قدردانی می‌گردد.

References

Abdullah Alfayez, A., Kunz, H., & Grace Lai, A. (2021). Predicting the risk of cancer in adults using supervised machine learning: a scoping review. *BMJ open, 11*(9), e047755.

Ahmed, M. A., AbdelMoety, A., & Soliman, A. M. A. (2025). Predicting cancer risk using machine learning on lifestyle and genetic data. *Scientific reports, 15*(1), 30458.

Alpaydin, E. (2020). *Introduction to machine learning*: MIT press.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4): Springer.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Brown, J. S., Amend, S. R., Austin, R. H., Gatenby, R. A., Hammarlund, E. U., & Pienta, K. J. (2023). Updating the definition of cancer. *Molecular Cancer Research, 21*(11), 1142–1147.

Burtscher, J., Millet, G. P., & Burtscher, M. (2023). Pushing the limits of strength training. *American journal of preventive medicine, 64*(1), 145–146.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21–27.

DeVita, V. T., Lawrence, T. S., & Rosenberg, S. A. (2008). *DeVita, Hellman, and Rosenberg's cancer: principles & practice of oncology* (Vol. 2): Lippincott Williams & Wilkins.

Du, G., Lv, H., Liang, Y., Zhang, J., Huang, Q., Xie, G., . . . Ye, J. (2025). Population-based colorectal cancer risk prediction using a SHAP-enhanced LightGBM model. *Frontiers in oncology, 15*, 1575844.

Felici, A., Peduzzi, G., Pellungrini, R., & Campa, D. (2025). Artificial intelligence to predict cancer risk, are we there yet? A comprehensive review across cancer types. *European Journal of Cancer, 222*, 115440.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Hastie, T. (2005). The elements of statistical learning data mining, inference, and prediction.

Kumar, V., Abbas, A. K., Fausto, N., & Aster, J. C. (2014). *Robbins and Cotran pathologic basis of disease, professional edition e-book*: Elsevier health sciences.

McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification.* Paper presented at the AAAI-98 workshop on learning for text categorization.

McGuire, S. (2016). World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO Press, 2015. *Advances in nutrition, 7*(2), 418–419.

Moore, S. C., Lee, I.-M., Weiderpass, E., Campbell, P. T., Sampson, J. N., Kitahara, C. M., . . . Hartge, P. (2016). Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults. *JAMA internal medicine, 176*(6), 816–825.

Patel, A. V., Friedenreich, C. M., Moore, S. C., Hayes, S. C., Silver, J. K., Campbell, K. L., . . . Fulton, J. E. (2019). American College of Sports Medicine roundtable report on physical activity, sedentary behavior, and cancer prevention and control. *Medicine and science in sports and exercise, 51*(11), 2391.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825–2830.

RE., K. (2025). Cancer Prediction Dataset. Retrieved from https://wwwkagglecom/datasets/rabieelkharoua/cancer-prediction-dataset. [El Kharoua, R. (2024). Cancer Prediction Dataset (Version 2). Kaggle. https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset]

Shreves, A. H., Small, S. R., Walmsley, R., Chan, S., Saint-Maurice, P. F., Moore, S. C., . . . Matthews, C. E. (2025). Amount and intensity of daily total physical activity, step count and risk of incident cancer in the UK Biobank. *British journal of sports medicine, 59*(12), 839–847.

Steyerberg, E. W. (2019). *Clinical prediction models* (Vol. 201): Springer.

Wu, X., Tu, H., Hu, Q., Tsai, S. P., Chu, D. T.-W., & Wen, C.-P. (2024). Novel machine learning algorithm in risk prediction model for pan-cancer risk: application in a large prospective cohort. *BMJ oncology, 3*(1), e000087.