Analyzing the Performance of Explainable Machine Learning Models in Risk Factor Identification for Major Cancers

Xiayuan Huang

Submitted to: JMIR Cancer
on: June 02, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Analyzing the Performance of Explainable Machine Learning Models in Risk Factor Identification for Major Cancers

Xiayuan Huang[1] Dr

[1]Yale University New Haven US

Corresponding Author:
Xiayuan Huang Dr
Yale University
100 College Street
New Haven
US

Abstract

Background: Cancer is a life-threatening disease and a leading cause of death worldwide, with an estimated 611,000 deaths and over 2 million new cases in the United States in 2024. The rising incidence of major cancers, including among younger individuals, highlights the need for early screening and monitoring of risk factors to manage and decrease cancer risk.

Objective: To identify pivotal factors essential for predicting the risk factors for four major cancer types (breast, colorectal, lung, and prostate) through the utilization of explainable machine learning techniques is imperative due to the increasing burden of cancer patients.

Methods: De-identified electronic health record data from MIMIC-III was used to identify patients with four types of cancer who had longitudinal hospital visits prior to receiving a cancer diagnosis. Their records were matched and combined with those of patients without cancer diagnoses using propensity scores based on demographic factors. Three advanced models, penalized Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP), were conducted to identify the rank of risk factors for each cancer type, with feature importance analysis for RF and MLP models. The Rank Biased Overlap was adopted to compare the similarity of ranked risk factors across cancer types.

Results: Our framework evaluated the prediction performance of explainable ML models, in which MLP achieved an AUC of 0.78 for breast cancer, 0.76 for colorectal cancer, 0.84 for lung cancer, and 0.78 for prostate cancer, respectively. In addition to demographic risk factors, the most prominent non-traditional risk factors overlapped across models and cancer types, including hyperlipidemia, diabetes, depressive disorders, heart diseases, and anemia. The similarity analysis indicated the unique risk factor pattern for lung cancer from other cancer types.

Conclusions: The study's findings demonstrate the effectiveness of explainable ML models in predicting non-traditional risk factors for major cancers and highlight the importance of considering unique risk profiles for different cancer types. These insights may contribute to efficient cancer screening and tailored cancer prevention strategies, which, in turn, offer fundamental support for clinical decision-making processes.

(JMIR Preprints 02/06/2024:62833)
DOI: https://doi.org/10.2196/preprints.62833

Preprint Settings

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Analyzing the Performance of Explainable Machine Learning Models in Risk Factor Identification for Major Cancers

Xiayuan Huang, PhD[1], Shushun Ren, BS[2], Elle Chen[2], Yuqi He, PhD[3], Yun Jiang, PhD, MS, RN, FAMIA[2]

1. Yale University School of Public Health, New Haven, CT
2. University of Michigan School of Nursing, Ann Arbor, MI
3. San Jose State University, King Library, San Jose, CA

Corresponding Authors:
Yun Jiang, Ph.D., MS, RN, FAMIA, University of Michigan School of Nursing, 400 North Ingalls Street, Ann Arbor, MI 48109, USA. Tel: +1 734-763-3705. Fax: +1 734-647-2416. Email: jiangyu@umich.edu
Xiayuan Huang, PhD, Yale University School of Public Health, 100 College Street, New Haven, CT 06510, USA. Email: xiayuan.huang@yale.edu

Abstract

Background

Cancer is a life-threatening disease and a leading cause of death worldwide, with an estimated 611,000 deaths and over 2 million new cases in the United States in 2024. The rising incidence of major cancers, including among younger individuals, highlights the need for early screening and monitoring of risk factors to manage and decrease cancer risk.

Objective

To identify pivotal factors essential for predicting the risk factors for four major cancer types (breast, colorectal, lung, and prostate) through the utilization of explainable machine learning techniques is imperative due to the increasing burden of cancer patients.

Methods

De-identified electronic health record data from MIMIC-III was used to identify patients with four types of cancer who had longitudinal hospital visits prior to receiving a cancer diagnosis. Their records were matched and combined with those of patients without cancer diagnoses using propensity scores based on demographic factors. Three advanced models, penalized Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP), were conducted to identify the rank of risk factors for each cancer type, with feature importance analysis for RF and MLP models. The Rank Biased Overlap was adopted to compare the similarity of ranked risk factors across cancer types.

Results

Our framework evaluated the prediction performance of explainable ML models, in which MLP achieved an AUC of 0.78 for breast cancer, 0.76 for colorectal cancer, 0.84 for lung cancer, and 0.78

for prostate cancer, respectively. In addition to demographic risk factors, the most prominent non-traditional risk factors overlapped across models and cancer types, including hyperlipidemia, diabetes, depressive disorders, heart diseases, and anemia. The similarity analysis indicated the unique risk factor pattern for lung cancer from other cancer types.

Conclusion

The study's findings demonstrate the effectiveness of explainable ML models in predicting non-traditional risk factors for major cancers and highlight the importance of considering unique risk profiles for different cancer types. These insights may contribute to efficient cancer screening and tailored cancer prevention strategies, which, in turn, offer fundamental support for clinical decision-making processes.

1 INTRODUCTION

Cancer is a lift-threatening disease and leading cause of death worldwide. In 2024, an estimated 611,000 people will die from cancer in the United States, and the estimated new cancer cases will reach more than 2 million for the first time.[1] This surge includes rising incidence rates for major cancers, including breast, prostate, lung, and colorectal cancers, which display the trend of increasingly affecting younger individuals who have many more years of life expectancy.[1] The U.S. Preventive Services Task Force modified the recommended age for colorectal cancer screening from 50 to 45 years for people at average risk in 2021[2] and adjusted the recommendation for breast cancer screening for all women to start at age 40 in 2024[3]. Similar upward trends in the incidence of early-onset cancers are observed in other developed countries, suggesting shared risk factors and exposures across these regions. However, besides those uncontrollable risk factors, such as previous cancer diagnosis, family history of cancer, and genetics or inherited cancer syndrome, many cancer risk factors, including lifestyle factors, are modifiable and can be managed to decrease people’s risk for cancer.[4]

Extensive evidence has shown the significant advantages of early identification of people at high risk for cancer, leading to improved cancer prevention and control to maximize treatment benefits, reduce cancer burden, and improve long-term survival.[5] In the context of breast cancer, it was estimated that early access to treatment services following breast cancer screening could have reduced breast cancer mortality by 25-40%.[6] Given the tremendous benefits of early identification of high-risk patients, an increasing number of cancer risk prediction models have been developed. However, traditional risk factor-based models, relying on methods like logistic regression or Cox regression, have low discrimination accuracy with the area under the receiver operating characteristic curve (AUC) between 0.53 and 0.64.[7] Some models heavily rely on family history and lack generalizability, and others can be biased when applied to specific subpopulations[8,9] Moreover, non-traditional factors, such as chronic diseases, are not usually included in the models, although chronic conditions are believed to raise cancer risk as much as lifestyle does[10]. Therefore, new methods and models are urgently needed to improve cancer risk predictions and facilitate the development of effective cancer prevention strategies.

Machine learning has demonstrated significant promise in cancer prediction by leveraging electronic health records (EHRs) data to identify potential risks.[11] Current applications include the development of predictive models for early cancer detection, personalized treatment recommendations, and outcome prediction based on diverse patient characteristics and biomarkers. However, machine learning in cancer prediction still faces several limitations[12]. One major challenge is the need for a comprehensive understanding of risk factors within and across cancer types[13]. As machine learning research delves deeper, the utilization of explainable machine learning marks a significant advancement in enhancing the efficacy of cancer prediction models[14-16]. The development and application of explainable machine learning not only provide accurate predictions or classifications but also offer insights into how those predictions are made in a transparent and interpretable manner[17]. In the medical context, explainable machine learning is particularly important because it allows healthcare professionals to understand the reasoning behind a model's predictions, which is crucial for trust, acceptance, and decision-making in clinical settings. By systematically discerning and excluding extraneous features, this approach holds the potential to mitigate unwanted noise and streamline the predictive process. However, it is noteworthy that feature selection algorithms frequently exhibit sensitivity to dataset characteristics, with minor fluctuations in the data yielding divergent outcomes[18]. Thus, the imperative task of identifying a subset of features that are most pertinent becomes paramount. This endeavor not only fosters a deeper comprehension of the dataset but also illuminates the comprehensive understanding of cancer, thereby enriching our knowledge and enhancing predictive accuracy.

Hence, this study presents comprehensive research aimed at uncovering the pivotal factors essential for predicting the risks of four major cancer types (breast, prostate, lung, and colorectal) through the utilization of explainable machine learning techniques on penalized Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP). Our primary objective is to pinpoint the significant features that exert an influence on the risks associated with these major cancers and to delineate the patterns of risk factors corresponding to each cancer type. Such insights hold immense potential in risk monitoring, cancer prevention, and improving early diagnosis, offering valuable guidance to clinicians in clinical decision-making support. By elucidating these critical factors and their associated risk factor patterns, we endeavor to provide clinicians with rigorous analysis for enhancing risk monitoring and patient care across various cancer types.

2 METHODS

2.1 Experimental Dataset

Our study was conducted using data from MIMIC-III, a comprehensive, structured, longitudinal EHR dataset that is publicly available[19]. This dataset contains de-identified, detailed clinical data from ICU admissions at Beth Israel Deaconess Medical Center in Boston, Massachusetts, and is accessible to the global research community under a data use agreement. We used the most recent version (v1.4) for this work which contains a broad spectrum of data, including information on individual patients' health and healthcare from various inpatient and outpatient visits, such as diagnoses, prescriptions, lab tests, and procedures. In total, this dataset contains 58,976 admissions of 46,520 patients.

2.2 Data Preprocessing

We included patients with four types of cancers (breast, colorectal, lung, and prostate) identified using ICD-9 codes associated with the diagnosis of each type of cancer (see Appendix **Table A1**).

We took a few steps to preprocess the experimental dataset, starting with the consolidation of three main tables from the MIMIC III database. These included: (1) foundational patient information, capturing demographics and initial hospital admission data; (2) a reference table for ICD9 Codes, detailing both codes and corresponding diagnostic labels; and (3) logs of patient visit sequences with associated ICD9 Codes. This consolidation linked the records via patient IDs to form a detailed longitudinal dataset. **Figure 1** illustrates the data processing workflow of this study. Patients' ages were determined by deducting their date of birth from their initial hospital admission date, with the result rounded to the nearest year. Any patient records missing demographic details (such as ethnicity, marital status, or religion) were omitted, narrowing the dataset to a total of 21,372 unique individuals. Our study focused on patients who had multiple hospital visits prior to receiving a cancer diagnosis to identify potential risk factors. After a cancer diagnosis was recognized, further visits were disregarded. These records were combined with those of patients without a cancer diagnosis. A label was created as 1 if a visit included an ICD-9 code for a cancer diagnosis and 0 if not. To ensure a balanced dataset in terms of cancer diagnosis, the study matched patients diagnosed with cancer with those without cancer using propensity score matching based on demographic factors. See Appendix **Table A2** for a detailed description of patient characteristics for four cancer types.

2.3 Methods

In this work, we applied three advanced models, penalized LR, RF, and MLP, based on their demonstrated accuracy and robustness in handling high-dimensional datasets. RF and MLP, in particular, excel at identifying complex, non-linear interactions among variables without requiring predefined interaction terms. This capability is crucial for analyzing interactions between risk factors and cancer outcomes. Our choice of RF and MLP was determined by a desire to balance complexity with interpretability, as well as to ensure computational efficiency. Both methods are straightforward and offer high interpretability, which makes them excellent foundational models for exploring how different features influence cancer risk.

Since the task aimed at forecasting cancer risk by considering important and relevant risk factors, we evaluated the efficacy of our methodologies by employing several critical performance metrics: AUC, accuracy, specificity, sensitivity, and the F1 score for each model. We partitioned the dataset into three sections for model development: 70% for training, 10% for validation, and 20% for testing. The model that exhibited the best results on the validation set was further subjected to an in-depth analysis on the test set, utilizing a 3-fold cross-validation technique to calculate its AUC precisely. To enhance our understanding of how our machine learning models contribute to cancer prevention, we also quantified the impact of each feature on the prediction of four cancer types. We then ranked these features according to their significance. All statistical analyses and model implementations were coded using Python, with the scikit-learn library serving as the foundation for our predictive framework[20].

To investigate the similarity of features ranking by different cancer types, we applied Rank Biased Overlap (RBO)[21], a similarity measure of two ranked lists. The RBO score ranges between 0 and 1, where a higher score indicates greater similarity between the lists. A score of 1 implies perfect overlap, meaning the two lists are identical in both order and content. On the other hand, a score of 0 suggests no overlap between the lists.

Mathematically, let x_i be the high-dimensional feature input. Let $y_i \in \{0,1\}$ be the corresponding label. $y_i = 0$ means not affected, and $y_i = 1$ means affected. Our goal is to learn a predictive function f that best classifies the data. We built three state-of-the-art models for four cancer types respectively in this study:

- Penalized Logistic Regression: given M training instances, we considered L1 regularized logistic regression by minimizing the following function: $\min_{\theta} \sum_{i=1}^{M} -\log p\left(y^{(i)} \middle| x^{(i)};\theta\right) + \beta ¿ \vee \theta \vee ¿_1$.
- Random Forest[22]: a robust ensemble learning method that constructs multiple decision trees during training to improve prediction accuracy and prevent overfitting, where f is the decision tree as base learners. The RF model was trained by iteratively selecting features from root to leaf nodes and aggregating multiple trees with the weights from a subset of the training instances. The nodes and the weights in the model reflect their importance to the final prediction.
- Multilayer Perceptron[23]: a type of artificial neural network that consists of at least three layers of nodes: an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, in one layer, connects with a certain weight to every node in the following layer, and nodes do not connect within the same layer. The non-linear activation functions, such as the sigmoid, or ReLU (Rectified Linear Unit), are applied to the weighted sum of inputs to a neuron, determining its output signal.

To rank the impact on predictive models of the features, relative to all three models, we used a permutation importance score[24] to rank all features in the training models for MLP. The scores were defined by the mean decrease in accuracy of the trained model when each feature was permuted.

3 RESULTS

3.1 Feature Selection

Our experiment's initial dataset comprised thousands of diagnosis codes intended for predicting cancer risk. Aware of some features' potential redundancy and less informative nature, we did a feature selection process. This involved assessing the relevance and importance of each feature in relation to four specific types of cancer. Through this rigorous analysis, we aimed to distill the dataset down to a more manageable and meaningful subset of features. After careful consideration and evaluation, we identified 33 features (re-categorized into 20 factors for further analysis, see **Table 1**) that emerged as particularly crucial for accurately predicting cancer risk. These features were meticulously curated, ensuring that only the most informative and pertinent variables were retained for our predictive models.

3.2 Model Performance

For each predicted cancer outcome, we carried out the experiment by predicting cancer using the entire diagnosis history of the patient by building LR, RF and MLP models. **Table 2** illustrates the accuracy, specificity, sensitivity, and F1 score of these three models for breast, colorectal, lung, and prostate cancers. **Figure 2** shows the ROC plots of three models for four types of cancer, respectively. Both Table 2 and Figure 2 show that within the three models, MLP performs the best, RF falls in the middle, and LR ranks last. It is worth noting that MLP achieved an AUC of 0.78 for breast cancer, 0.76 for colorectal cancer, 0.84 for lung cancer, and 0.78 for prostate cancer, demonstrating a higher AUC over traditional risk factor-based models and a statistically significant superiority over random chance. The underwhelming results from the LR model led us to investigate the complexity of risk factors for prediction. Compared to LR, MLP reveals the intricate, non-linear associations between risk factors and the likelihood of cancer, offering meaningful insights into the collective influence of these risk factors on cancer risk.

3.3 Feature Importance Analysis

We analyzed the feature importance for each cancer type further to investigate the potential impact of risk factors on cancer. **Table 3** presents the feature importance analysis of RF and MLP, showcasing the top-ranked risk factors for each type of cancer. The ranks of these factors were relatively different by model and cancer type, although some consistency can be observed across cancer types. Age emerged as the top risk factor across all four types of cancer; race/ethnicity ranked among the top 10 factors for all cancers from all models except for the RF-based lung cancer and prostate cancer models; gender was ranked among the top 10 in MLP-based models but not in any RF-based models; marital status and religion were presented for some types of cancer in some of the models, and tobacco use as an important factor for lung and prostate cancer patients exclusively. However, all these demographic risk factors were included in the top 20 factors for all cancer types (see Appendix **Table A3**). Similarly, RF-based models identified hypertension, heart diseases, respiratory/pulmonary diseases, and acute kidney failure as the common top risk factors for all types of cancers, while MLP-based models highlighted hyperlipidemia, diabetes, depressive disorder, and heart diseases. In MLP-based models, respiratory/pulmonary diseases and acute kidney failure were only presented as the top 10 for lung cancer. Both RF and MLP-based models pinpointed anemia as the top risk for breast cancer. **Figure 3** shows the RBO similarity scores of risk factors for four types of cancer according to MLP-based models. Low similarity scores are presented between lung cancer and any other three cancer types, all around 0.58, suggesting distinct patterns of risk factors associated with lung cancer. Risk factors for breast and prostate cancers show the most similar ranking with an RBO similarity score of 0.76. A moderate similarity score between colorectal and breast cancers is about the same as the score between colorectal and prostate cancer, both around 0.70.

4 DISCUSSION

This study used comprehensive patient diagnosis histories to evaluate the efficacy of penalized LR, RF, and MLP models in predicting cancer risks. The analysis identified the top-ranking risk factors, including non-traditional risk factors such as the diagnoses of hyperlipidemia, diabetes, depressive disorders, heart diseases, and anemia, in addition to demographic factors such as age, sex, race/ethnicity, for the most prevalent four types of cancer, including breast, colorectal, lung, and prostate cancers. The model performance evaluation revealed the significant potential of neural network-based models, especially MLPs, in oncology for predicting cancer risks across cancer types. Demonstrating superior capability to model complex, non-linear interactions among diverse risk factors, MLPs emerge as crucial tools for cancer early detection and intervention. This advantage is particularly important given the model's capacity to integrate and interpret the intricate relationships between clinical factors present in EHRs. In contrast to simpler models like LR, which struggle with the multidimensional nature of cancer risk factors, MLPs offer a more detailed and comprehensive analysis, enhancing our understanding of how these factors collectively impact cancer risk and improving the precision of preventive strategies in clinical settings.

Understanding the relationships between various risk factors and cancer risk is pivotal for the early detection and prevention of cancer. In this context, our feature importance analysis using RF and MLP models pinpointed critical risk factors for different cancer types and explored patterns of these risk factors across various cancers. Although the ranks of risk factors for cancers were slightly different by the RF and MLP-based models, similar patterns were presented among the top 10 factors (Table 3), which are interpretable and supported by the literature. Both models highlighted age as the predominant risk factor across all four types of cancer, which is evident that as age increases, the incidence rates for cancer overall climb steadily[25], and alongside age, demographic variables such as gender, race/ethnicity, marital status, and religion emerged within the top 10 features. Racial/ethnic disparities in cancer incidence and outcomes are well-known[26]. Although there may not be existing evidence to confirm that marital status is an independent risk factor for cancer, observational studies demonstrate that married status is associated with reduced risk of cancer-specific and all-cause mortality[27,28]. Religion and spirituality are important in patient cancer care, and specifically, a systematic review suggests a positive association between religious attendance and cancer screening utilization[29]. Our models not only confirmed the significance of these risk factors for each cancer type but also our RF-based model facilitated an interpretable analysis, allowing us to clearly rank the significance of each risk factor, while the MLP-based model provided deeper insights into complex, non-linear interactions among the risk factors. This approach enriches our understanding of how specific risk factors influence cancer risk, enhancing the potential for developing tailored intervention strategies that address the unique risk profiles associated with different cancer types and potentially shared risk patterns across prevalent cancer types.

Chronic diseases are often overlooked as risk factors for cancer, and they are not often targeted in cancer prevention strategies. As non-traditional risk factors, the influence of these conditions on cancer has been brought to researchers' attention in the past decade. A prospective cohort study with 405,878 participants followed for an average of 8.7 years demonstrated eight common chronic diseases accounted for more than 20% of cancer risk, which are comparable to five major lifestyle factors, such as smoking and lack of physical activity.[28] These eight chronic diseases or markers included blood pressure, total cholesterol, heart rate, diabetes, proteinuria, glomerular filtration rate, pulmonary disease, and gouty arthritis marker.[10] However, as these diseases or markers were pre-selected by the researchers based on their disease burden worldwide, some other essential influential conditions might be missed. Our models confirmed most of these eight diseases as the top-ranking risk factors. Additionally, some new conditions were revealed in our models among the top 10 factors for four types of cancer, such as depressive disorder, anemia, hypothyroidism, sepsis, urinary tract infection, and acidosis, which encourages further exploration. Notably, tobacco usage and respiratory/pulmonary diseases emerged as pivotal risk factors, specifically for lung cancer, which is not surprising based on our knowledge in the field. Diabetes and anemia were highlighted as significant risk factors for colorectal cancer, which is congruent with the literature [30, 31]. Iron deficiency has been recognized long-term as an independent predictor of colorectal cancer, which may be due to chronic blood loss from the gastrointestinal tract and the inflammation associated with malignancy[32,33]. These conditions may have shared risk factors with cancer. However, emerging evidence implies that they may have more complicated relationships, including shared pathophysiological mechanisms that need further exploration[34]. Moreover, cancer prevention strategies should consider the impact of comorbid conditions on the incidence of cancer and particularly their joint impact on cancer risk.[28]

The analysis of the similarity among risk factors for four types of cancer also revealed interesting findings. As breast and prostate cancer are both hormone-dependent cancers, it is understandable that their importance-ranked risk factors share a high level of similarity. However, lung cancer had more unique ranked risk factors than other types of cancer, which may be because lung cancer is more sensitive to environmental risk factor exposure. Notably, acute kidney failure was presented among the top 10 factors for lung cancer only. Previous studies have reported the incidence of acute kidney injury is higher in lung cancer patients than in other malignancies[35], and it is believed that acute kidney injury can negatively affect lung physiology by altering fluid balance, acid-base balance, and vascular tone[36]. The presence of albuminuria has been found to be associated with the incidence of lung cancer in large-size observational studies[37],[38]. The findings from our analysis underscore the heterogeneous nature of cancer and highlight the importance of considering unique risk profiles for different cancer types. This also urges us to address the fundamental mechanism of risk factors leading to cancers. Such insights are crucial for developing tailored prevention strategies, optimizing screening protocols, and informing personalized treatment approaches to mitigate the burden of lung cancer and improve patient outcomes.

Using the MIMIC-III dataset in the study on explainable machine learning for cancer risk prediction introduces certain limitations that could affect the generalizability of the findings. To enhance the robustness of future research, integrating more recent and varied data sources and validating findings across different cohorts are essential steps. Another limitation comes from the application of explainable machine learning models for cancer risk prediction. Employing advanced techniques like penalized LR, RF, and MLP, this research seeks to optimize predictive accuracy. However, each model inherently embodies trade-offs: while more complex models, such as multi-layer perceptrons, may enhance performance, they often compromise on interpretability. This presents significant challenges in clinical settings, where understanding the reasoning behind model predictions is crucial for acceptance and trust by medical practitioners. Ensuring that these advanced models can convey their decision-making process in a transparent and comprehensible manner remains a key hurdle in bridging the gap between machine learning capabilities and practical clinical application.

In conclusion, our study established a predictive framework using EHR data to assess the efficacy of explainable ML models in predicting the risk of major cancer types. We reported critical non-traditional chronic condition risk factors in addition to common demographic risk factors and outlined distinct patterns for each of the four cancer types studied. Additionally, we explored the similarities and differences in risk factor patterns across these cancers. These insights enhance understanding of cancer prevention strategies and improve early diagnosis, providing valuable support for clinical decision-making.

Author contributions

X.H. and Y.J. conceived the study. X.H. and S.R. implemented the algorithm, conducted the experiments, and performed all the analyses. S.R. generated results visualization. X.H. and Y.J. supervised the study. X.H., S.R., E.C., Y.H. and Y.J. wrote the manuscript. All authors provided feedback and approved the manuscript.

REFERENCES

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. Jan-Feb 2024;74(1):12-49. doi:10.3322/caac.21820

2. Colorectal Cancer: Screening. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/colorectal-cancer-screening

3. Breast Cancer: Screening. 2024;

4. Mansour R, Al-Ani A, Al-Hussaini M, Abdel-Razeq H, Al-Ibraheem A, Mansour AH.

Modifiable risk factors for cancer in the middle East and North Africa: a scoping review. *BMC Public Health*. Jan 18 2024;24(1):223. doi:10.1186/s12889-024-17787-5

5. Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. *Nat Med*. Apr 2022;28(4):666-677. doi:10.1038/s41591-022-01746-x

6. Duffy SW, Tabar L, Yen AM, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer*. Jul 1 2020;126(13):2971-2979. doi:10.1002/cncr.32859

7. Gao Y, Li S, Jin Y, et al. An Assessment of the Predictive Performance of Current Machine Learning-Based Breast Cancer Risk Prediction Models: Systematic Review. *JMIR Public Health Surveill*. Dec 29 2022;8(12):e35750. doi:10.2196/35750

8. Guan Z, Huang T, McCarthy AM, et al. Combining Breast Cancer Risk Prediction Models. *Cancers (Basel)*. Feb 8 2023;15(4)doi:10.3390/cancers15041090

9. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis*. Mar 2019;11(Suppl 4):S574-S584. doi:10.21037/jtd.2019.01.25

10. Tu H, Wen CP, Tsai SP, et al. Cancer risk associated with chronic diseases and disease markers: prospective cohort study. *BMJ*. Jan 31 2018;360:k134. doi:10.1136/bmj.k134

11. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform*. Sep 2018;22(5):1589-1604. doi:10.1109/JBHI.2017.2767063

12. Wang R, Li Z, Liu S, Zhang D. Global, regional and national burden of inflammatory bowel disease in 204 countries and territories from 1990 to 2019: a systematic analysis based on the Global Burden of Disease Study 2019. *BMJ Open*. Mar 28 2023;13(3):e065186. doi:10.1136/bmjopen-2022-065186

13. Steinberg J, Yap S, Goldsbury D, et al. Large-scale systematic analysis of exposure to multiple cancer risk factors and the associations between exposure patterns and cancer incidence. *Sci Rep*. Jan 27 2021;11(1):2343. doi:10.1038/s41598-021-81463-6

14. Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. *Front Big Data*. 2021;4:688969. doi:10.3389/fdata.2021.688969

15. Gurmessa DK, Jimma W. Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review. *BMJ Health Care Inform*. Feb 2 2024;31(1)doi:10.1136/bmjhci-2023-100954

16. Shulha M, Hovdebo J, D'Souza V, Thibault F, Harmouche R. Integrating Explainable Machine Learning in Clinical Decision Support Systems: Study Involving a Modified Design Thinking Approach. *JMIR Form Res*. Apr 16 2024;8:e50475. doi:10.2196/50475

17. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep*. Mar 26 2021;11(1):6968. doi:10.1038/s41598-021-86327-7

18. Huang P, Kong Z, Wang L, Han X, Yang X. Efficient and Stable Unsupervised Feature Selection Based on Novel Structured Graph and Data Discrepancy Learning. *IEEE Trans Neural Netw Learn Syst*. Apr 15 2024;PPdoi:10.1109/TNNLS.2024.3385838

19. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. May 24 2016;3:160035. doi:10.1038/sdata.2016.35

20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. Oct 2011;12:2825-2830.

21. Sarica A, Quattrone A, Quattrone A. Introducing the Rank-Biased Overlap as Similarity Measure for Feature Importance in Explainable Machine Learning: A Case Study on Parkinson's Disease. *Lect Notes Artif Int*. 2022;13406:129-139. doi:10.1007/978-3-031-15037-1_11

22. Breiman L. Random forests. *Mach Learn*. Oct 2001;45(1):5-32. doi:Doi 10.1023/A:1010933404324

23. Rosenblatt F. The Perceptron - a Probabilistic Model for Information-Storage and Organization in the Brain. *Psychol Rev*. 1958;65(6):386-408. doi:DOI 10.1037/h0042519

24. Nirmalraj S, Antony ASM, Srideviponmalar P, et al. Permutation feature importance-based fusion techniques for diabetes prediction. *Soft Comput*. Apr 24 2023;doi:10.1007/s00500-023-08041-y

25. Institute NC. Age and Cancer Risk was originally published by the National Cancer Institute.

26. Zavala VA, Bracci PM, Carethers JM, et al. Cancer health disparities in racial/ethnic minorities in the United States. *Br J Cancer*. Jan 2021;124(2):315-332. doi:10.1038/s41416-020-01038-6

27. Zhu S, Lei C. Association between marital status and all-cause mortality of patients with metastatic breast cancer: a population-based study. *Sci Rep*. Jun 5 2023;13(1):9067. doi:10.1038/s41598-023-36139-8

28. Chen ZH, Yang KB, Zhang YZ, et al. Assessment of Modifiable Factors for the Association of Marital Status With Cancer-Specific Survival. *JAMA Netw Open*. May 3 2021;4(5):e2111813. doi:10.1001/jamanetworkopen.2021.11813

29. Kretzler B, Konig HH, Brandt L, Weiss HR, Hajek A. Religious Denomination, Religiosity, Religious Attendance, and Cancer Prevention. A Systematic Review. *Risk Manag Healthc Policy*. 2022;15:45-58. doi:10.2147/RMHP.S341085

30. Stan MC, Georgescu D, Mirestean CC, Badulescu F. Cancer and Diabetes: Predictive Factors in Patients with Metabolic Syndrome. *Diagnostics (Basel)*. Aug 11 2023;13(16)doi:10.3390/diagnostics13162647

31. Soltani G, Poursheikhani A, Yassi M, Hayatbakhsh A, Kerachian M, Kerachian MA. Obesity, diabetes and the risk of colorectal adenoma and cancer. *BMC Endocr Disord*. Oct 29 2019;19(1):113. doi:10.1186/s12902-019-0444-6

32. Chardalias L, Papaconstantinou I, Gklavas A, Politou M, Theodosopoulos T. Iron Deficiency Anemia in Colorectal Cancer Patients: Is Preoperative Intravenous Iron Infusion Indicated? A Narrative Review of the Literature. *Cancer Diagn Progn*. Mar-Apr 2023;3(2):163-168. doi:10.21873/cdp.10196

33. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng KK, Marshall T. The importance of anaemia in diagnosing colorectal cancer: a case-control study using electronic primary care records. *Br J Cancer*. Jan 29 2008;98(2):323-7. doi:10.1038/sj.bjc.6604165

34. de Boer RA, Meijers WC, van der Meer P, van Veldhuisen DJ. Cancer and heart disease: associations and relations. *Eur J Heart Fail*. Dec 2019;21(12):1515-1525. doi:10.1002/ejhf.1539

35. Park N, Kang E, Park M, et al. Predicting acute kidney injury in cancer patients using heterogeneous and irregular data. *PLoS One*. 2018;13(7):e0199839. doi:10.1371/journal.pone.0199839

36. Cho S, Kang E, Kim JE, et al. Clinical Significance of Acute Kidney Injury in Lung Cancer Patients. *Cancer Res Treat*. Oct 2021;53(4):1015-1023. doi:10.4143/crt.2020.1010

37. Mok Y, Ballew SH, Sang Y, et al. Albuminuria, Kidney Function, and Cancer Risk in the Community. *Am J Epidemiol*. Sep 1 2020;189(9):942-950. doi:10.1093/aje/kwaa043

38. Lees JS, Ho F, Parra-Soto S, et al. Kidney function and cancer risk: An analysis using creatinine and cystatin C in a cohort study. *EClinicalMedicine*. Aug 2021;38:101030. doi:10.1016/j.eclinm.2021.101030

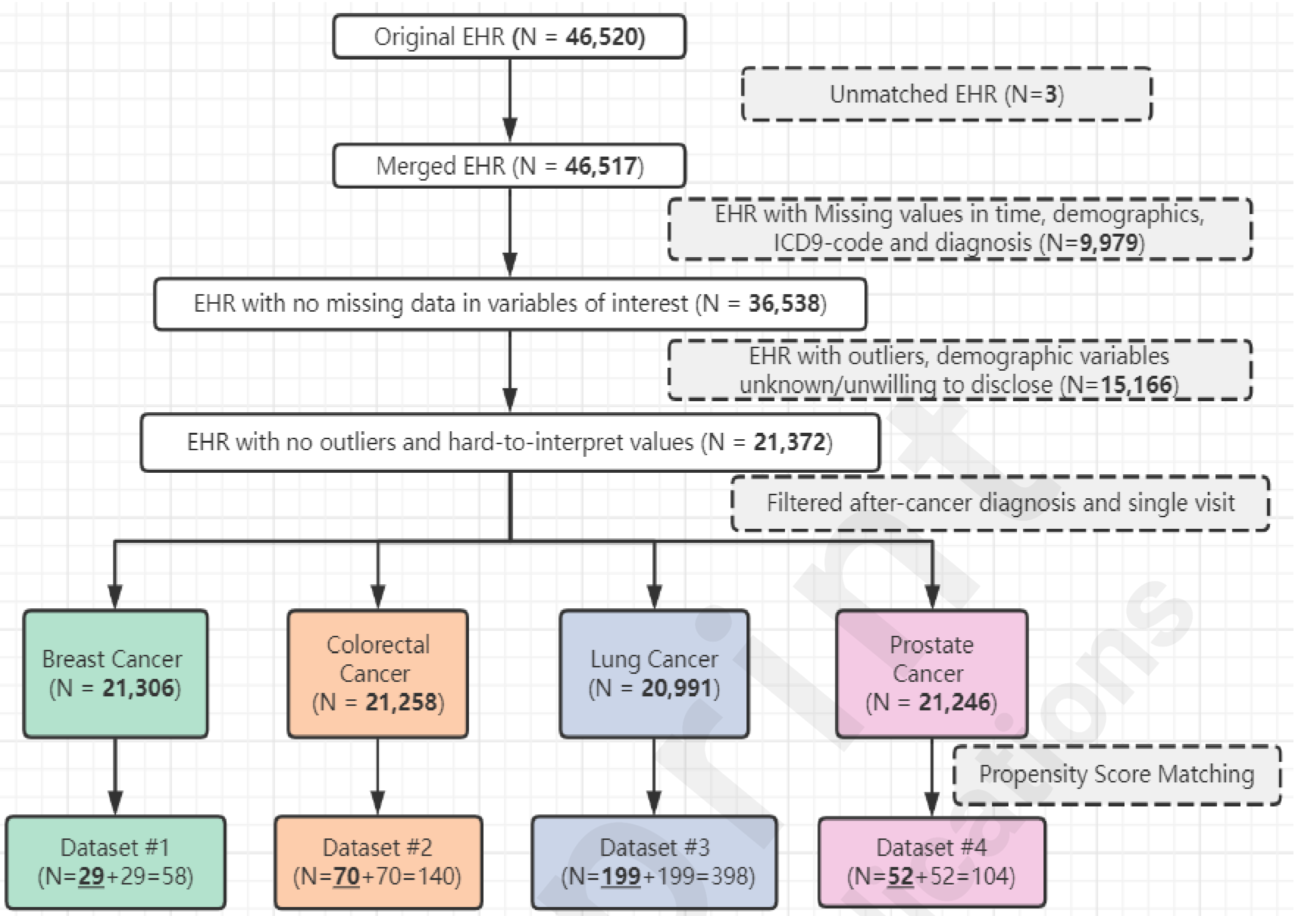
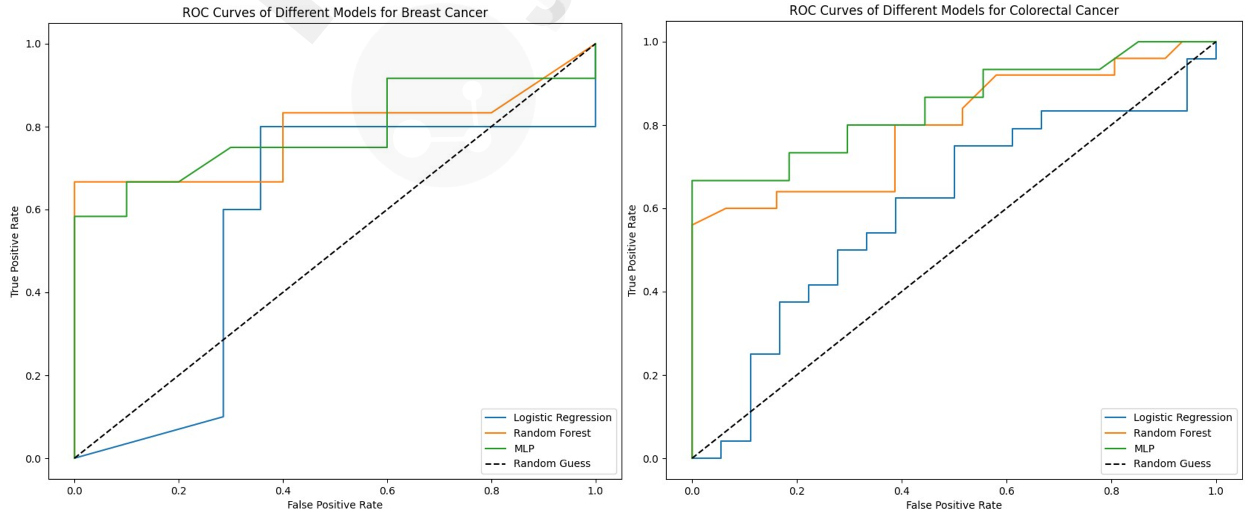

Original EHR (N = 46,520)
Unmatched EHR (N=3)
Merged EHR (N = 46,517)
EHR with Missing values in time, demographics, ICD9-code and diagnosis (N=9,979)
EHR with no missing data in variables of interest (N = 36,538)
EHR with outliers, demographic variables unknown/unwilling to disclose (N=15,166)
EHR with no outliers and hard-to-interpret values (N = 21,372)
Filtered after-cancer diagnosis and single visit
Breast Cancer (N = 21,306)
Colorectal Cancer (N = 21,258)
Lung Cancer (N = 20,991)
Prostate Cancer (N = 21,246)
Propensity Score Matching
Dataset #1 (N=29+29=58)
Dataset #2 (N=70+70=140)
Dataset #3 (N=199+199=398)
Dataset #4 (N=52+52=104)

Figure 1. MIMIC III data processing pipeline.



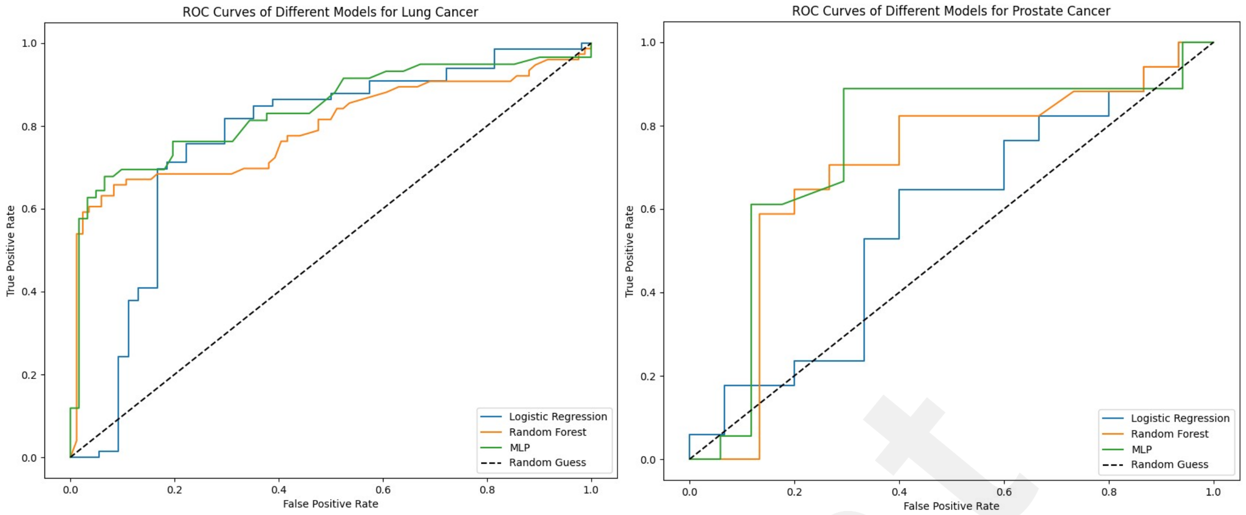


Figure 2. AUCs performance of the three binary classification models (LR, RF and MLP). The figure shows AUC curves of breast cancer, colorectal cancer, lung cancer and prostate cancer for LR, RF and MLP, respectively.

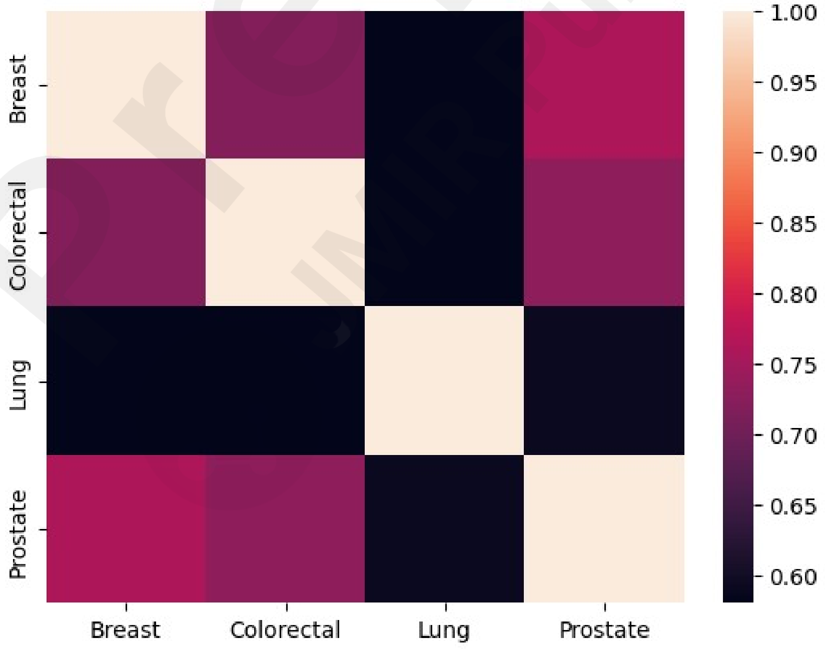


Figure 3. Rank Biased Overlap similarity score of risk factors for four cancer types. High value represents high similarity, low value represents low similarity of risk factor ranks between two cancer types.

Table 1. Features selected for predicting cancer risks.

Features	Factors
Acidosis	Acidosis
Acute kidney failure, unspecified	Acute kidney failure
Age	Age
Anemia, unspecified	Anemia
Acute posthemorrhagic anemia	Anemia
Depressive disorder, not elsewhere classified	Depressive disorder
Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	Diabetes
Esophageal reflux	Esophageal reflux
Ethnicity	Ethnicity
Gender	Gender
Cardiac complications, not elsewhere classified	Heart disease
Aortocoronary bypass status	Heart disease
Coronary atherosclerosis of native coronary artery	Heart disease
Old myocardial infarction	Heart disease
Congestive heart failure, unspecified	Heart disease
Atrial fibrillation	Heart disease
Subendocardial infarction, initial episode of care	Heart disease
Pure hypercholesterolemia	Hyperlipidemia
Other and unspecified hyperlipidemia	Hyperlipidemia
Unspecified essential hypertension	Hypertension
Other iatrogenic hypotension	Hypotension
Unspecified acquired hypothyroidism	Hypothyroidism
Marital status	Marital status
Religion	Religion
Acute respiratory failure	Respiratory/pulmonary diseases
Unspecified pleural effusion	Respiratory/pulmonary diseases
Pneumonia, organism unspecified	Respiratory/pulmonary diseases
Pneumonitis due to inhalation of food or vomitus	Respiratory/pulmonary diseases
Pulmonary collapse	Respiratory/pulmonary diseases
Chronic airway obstruction, not elsewhere classified	Respiratory/pulmonary diseases
Unspecified septicemia	Sepsis
Personal history of tobacco use	Tobacco use
Urinary tract infection, site not specified	Urinary tract infection (UTI)

Table 2. Comparison of model performance across four types of cancer

	Breast cancer			Colorectal cancer			Lung cancer			Prostate cancer		
	LR	RF	MLP	LR	RF	MLP	LR	RF	MLP	LR	RF	MLP
Accuracy	0.56	0.73	0.78	0.60	0.70	0.76	0.74	0.80	0.83	0.59	0.72	0.78

Specificity	0.45	0.70	0.80	0.67	0.61	0.81	0.61	0.92	0.87	0.53	0.80	0.84
Sensitivity	0.71	0.75	0.75	0.54	0.80	0.73	0.85	0.68	0.80	0.65	0.65	0.72
F1-score	0.56	0.75	0.75	0.60	0.70	0.79	0.78	0.78	0.84	0.63	0.71	0.76

*: LR: Logistic Regression, RF: Random Forest, MLP: Multilayer Perceptron

Table 3. Top-10 ranked features generated across four different cancer types.

Model: RF

Ranking	Breast Cancer	Colorectal Cancer	Lung Cancer	Prostate Cancer
1	Age	Age	Age	Age
2	Hypertension	Respiratory/Pulmonary diseases	Hypertension	Hypertension
3	Religion	Hypertension	Religion	Religion
4	Marital status	Acute kidney failure	Hyperlipidemia	Heart diseases
5	Respiratory/ Pulmonary diseases*	Diabetes	Heart diseases	Marital status
6	Heart diseases**	Heart diseases	Acute kidney failure	UTI
7	Race/Ethnicity	Hyperlipidemia	UTI	Respiratory/Pulmonary
8	Depressive disorders	Race/Ethnicity	Respiratory/Pulmonary	Anemia
9	Acute kidney failure	Religion	Marital status	Hyperthyroidism
10	Anemia	Acidosis	Anemia	Diabetes

Model: MLP

Ranking	Breast Cancer	Colorectal Cancer	Lung Cancer	Prostate Cancer
1	Age	Age	Tobacco use	Age
2	Gender	Diabetes	Age	Gender
3	Hyperlipidemia	Anemia	Respiratory/Pulmonary diseases	Race/Ethnicity
4	Heart diseases	Acidosis	Gender	Tobacco use
5	Race/Ethnicity	Hyperlipidemia	Race/Ethnicity	Diabetes
6	Marital status	Sepsis	Diabetes	Hyperlipidemia
7	Depressive disorder	Gender	Hyperlipidemia	Heart diseases
8	Religion	Race/Ethnicity	Hypertension	Marital status
9	Anemia	Marital status	Heart diseases	Religion
10	Hypothyroidism	Depressive disorder	Acute kidney failure	Depressive disorder

*: Respiratory/Pulmonary diseases include pneumonia, acute respiratory failure, chronic airway obstruction, and other respiratory or pulmonary complications.

**: Heart diseases include atrial fibrillation, myocardial infarction, congestive heart failure, coronary atherosclerosis, and other cardiac complications.