International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7, 2025
https://theaspd.com/index.php

Cancer Detection Using Multifactorial Analysis for Examining the Impact of Lifestyle

Dr Nilam Upasani[1], Yajushrestha Shukla[2], Sanika Upasani[3], Swayam Pedgaonkar[4], Dr. Deepali Joshi[5]
[1]Associate Professor, Balaji Institute of Technology & Management, Sri Balaji University, Pune, nilamupasani@gmail.com, Orcid Id- 0000-0001-5937-737X
[2]Intern, Tracelink, yajushreshtha@gmail.com
[3]Software engineer, HP inc., sanikaupasani2212@gmail.com
[4]Associate Software Engineer, A.P. Moller Maersk, Swayam.pendgaonkar@gmail.com
[5]Assistant professor, Vishwakarma Institute of Technology, Pune, deepali.joshi@vit.edu, Orcid Id - 0000-0002-8832-9294

Abstract: *The increasing worldwide worry about cancer makes us need an ingenious prediction model for early diagnosis. In the realm of cancer prediction, the escalating global concern demands innovative solutions for early diagnosis. This paper addresses the pressing need by introducing a modern approach grounded in multimodal analysis. It includes various elements that affect lifestyles, genetics, and behavioral distinctions to improve people's health and well-being, such as diet, height, weight, blood groups, marital status, smoking, and alcohol consumption. Our model based on the above-mentioned factors predicts cancer risk and offers a complete and well-timed assessment. Unlike traditional models, our pioneering method goes beyond singular indicators, offering a holistic prediction framework. This novel approach envisions a paradigm shift where individuals and healthcare professionals proactively assess and manage cancer risks. By leveraging AI and machine learning, our research propels the development of user-centered, comprehensive predictive models, promising transformative impact on public health and contributing to the evolution of healthcare systems.*
Keywords: *Cancer Risk Prediction Model, Genetics Cancer Prediction, Lifestyle, Cancer Detection, Multifactorial Analysis*

1. INTRODUCTION

Cancer, a word that has taken over 100 million human lives in the last decade [1] or so, is a disease that goes across borders and cultures, causing havoc on countless lives for generations. This persistent effort by this enemy has driven us to seek better and newer ways to fight it. With each passing day, we strive to understand it better, diagnose it earlier, and consequently conquer it. In our fast moving, modern world, where time is of absolute importance and our health is cardinal, we found ourselves in need of an ingenious solution—a predictive model for early detection of cancer [2, 3]. This research is the result of that need—the incorporation of our collective aspirations to face cancer. The initiation of our exploration into this predictive model commences with the recognition that cancer manifests as a multifaceted entity, analogous to the complexities inherent in life itself, rather than existing in isolation. Its etiology is shaped by behavioral patterns, influenced by genetic predispositions, and intricately interwoven with lifestyle factors. Therefore, we must blend various factors to create a comprehensive prediction model for cancer.

The factors are the eating habits that is whether we prefer a vegetarian or a non-vegetarian diet, the physical attributes like height and weight, blood group, gender, marital status, and other habits like whether smoking and alcohol consumption, and how often and how much. With these factors in perspective, we propose an all-inclusive assessment of the risk of cancer. The tools at our disposal have evolved, owing to the incredible developments in data science, and machine learning. These technological advancements give us the opportunity to provide early detection capabilities and timely mediation.

Imagine a future where medical professionals and individuals themselves can have a perspective on their health. A future where the variations in our lives- the "what we eat, who we are, and how we live"—are examined carefully to provide us with a personalized alarm system for fatal diseases. We aim to move beyond the boundaries of a medicine based on the reaction to the arrival of conditions like cancer to a proactive approach where we anticipate its arrival and act accordingly.

Our approach to being user-centered believes that individuals and their well-being are to be placed as the topmost priority. We are not just attempting to predict cancer, but also to improve people's lives by contributing to their health and well-being. We want to traverse the gap between the intricacies of life, genetics, and behavioral separations, offering apprehensions that are both meaningful and practical. Our research stresses the need for a comprehensive prediction model, one that summarizes the entirety of what makes us unique yet intertwined. Our control flow has been described in Figure 1.

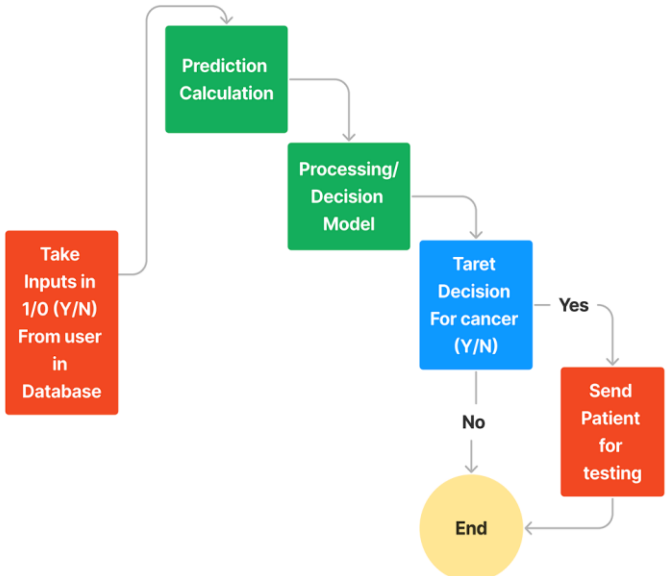

Figure 1. Flow of choice depending on User Habits

It is a call to visualize the possibilities, to see beyond the current line of advancements, and to consider a world where cancer is not a mountain we cannot climb but a challenge that we can tackle with a well-informed strategy. In the pages that follow, we dive into this very approach to exploring the complexities of multifactorial predictive detection of cancer. We comb through existing research, throw light on existing methodological considerations, and reveal the expected outcomes of our journey.

1.1Problem Statement

The problem at hand is the need for a prediction model that should be capable of early detection of cancer by utilizing multifactorial analysis. This model must ease all-inclusive assessments of cancer risk. This work is directed at providing a complete, user centered approach to predicting and preventing the onset of cancer. The final goal is to entitle individuals and medical professionals to a timely mediation plan of action, thus contributing to public welfare and well-being and completely changing the landscape of cancer diagnosis and prevention.

1.2Objectives

The aim of this work is to enhance the public health through personalized risk assessment and timely action. Following are the objectives of work.

- To develop a multifaceted approach to assess cancer risk that encompasses multiple predictive factors, including diet, weight, height, blood type, gender, marital status, smoking, and alcohol intake.
- To empower medical professionals and individuals to assess the risk of cancer at the early stage using machine learning, therefore allowing them to intervene when necessary.
- To provide a user-centered approach to enhance public health by making cancer prediction more relevant and accessible to individuals.

The remaining paper is organized as follows. Section 2 discusses the work related to the cancer detection. Section 3 discusses the proposed methodology. The experimental results of the proposed methodology and its comparison with other standard classifiers is presented in Section 4. The conclusions are drawn in Section 5 and future scope is discussed in Section 6.

2. LITERATURE REVIEW

Xie et al. [4] developed an advanced 2D CNN-based system for pulmonary nodule detection. By utilizing an upgraded 2D CNN architecture to lower false positives and integrating a faster R-CNN for nodule classification, the system improved the CT reading process. Their approach achieved a sensitivity of 86.42%

when evaluated in the LUNA16 dataset, presenting a substantial step forward in early lung cancer diagnosis. Jiang et al. [5] effectively delineated lung nodules through the utilization of a multi-patch methodology, amalgamating expertise from radiologists alongside the implementation of a four-channel neural network model. The method delivered significant progress in lung nodule detection for early cancer diagnosis. The research work presented in [6], proposed a deep learning strategy that involved extracting attributes using multiple methods and selecting the best features with the Fuzzy Particle Swarm Optimization algorithm. Their resulting dataset was employed to train a CNN with reduced computational complexity, achieving a remarkable sensitivity of 97.93%. This research signifies the potential of advanced feature extraction and selection techniques for improving early lung cancer diagnosis.

This review [7] delves into the intricacies of oral squamous cell carcinoma (OSCC), exploring the changing landscape of its incidence, particularly among young patients. While traditionally linked to tobacco and alcohol consumption, the research highlights the challenge of ascertaining the underlying causes of OSCC in younger individuals. Genetic factors, predispositions to genetic instability, and the presence of risk factors except tobacco and alcohol are examined in-depth. The review emphasizes the importance of investigating potential environmental carcinogens, familial cancer history, viral infections, and stress as contributors to OSCC. Additionally, it discusses the application of non-invasive optical and photodynamic diagnostic methods, shedding light on potential advancements in OSCC detection and diagnosis [8].

Early detection is crucial to reducing lung cancer morbidity and mortality [9]. High-risk ever-smokers may benefit from screening using low-dose computed tomography (LDCT), according to the National Lung Screening Trial (NLST). However, the requirements for NLST selection are suboptimal, excluding many early-stage lung cancer patients based on smoking history and age [10]. Just 26.7% of US lung cancer cases meet NLST criteria, indicating room for improving precision in patient selection [11]. Alternative approaches, like stereotactic ablative body radiotherapy (SABR), show promise, particularly for stage I lung cancer patients ineligible for surgery [12]. Determining the optimal frequency of CT scans remains a challenge, considering concerns about radiation-induced cancers. Mitigating risks involves refining risk assessment tools and imaging protocols [13].

LDCT has limitations, including high false-positive rates and late-stage cancer emergence between screenings, necessitating complementary tests. Biomarkers like plasma microRNAs and circulating tumor cells (CTCs) have potential for reducing false positives and aiding in the diagnosis of aggressive cancers [14]. A deeper understanding of early molecular events in lung tumorigenesis may lead to novel biomarker development [15].

The study [16] delves into genetic marker analysis with the aim of improving early and precise lung cancer diagnosis. Exploring potential biomarkers, such as microsatellite changes, DNA hypermethylation, gene mutations (p53 and KRAS), and microRNA expression,, the research investigates avenues for more effective lung cancer detection. It highlights the significant promise of microRNAs, particularly their expression profiles, as potential genetic markers capable of predicting lung cancer even 24 months earlier than traditional methods. Standardizing the quantification of circulating microRNAs [17] is emphasized as a crucial step for future clinical applications.

This study [18] focused on identifying metabolic subtypes of gastric cancer and assessing their prognostic significance. Researchers carefully analyzed data from the Cancer Genome Atlas (TCGA) database [19], and clinical follow-up information. The analysis presented in [20, 21] of patient prognosis unveiled notable disparities, with the cholesterol subtype displaying a less favorable outlook compared to the glycolysis subtype. These differences were further supported by observed variations in the expression patterns of cholesterol and glycolysis genes within these subtypes.

For the purpose of assessing the risk of colorectal cancer (CRC), the study [22] utilized data from the Darmkrebs: Chancen der Verhütung durch Screening (DACHS) project, a continuous population-based case-control investigation that has been conducted in southwest Germany since 2003. Eligibility for participation was determined by having a confirmed initial diagnosis of CRC, with participants aged at least 30 years and conversant in German. The control group was selected randomly from population registries, employing age, sex, and county of residence as matching criteria. Thorough interviews were conducted to capture lifestyle and medical information, encompassing factors such as body mass index, eating habits, drinking, smoking, and physical activity. Genomic data were obtained from either blood samples or buccal cells, according to the participants, with DNA extraction performed via standard procedures. A significant factor in the CRC risk assessment was the recent colonoscopy history [23]. To ascertain the association between CRC risk and lifestyle factors, polygenic risk scores, and recent colonoscopy, the study employed multiple logistic regression, taking sex and age into account. The investigation also encompassed the estimation of the 30-year absolute risk of developing CRC in 50-year-old individuals, founded on diverse risk profiles [24]. The study, involving 4220 CRC patients and 3338 control participants, provided invaluable insights into the interconnectedness of lifestyle components, genetic predisposition, and colonoscopy history in the context of assessing CRC risk. In this study [25], the total contribution of alcohol-based cancer in 2020 was found. The method involved the selection of types of cancer with a significant result of a causal relationship with alcohol consumption, using data from the IARC and GIS on Alcohol and Health. The findings revealed an astonishing estimated 741,300 new cancer cases in 2020 (4.1% of all new cases) were because of alcohol consumption, with the highest PAFs observed for oesophageal, pharyngeal, and lip cavity cancer. The study stresses the worldwide spread of cancer owing to alcohol consumption, highlighting the need for very effective and immediate policies and interventions to reduce alcohol-related cancer risks and overall alcohol consumption.

In the study [26], 3309 participants enrolled in a currently going prospective cohort study, and their data was analyzed, which focused on the possible connection between lifestyle and psychosocial job qualities associated with cancer risk. The study gauged job demands, job strain, and iso-strain to investigate their influence on cancer risk, along with lifestyle activities like smoking, consuming alcohol, engaging in physical activity and eating vegetables and fruits [27]. The results did not reveal significant inconsistencies in the presence of lifestyle risk factors for cancer, among other work characteristics that were investigated [28]. Furthermore, there was little evidence to indicate a link between occupational strain, iso-strain, and lifestyles linked to cancer when multivariate analysis was used. Finally, the outcomes of the study do not substantiate the notion that job strain, or iso-strain is associated with adopting a cancer related lifestyle.

The study population involved individuals who were employed at the time of completing the Health and Life Experiences Questionnaire, which is the reason for the higher representation of younger participants with an elevated level of education. The analysis of sociodemographic variables and lifestyle characteristics, stratified by gender, revealed considerable differences, particularly in alcohol consumption, where men displayed a notably higher daily intake than women. Furthermore, the results did not identify a pronounced association between lifestyle factors, occupational strain, and iso-strain across the years 1993 and 1996 [27, 28]. Conclusively, this study has probed the intricate relationship between psychosocial job characteristics and cancer-related lifestyle risk factors, yielding inconclusive findings. The study serves as a stepping stone for deeper research into other psychosocial aspects and their potential mediation in the complex interplay between work environment and health behaviors. While the investigation adds valuable insights to this understudied area, it also highlights the imperative need for future comprehensive research endeavours.

With a 5-year survival rate of less than 1/5th, esophageal cancer is a devastating and complex disease primarily brought on by delayed diagnosis. The vast landscape of esophageal cancer presents numerous challenges, with varied commonalities and risk factors across different populations. While there have been significant risk factors like tobacco, alcohol, and reflux esophagitis for a long time, dietary practices and nutrition have an influence. This review provided us with a thorough summary of how diet and nutrition affect the risk of esophageal cancer, which is still less understood.

Further study resulted in the understanding that the causation of the disease was attributed to a wide range of dietary elements like fruits, vegetables, vitamins and minerals, fats, meats, salted foods, carcinogens, nitrogen compounds, mycotoxins, and also the temperature of the food consumed. Using this derived understanding is crucial for advancing our strategies to tackle esophageal cancer and devising effective prevention plans. By highlighting the importance of tobacco and alcohol cessation, a diet absolutely rich in vegetables and fruits, and the significant role of nutrition in at-risk populations, this paper [29] highlights the need for comprehensive efforts to reduce the impact of esophageal cancer on a broader population.

In this study [30], they aimed to scrutinize the relationship between extended exposure to air pollution and the degree of lung cancer among Koreans, since it is Korea's primary cause of cancer-related mortality. They observed an increasing trend in lung cancer cases, especially with shifts in biological types, such as an increase in the presence of adenocarcinomas, and a decrease in squamous cell carcinomas. Even though the key cause of lung cancer is tobacco use, high lung cancer rates amongst none and never smokers have raised questions about other factors like occupational exposures, environmental tobacco dust, lower socioeconomic status, and air pollution. This study offers critical insights into the role of air pollution in the emergence of lung cancer, emphasizing the importance of considering histological subtypes when assessing its association with air pollution.

Table 1: Summarized Form of Literature Review

Reference number	Method used	Dataset used	Finding/ Limitation/ Future Scope/Performance parameters
[4]	3 CNN in MCNN	LIDC-IDRI	1. Enhanced nodule detection. 2. High sensitivity, low FP. 3. Diagnostic accuracy: 86.84%.
[7]	Cytologic diagnosis during OSCC development	Web-based search initiated using Medline/PubMed	Genetics, alcohol, smoking, carcinogens, and various factors contribute to oral cancer.
[9]	Low Dose CT scan	NLST selections	Increasing population selection accuracy; SABR shows potential for non-surgical patients.
[16]	Genetic Marker Analysis	Biomarkers Studied	Potential microRNA biomarkers
[18]	Data processing, classification.	TCGA gastric cancer samples dataset.	Metabolic subtypes impact via gene expression variations
[22]	Data analysis and calculations	DACHS population-based study	Lifestyle, colonoscopy reduce CRC risk with risk reduction measures.
[26]	Cross-sectional survey analysis	Dutch cohort study	Logistic, linear regression found no job-cancer link; future explores psychosocial factors
[30]	Case-control study, regression models	908 lung cancer patients	Air pollution links, odds ratios, confidence intervals for lung cancer

3. PROPOSED METHODOLOGY

In this section, we discuss the proposed method that offers a holistic prediction framework. This novel approach not only significantly enhances early detection but also envisions a paradigm shift where individuals and healthcare professionals proactively assess and manage cancer risks.

3.1Dataset Description

The Health Assessment Dataset comprises information gathered from 8322 individuals, serving as a robust resource for evaluating various medical conditions and associated risks. These comprise numerical data like alcohol consumption levels and BMI, as well as categorical variables like age, gender, and smoking status. The information also contains binary indicators for genetic vulnerabilities to certain diseases and environmental exposures like air pollution. As categorical variables, symptoms pertaining to cardiovascular and respiratory health are also recorded. The total health assessment level of each individual is an ordinal data and divided into three categories: low, medium, and high.

Risk Factors for lung cancer are as follows:

●Gender: This field represents the gender of individuals in the dataset, typically categorized as male or female. Gender can be a significant factor in assessing lung cancer risk.
●Age: Age is a number characteristic that represents each person's age within the dataset. Given that the incidence of lung cancer tends to rise with age, age has a significant role in determining one's risk for the disease.
●Smoking: This categorical attribute indicates whether an individual is a smoker, typically classified as a binary choice, such as "smoker" or "non-smoker." Smoking is a well-established risk factor for lung cancer.
●Yellow fingers: This attribute may describe whether an individual has yellow fingers, which can be associated with smoking due to nicotine staining.
●Anxiety: Anxiety is a categorical attribute that signifies whether an individual experiences anxiety, a psychological factor that may contribute to smoking and other behaviors related to lung cancer risk.
●Peer pressure: Peer pressure is a categorical attribute that can reveal whether or not a person is persuaded by their peers to take up risky behaviors like smoking or other habits.
●Chronic disease: This field may describe whether an individual has any chronic diseases or conditions, can affect their overall health and, indirectly, their lung cancer risk.
●Fatigue: Fatigue is a categorical attribute that can signal whether an individual experiences excessive tiredness, which might be linked to various lifestyle factors and health conditions.
●Allergy: This attribute denotes whether an individual has allergies, which can affect respiratory health and potentially influence lung cancer risk.
●Wheezing: Wheezing is a categorical attribute that can signify whether an individual experiences wheezing or breathing difficulties, which may affect lung health.
●Alcohol consumption: This field indicates whether an individual consumes alcohol that is one of the factors in assessing lung cancer risk.
●Coughing: Coughing is a categorical attribute that represents whether an individual experiences persistent coughing, which may affect respiratory health.
●Shortness of breath: Shortness of breath is a categorical attribute that can denote whether an individual experiences difficulty breathing, which is another important respiratory health factor.
●Swallowing difficulty: This attribute may describe whether an individual has difficulty swallowing, which could be linked to various health issues.
●Chest pain: Chest pain is a categorical attribute that signifies whether an individual experiences chest pain, which can have various causes, including those related to lung health.
●Lung cancer: This binary attribute is the target variable, indicating whether an individual has been diagnosed with lung cancer or not

3.2Implementation

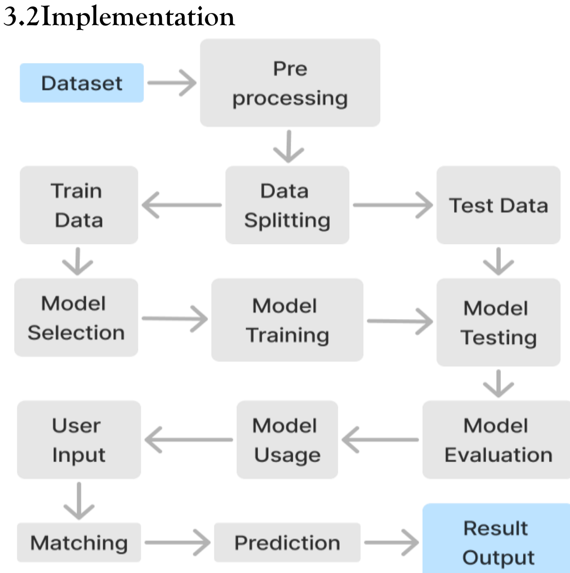

Figure 2. Proposed Methodology

The proposed methodology is shown in Figure 2. The subsequent phase involves collecting essential patient data through a structured questionnaire. The data encompasses critical elements such as age, dietary patterns, smoking history, fatigue levels, allergies, familial cancer history, respiratory conditions, and other pertinent variables. The comprehensiveness and quality of this data directly influence the model's precision and performance. Once the data has been gathered, it is divided into training and testing subsets to prevent overfitting and enable the model to generalize to new, unseen data effectively. Data preprocessing is a pivotal step in cleaning and preparing the data for analysis. This process includes handling missing values, addressing outliers, and standardizing data formats. For instance, in medical data, addressing missing values in patient records is crucial for a complete dataset.

Categorical data exploration entails determining the distribution and linkages within the dataset. Visualizing the spread of different cancer kinds, for example, can reveal trends that might help with early identification. Analyzing age-related data reveals information about age groups' susceptibility to cancer and other important patterns, which informs feature selection and model design.

A comparison of age and categorical variables reveals correlations that influence the model's predictive ability. These visualizations show how risk variables change with age and allow for more educated feature selection. Creating a heatmap to visualize variable correlations can help with feature selection and identifying the features that have a major influence on cancer prediction. For example, research might indicate associations between lifestyle decisions and cancer risk.

Addressing class imbalance by oversampling guarantees that the model does not favor the dominant class, which is crucial for cancer prediction in areas where positive cases are rare. Data splitting into training and testing sets provides model validation, whereas data scaling is critical for certain model types.

Choosing an appropriate machine learning model entails testing with several algorithms, followed by training and fine-tuning for optimum performance. Model evaluation is comparing the model's predictions to actual outcomes using measures such as precision, accuracy, recall, and F1-score to determine its usefulness in predicting cancer cases. In the last step, we explore future development opportunities, such as incorporating real-time patient data and adding additional features to improve prediction precision and relevance.

4. RESULT ANALYSIS

Our findings are summarized in the following tables, graphs, and figures. Each provides a different aspect of the analysis and model performance. The analysis compares various classification models, each evaluated based on various performance metrics.

4.1Model Comparison/Selection

1) Support Vector Machine (SVM):

The SVM model demonstrates exceptional accuracy (99%), achieving near-perfect precision (98%) and recall (99%) for both class 0 and class 1 instances. This makes it highly suitable for binary and multiclass classification in high-dimensional spaces.

2) K-Nearest Neighbors (KNN):

KNN displays excellent accuracy and a strong balance between precision and recall, particularly excelling in accurately identifying class 0 instances.

3) Random Forest:

The Random Forest model excels in achieving excellent recall and precision for both class 0 and class 1 instances. Its ensemble approach, utilizing multiple decision trees, effectively reduces overfitting while maintaining high accuracy. This makes Random Forest particularly robust and suitable for complex classification tasks.

4) Gradient Boosting:

The Gradient Boosting model demonstrates exceptional accuracy and precision, with a focus on minimizing prediction errors through the additive model it builds. It achieves a good balance between precision and recall for both class 0 and class 1 instances, showcasing its effectiveness in various predictive modelling tasks.

Table 2: Model Comparison

Model	Recall	Precision	F-1 Score	ROC AUC
SVM	**98%**	**97%**	**97%**	**97.31%**
KNN	94%	95%	94%	96.39%
RANDOM	95%	95%	95%	96.46%
GRADIENT	96%	96%	96%	97.01%
Our Method SVM (ADABoost In further Steps)	**99%**	**98%**	**98%**	**98.13%**

Based on the model comparison presented in Table 2, the SVM stands out as the most suitable choice for the classification task. SVM achieves the highest recall, precision, and F1 score among all models, with precision and recall rates of 97% and 98% respectively, resulting in an overall F1 score of 97%. Additionally, SVM demonstrates a robust performance with an ROC AUC of 97.31%.
This exceptional performance indicates SVM's ability to effectively classify instances across different classes with high accuracy while maintaining a strong balance between precision and recall. The high ROC AUC further validates its capability to distinguish between positive and negative classes efficiently.
Therefore, based on its superior performance across multiple metrics, particularly in precision and recall, SVM is chosen as the right model to use for the classification task. It offers a reliable and accurate solution for effectively categorizing instances within the dataset.

4.2RESULTS AND DISCUSSION

In our paper, we utilized the Surveyed_Cancer dataset, employing Python and Orange as our tools of choice. We implemented SVM with plans to further enhance its performance through ADA Boost. The achieved accuracy of 98.13% underscores the effectiveness of our approach. We employed various data preprocessing techniques, tested multiple machine learning models, and selected the optimal model based on precision.
ROC Curve (Receiver Operating Characteristic Curve): The ROC curve is graphical depiction of a model's capacity for class differentiation. Within the Random Forest framework, the ROC curve facilitates the visualization of the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) by modifying the model's classification threshold.
AUC (Area Under the ROC Curve): AUC is a single metric that quantifies a model's overall performance. The area under the ROC curve is measured, with a greater AUC indicating better model performance. AUC values vary from 0 to 1, with 0.5 representing chance and 1 representing a perfect classifier. A high AUC in the context of our model indicates that the model is competent at differentiating between classes and producing accurate predictions

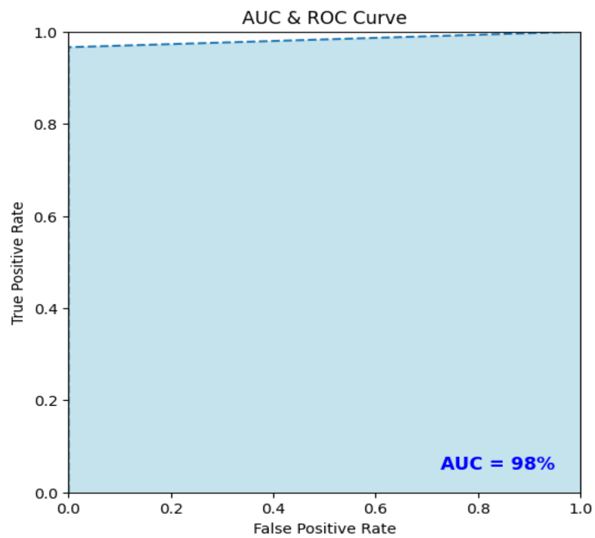

Figure 3. AUC & ROC Curve for our model

The ROC curve and AUC analysis, as shown in Figure 3 illustrate the excellent performance of the proposed model. The curve's trajectory, which leans towards the top-left corner, indicates the model's strong ability to distinguish between classes. The AUC value, quantified at 98%, underscores the model's exceptional precision in classifying data. This visual representation, along with the high AUC percentage, confirms the model's robustness in distinguishing between the two classes, making it a reliable choice for cancer disease detection.

5. CONCLUSION

In summary, our research was dedicated to the development of a robust model for the early detection of cancer in patients. Our method was anchored by thorough data preprocessing, feature selection, and a systematic examination of several machine learning algorithms. We systematically tested and compared several models, including the KNN, Random Forest, SVM, and Gradient Boosting, with the objective of identifying the most effective model. After a thorough assessment, our findings revealed that the SVM model exhibited superior performance when compared to other models. We implemented SVM with plans to further enhance its performance through ADA Boost. It achieved a remarkable accuracy rate of 98.13% and a precision of 98%, which indicates its exceptional ability to accurately discern patients with cancer while maintaining a very low misclassification rate. Furthermore, our study included a comprehensive analysis of ROC curves and AUC scores. The AUC score of 98.31% exemplifies the SVM model's outstanding predictive capabilities and its capacity to effectively differentiate between cancer and non-cancer cases. Our research demonstrates that the meticulous application of machine learning techniques, combined with rigorous data analysis, holds significant promise in advancing early cancer detection. These results underline the potential for these models to provide invaluable support to healthcare professionals, contributing to timely and precise diagnoses. This research provides a robust foundation for further exploration within the realm of medical diagnostics and underscores the transformative potential of data-driven methodologies in the field of healthcare.

6. FUTURE SCOPE

Building on current progress, we anticipate advancements in early cancer detection and personalized healthcare. Our team aims to create an innovative diagnostic gadget that analyzes blood, saliva, and sweat for a comprehensive health assessment. The objective is a non-invasive, user-friendly instrument for real-time analysis, revolutionizing early detection, reducing testing costs, and enhancing healthcare accessibility, especially in remote or impoverished areas.

Early diagnosis and timely intervention can significantly impact outcomes, potentially saving lives. We strive to usher in an era where individuals monitor their health proactively, and healthcare professionals rely on improved, non-invasive instruments for early disease detection. This disruptive approach would be contributing to a more sustainable healthcare system.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Crosby, D., Bhatia, S., Brindle, K.M., Coussens, L.M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R.C., Gambhir, S.S., Kuhn, P. and Rebbeck, T.R., 2022. Early detection of cancer. Science, 375(6586), p.eaay9040.

[3] Ritchie, J. B., Welch, B. M., Allen, C. G., Frey, L. J., Morrison, H., Schiffman, J. D., Alekseyenko, A. V., Dean, B., Halbert, C. H., & Bellcross, C. (2022, May 16). Comparison of a Cancer Family History Collection and Risk Assessment Tool – ItRunsInMyFamily – with Risk Assessment by Health-Care Professionals. Public Health Genomics, 25(3-4), 80–88. https://doi.org/10.1159/000520001

[4] Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. Journal of Infection and Public Health, 13(9), 1274-1289. https://doi.org/10.1016/j.jiph.2020.06.033.

[5] H. Jiang, H. Ma, W. Qian, M. Gao and Y. Li, "An Automatic Detection System of Lung Nodule Based on Multigroup Patch-Based Deep Learning Network," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 4, pp. 1227-1237, July 2018, doi: 10.1109/JBHI.2017.2725903.

[6] Asuntha, A., Srinivasan, A. Deep learning for lung Cancer detection and classification. Multimed Tools Appl 79, 7731–7762 (2020). https://doi.org/10.1007/s11042-019-08394-3

[7] Zygogianni, A.G., Kyrgias, G., Karakitsos, P. et al. Oral squamous cell cancer: early detection and the role of alcohol and smoking. Head Neck Oncol 3, 2 (2011). https://doi.org/10.1186/1758-3284-3-2

[8] Harris, D. M., & Werkhaven, J. (1987). Endogenous porphyrin fluorescence in tumors. Lasers in surgery and medicine, 7(6), 467–472. https://doi.org/10.1002/lsm.190007060

[9] Knight, S., Crosbie, P., Balata, H., Chudziak, J., Hussell, T., & Dive, C. (2017). Progress and prospects of early detection in lung cancer. Open Biology, 7(9), 170070. https://doi.org/10.1098/rsob.170070

[10] Taiwo, E. O., Yorio, J. T., Yan, J., & Gerber, D. E. (2012). How have we diagnosed early-stage lung cancer without radiographic screening? A contemporary single-center experience. PloS one, 7(12), e52313. https://doi.org/10.1371/journal.pone.0052313

[11] Pinsky, P. F., & Berg, C. D. (2012). Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered?. Journal of medical screening, 19(3), 154–156. https://doi.org/10.1258/jms.2012.012010

[12] Zheng, X., Schipper, M., Kidwell, K., Lin, J., Reddy, R., Ren, Y., Chang, A., Lv, F., Orringer, M., & Spring Kong, F. M. (2014). Survival outcome after stereotactic body radiation therapy and surgery for stage I non-small cell lung cancer: a meta-analysis. International journal of radiation oncology, biology, physics, 90(3), 603–611. https://doi.org/10.1016/j.ijrobp.2014.05.055

[13] Rampinelli, C., De Marco, P., Origgi, D., Maisonneuve, P., Casiraghi, M., Veronesi, G., Spaggiari, L., & Bellomi, M. (2017). Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis. BMJ (Clinical research ed.), 356, j347. https://doi.org/10.1136/bmj.j347

[14] Sozzi, G., Boeri, M., Rossi, M., Verri, C., Suatoni, P., Bravi, F., Roz, L., Conte, D., Grassi, M., Sverzellati, N., Marchiano, A., Negri, E., La Vecchia, C., & Pastorino, U. (2014). Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 32(8), 768–773. https://doi.org/10.1200/JCO.2013.50.4357

[15] Kwon, M. C., & Berns, A. (2013). Mouse models for lung cancer. Molecular oncology, 7(2), 165–177. https://doi.org/10.1016/j.molonc.2013.02.010

[16] Wadowska K, Bil-Lula I, Trembecki Ł, Śliwińska-Mossoń M. Genetic Markers in Lung Cancer Diagnosis: A Review. International Journal of Molecular Sciences. 2020; 21(13):4569. https://doi.org/10.3390/ijms21134569

[17] Shen, C., Wang, X., Tian, L., & Che, G. (2014). Microsatellite alteration in multiple primary lung cancer. Journal of thoracic disease, 6(10), 1499–1505. https://doi.org/10.3978/j.issn.2072-1439.2014.09.14

[18] Zhu, Z., Qin, J., Dong, C., Yang, J., Yang, M., Tian, J., & Zhong, X. (2021). Identification of four gastric cancer subtypes based on genetic analysis of cholesterogenic and glycolytic pathways. Bioengineered, 12(1), 4780-4793. https://doi.org/10.1080/21655979.2021.1956247

[19] Wang, Z., Jensen, M. A., & Zenklusen, J. C. (2016). A Practical Guide to The Cancer Genome Atlas (TCGA). Methods in molecular biology (Clifton, N.J.), 1418, 111–141. https://doi.org/10.1007/978-1-4939-3578-9_6

[20] Jiang, J., Zheng, Q., Zhu, W., Chen, X., Lu, H., Chen, D., Zhang, H., Shao, M., Zhou, L., & Zheng, S. (2020). Alterations in glycolytic/cholesterogenic gene expression in hepatocellular carcinoma. Aging, 12(11), 10300–10316. https://doi.org/10.18632/aging.103254

[21] Karasinska, J. M., Topham, J. T., Kalloger, S. E., Jang, G. H., Denroche, R. E., Culibrk, L., Williamson, L. M., Wong, H. L., Lee, M. K. C., O'Kane, G. M., Moore, R. A., Mungall, A. J., Moore, M. J., Warren, C., Metcalfe, A., Notta, F., Knox, J. J., Gallinger, S., Laskin, J., Marra, M. A., ... Schaeffer, D. F. (2020). Altered Gene Expression along the Glycolysis-Cholesterol Synthesis Axis Is Associated with Outcome in Pancreatic Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research, 26(1), 135–146. https://doi.org/10.1158/1078-0432.CCR-19-1543

[22] Carr, P. R., Weigl, K., Edelmann, D., Jansen, L., Chang-Claude, J., Brenner, H., & Hoffmeister, M. (2020). Estimation of Absolute Risk of Colorectal Cancer Based on Healthy Lifestyle, Genetic Risk, and Colonoscopy Status in a Population-Based Study. Gastroenterology, 159(1), 129-138.e9. https://doi.org/10.1053/j.gastro.2020.03.016

[23] Brenner, H., Stock, C., & Hoffmeister, M. (2014). Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomized controlled trials and observational studies. BMJ, 348, g2467. doi:10.1136/bmj.g2467

[24] Pfeiffer, R. M., & Petracci, E. (2011). Variance computations for functionals of absolute risk estimates. Statistics & Probability Letters, 81(7), 807-812. https://doi.org/10.1016/j.spl.2011.02.002

[25] Rumgay, H., Shield, K., Charvat, H., Ferrari, P., Sornpaisarn, B., Obot, I., et al. (2021). Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. The Lancet Oncology, 22(8), 1071-1080. https://doi.org/10.1016/S1470-2045(21)00279-5

[26] A Jeanne M van Loon, Marja Tijhuis, Paul G Surtees, Johan Ormel, Lifestyle risk factors for cancer: the relationship with psychosocial work environment, International Journal of Epidemiology, Volume 29, Issue 5, October 2000, Pages 785–792, https://doi.org/10.1093/ije/29.5.785

[27] Pols MA, Peeters PHM, Ocké MC, Slimani N, Bueno de Mesquita HB, Collette HJA. Estimation of reproducibility and relative validity of the questions included in the EPIC physical activity questionnaire. Int J Epidemiol1997;26:S181–89.

[28] Baecke JAH, Burema J, Frijters JER. A short questionnaire for the measurement of habitual physical activity in epidemiological studies. Am J Clin Nutr1982;36:936–

[29] A. G. Palladino-Davis, B. M. Mendez, P. M. Fisichella, C. S. Davis, Dietary habits and esophageal cancer, Diseases of the Esophagus, Volume 28, Issue 1, 1 January 2015, Pages 59–67, https://doi.org/10.1111/dote.12097

[30] Lamichhane, D. K., Kim, H. C., Choi, C. M., Shin, M. H., Shim, Y. M., Leem, J. H., Ryu, J. S., Nam, H. S., & Park, S. M. (2017). Lung Cancer Risk and Residential Exposure to Air Pollution: A Korean Population-Based Case-Control Study. Yonsei medical journal, 58(6), 1111–1118. https://doi.org/10.3349/ymj.2017.58.6.1111

[31] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[32] B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

[33] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhannah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.

[34] Kibeom Jang, Minsoon Kim, Candace A Gilbert, Fiona Simpkins, Tan A Ince, Joyce M Slingerland "WEGFA activates an epigenetic pathway regulating ovarian cancer initiating cells" Embo Molecular Medicines Volume 9 Issue 3 (2017)

[35] Joseph A. Cruz and David S. Wishart "Applications of Machine Learning in cancer prediction and prognosis Cancer informatics" 2(3):59-77 · February 2007

[36] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.

[37] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"

[38] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.