INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 03 | MARCH - 2024 SJIF RATING: 8.176 ISSN: 2582-3930

Cancer Risk Identification using Machine Learning

Dr. Shanta Sondur
Professor
Department of Information Technology
Vivekanand Education Society, Institute of Technology.
Mumbai 400074, India
shanta.sondur@ves.ac.in

Vatsal Gadaria
Department of Information Technology
Vivekanand Education Society, Institute of Technology.
Mumbai 400074, India
2020.vatsal.gadaria@ves.ac.in

Nitish Jaiswal
Department of Information Technology
Vivekanand Education Society, Institute of Technology.
Mumbai 400074, India
2020.nitish.jaiswal@ves.ac.in

Bharat Kotwani
Department of Information Technology
Vivekanand Education Society, Institute of Technology.
Mumbai 400074, India
2015bharat.kotwani@ves.ac.in

Om Tandel
Department of Information Technology
Vivekanand Education Society, Institute of Technology.
Mumbai 400074, India
2020.om.tandel@ves.ac.in

***Abstract*—In this comprehensive study, we explore the SEER (Surveillance, Epidemiology, and End Result) dataset to meticulously craft a predictive model aimed at understanding patient survivability, focusing keenly on outcomes like patient survival diagnosis and treatments. Employing a good approach involving rigorous data cleaning, intricate feature extraction, and detailed correlation analysis, we identify the dominant attributes that significantly influence patient outcomes. Our methodology integrates classical machine learning models, including decision trees and random forests, with modern techniques such as neural networks. This mix of methodologies allows us to accurately predict patient survivability. Furthermore, our study delves into the intricate relationships between attributes, models, and algorithms, aiming to identify the dominant factors that influence the outcomes. Our groundbreaking findings underscore the enormous potential of integrating these dominant attributes, paving the way for the creation of exceptionally robust predictive models. These models will substantially enhance medical decision-making processes and, in turn, elevate the overall quality of patient care.**

Keywords - SEER dataset, patient survivability, machine learning, decision trees, random forests, neural networks, data cleaning, feature extraction, clinical data, medical decision-making.

Introduction

In the field of modern healthcare, the integration of advanced technology and medical expertise has revolutionized our comprehension, diagnosis, and treatment of diseases, particularly cancer. We have meticulously examined the extensive SEER (Surveillance, Epidemiology, and End Results) [1] data provided by 'The National Cancer Institute' [2] in the USA. Utilizing Artificial Intelligence and Machine Learning, we have developed a precisely crafted model that provides invaluable insights into cancer prognosis. This dataset is thoroughly organized by age, sex, race, year of diagnosis, and geographic areas, offering a wealth of information. In this pioneering project, our focus lies in understanding the intricate relationship between cancer diagnosis stages and race/ethnicity. Our exploration does not conclude here. We compute survival rates through a detailed analysis of critical factors such as diagnosis stage, age, and tumour grade or size [3]. This thorough examination reveals patterns and disparities, equipping medical experts with actionable intelligence. Thanks to our ML-powered predictions, healthcare professionals can anticipate challenges, tailor treatments, and significantly improve the quality of care for cancer patients. By utilizing advanced technology, we aim to provide personalized care and support to individuals of all backgrounds facing a cancer diagnosis. Our ultimate goal is to contribute to reducing health disparities and improving outcomes for all patients, regardless of race or ethnicity.

I. AIM

To propose a comprehensive framework designed to aid doctors in the treatment of cancer patients. This framework aims to assist healthcare professionals in understanding the correlation between different attributes with the help of historical databases (SEER) generated by cancer hospitals

II. OBJECTIVES

In our exploration of the SEER dataset, we delved into multiple perspectives to extract valuable insights concerning various outcomes related to cancer:

(i) Identification of the Outcome: -

Our analysis aimed at identifying key patterns and trends within the dataset. By employing techniques such as feature importance, correlation matrices, we gained a comprehensive understanding of the factors influencing different outcomes.

(ii) Prediction: -

Utilizing advanced machine learning models, including traditional algorithms and neural networks, we aim to develop models to predict survivability

(iii) Diagnosis: -

In the realm of diagnosis, we focused on leveraging the dataset to improve early detection and accuracy to aid doctors.

(iv) Treatment: -

To help doctors in choosing accurate treatments and cost effective one.

III. MOTIVATION FOR THE WORK

Cancer has always been a daunting challenge, prompting relentless efforts in the medical field to eliminate or mitigate its effects. With the continuous progress of technology, we being the professionals in the field of IT, recognizing the struggles faced by oncologists, have undertaken the task of aiding doctors with advanced technology. One of the significant challenges faced by doctors is the effective utilization of vast datasets, such as the extensive SEER dataset. The sheer volume of information can be overwhelming, making it arduous to extract pertinent insights efficiently. Another critical hurdle is selecting the appropriate attributes or variables from the multitude of available data. For oncologists, determining which factors—such as age, stage, genetic markers, lifestyle choices, or environmental exposures—significantly impact the disease's progression and treatment response proves to be a complex endeavour. Furthermore, identifying and interpreting trends within these chosen attributes poses yet another challenge. Cancer is a highly dynamic disease, influenced by numerous factors. Recognizing patterns and trends, such as understanding the evolution of certain genetic markers over time or correlating specific treatments with patient outcomes, demands advanced data analysis techniques. Therefore, employing Artificial Intelligence and Data Science, our aim is to support doctors by facilitating these intricate tasks and easing their workload.

IV. SCOPE OF PROJECT

The initial step involves data cleaning to enhance data quality by addressing missing values, outliers, and inconsistencies in the dataset. Subsequently, relevant features are extracted, taking into account attributes related to patient information, medical history, and treatments. The focus shifts to identifying specific outcomes of interest, such as patient survivability, and mapping corresponding attributes crucial for prediction. A correlation matrix is then generated to comprehend relationships among different variables, helping to identify potential predictors for the chosen outcome. Following this, a specific outcome, aligned with research objectives and stakeholder requirements (e.g., patient survivability), is selected. The process proceeds to training machine learning models tailored for the chosen outcome, employing algorithms like decision trees, random forests, or neural networks.

V. CONTRIBUTION

Cancer Risk Identification holds the premise of making significant contributions to healthcare and beyond. It aims to revolutionize cancer care by enabling early detection through machine learning model that can discern subtle risk factors [4]. This innovation has the potential to improve patient outcomes, reduce the global cancer burden by offering scalable solutions, and facilitate personalized treatment plans. By empowering healthcare professionals with advanced decision support tools and patients with personalized risk assessments, the project fosters greater awareness and proactive health measures. Moreover, it addresses ethical considerations and encourages cross-disciplinary collaboration, ensuring continuous innovation and ethical best practices in the field. Ultimately, this project represents a comprehensive effort to advance health-care through data-driven solutions, with far-reaching implications for cancer prevention and patient care.

VI. LITERATURE SURVEY

1) INTRODUCTION

In this comprehensive review of influential studies, significant advancements in the field cancer prognosis using machine learning techniques were explored. One notable study, conducted by Min Seob Kwak, Young-Gyu Eun, Jung-Woo Lee, and Young Chan Lee and published on March 16, 2021 [5], focused on predicting nodal metastasis in early T classification oral squamous cell carcinoma (OSCC). The research emphasized the critical nature of accurate lymph node metastasis (LNM) prediction, as it significantly impacts overall survival rates. Traditional methods, including imaging tests, often proved insufficient, leading the researchers to incorporate machine learning models. Six distinct models, namely XGBoost, Logistic Regression (LR), Support Vector Machine (SVM), Classification and Regression Trees (CART), Random Forest (RF), and k-Nearest Neighbours (kNN), were employed. These models, when applied to factors such as depth of invasion (DOI), patient's sex, tumor location, and histological features, demonstrated varying accuracies (XGBoost: 88.879 percent, LR: 70.656 percent, SVM: 69.666 percent, CART: 84.167 percent, RF: 83.331 percent, kNN: 77.476 percent). Challenges encountered included retrospective data limitations and missing information, crucial for machine learning algorithms.

Another noteworthy study, conducted by Rasheed Omobolaji Alabi, Alhadi Almangush, Elmusrati, Ilmo Leivo, and Antti A. M¨akitie and published on October 7, 2022 [6], focused on oropharyngeal squamous cell carcinoma (OPSCC). Researchers aimed to enhance OPSCC patient management using machine learning. Employing advanced techniques such as the voting ensemble and algorithms like XGBoost and Random Forest, a model was developed utilizing data from 3164 patients. This model effectively stratified patients into risk groups for overall survival, achieving an impressive accuracy of 88.3 percent. Key influencing factors included HPV status, patient age, cancer stage, marital status, N stage, and specific treatment methods. The study provided in-depth explanations for the model's decisions using frameworks like SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), empowering clinicians to make well-informed and personalized interventions for OPSCC patients. Challenges in this study arose from data extraction complexities, specifically in distinguishing specific cancer types within the heterogeneous data from the SEER Database.

As our model strives for prediction and projection, a study was referenced to estimate mortality due to cancer. The American Cancer Society (ACS) and the NCI work every 5 to 8 years to update the techniques for predicting the numbers of new cancer cases and fatalities in the current year for the U.S. and specific states. In this study [7], researchers compared existing approaches employed by the ACS and the NCI for predicting new cancer cases and fatalities with a new generation of statistical models. They performed a validation analysis utilizing incidence and death data from 1996 to 2010 and observed data projected ahead to 2014 for incidence and 2012–2015 for mortality. The assessment parameter employed was the average absolute relative deviation (AARD) between observed counts and estimations for different cancer locations nationwide and by state. The findings demonstrated that a unique Joinpoint model functioned well for both incidence and mortality, notably for the most frequent malignancies in the U.S. The AARD for tumors with cases over 49,000 in 2014 was much lower utilizing this approach compared to the existing techniques. The data-driven Joinpoint algorithm shows adaptable performance at both national and state levels and is intended

to replace the present ACS approaches. Overall, this new technique gives more precise estimates of cancer statistics for the current year, which is vital for advocacy, research, and public health planning initiatives.

Another study [8] presents a summary of the yearly estimates of new cancer cases and deaths in the United States produced by the American Cancer Society, employing data acquired up to 2018 for incidence and up to 2019 for mortality from different cancer registries. Projections for 2022 suggest 1,918,030 new cancer cases and 609,360 fatalities, with lung cancer being the major cause of cancer-related deaths, accounting for around 350 deaths each day. The data indicate a gradual increase in female breast cancer incidence (0.5% yearly) and constant rates for prostate cancer, despite a surge in advanced disease cases since 2011, leading to an increasing percentage of distant-stage diagnoses for prostate cancer. Conversely, lung cancer incidence exhibited a significant reduction for advanced stages while growing for localized stages (4.5% yearly), leading to better 3-year relative survival rates. Mortality rates matched similar incidence trends, with lung cancer witnessing faster drops, breast cancer seeing slower declines, and prostate cancer stable. The research underlines the need for focused cancer control programs, increased early detection techniques, and investment in therapeutic developments to further lower cancer death rates.

In a study by Xin Zhang, Guihong Liu and Xingchen Peng [19], we addressed the scarcity of integrated survival prediction tools for head and neck non-squamous cell carcinoma (HNnSCC), which is less common compared to squamous cell carcinoma. Leveraging data from 4458 HNnSCC patients obtained from the SEER database, we developed a novel prediction model for overall survival (OS) and disease-specific survival (DSS) at 3 and 5 years. The dataset was randomly divided into train & validation (70%) and test cohorts (30%), with tenfold cross-validation employed for model establishment. Multivariate analyses identified key prognostic factors, enabling the construction of a robust prediction model. Evaluation on the test cohort demonstrated high performance, with area under the curve (AUC) values of 0.866 (3-year OS), 0.862 (5-year OS), 0.902 (3-year DSS), and 0.903 (5-year DSS). Notably, the model's net benefit surpassed that of traditional prediction methods, underscoring its clinical utility. Pathology, involvement of cervical nodes level, and tumor size emerged as significant predictors contributing to variance in the model. Finally, the developed model was made accessible online for easy utilization by clinicians, representing a significant advancement in providing personalized prognostic information for post-treatment HNnSCC patients.

Lastly, a pioneering research paper by Siow-Wee Chang, Sameem Abdul-Kareem, Amir Feisal Merican, and Rosnah Binti Zain, published on May 31, 2013 [9], delved into the integration of clinicopathologic and genomic markers for oral cancer prognosis. Traditional prognostic decisions, reliant on clinicopathologic markers, often lacked precision. To address this, the study employed five feature selection methods to identify relevant parameters from a dataset encompassing both types of markers. Four machine learning classifiers, including ANFIS, artificial neural networks, support vector machines, and logistic regression, were then trained and rigorously tested using these selected features. The meticulous application of k-fold cross-validation ensured the robustness of the predictions. Remarkably, the hybrid model called ReliefF-GA-ANFIS, incorporating features such as drink, invasion, and p63, achieved an outstanding accuracy of 93.81 percent and an impressive AUC of 0.90 for oral cancer prognosis. This innovative approach showcases the potential of combining traditional clinical markers with genomic information, offering a significant advancement in oral cancer studies and paving the way for more accurate and personalized patient care.

2) PROBLEM DEFINITION

Upon careful consideration of the information presented, it is evident that early detection of cancer holds immense significance [4], potentially saving the lives of thousands of patients. Developing a machine learning model to assist doctors in patient diagnosis not only ensures accurate assessments but also aids in selecting the most suitable and cost-effective treatments. Leveraging historical data from SEER (Surveillance, Epidemiology, and End Result) is instrumental in achieving this goal. By doing so, we can enhance medical decision-making, leading to improved patient outcomes and advancing the field of cancer diagnosis and treatment.

VII. DESIGN IMPLEMENTATION

1. PROPOSED SYSTEM

In our proposed model, we integrate both traditional machine learning models and advanced neural networks. These models are meticulously crafted to analyze and utilize the dominant attributes identified through a thorough correlation matrix and feature importance analysis [10]. By harnessing the inherent power of these key features, our model delivers precise predictions and classifications. The primary objective of this approach is to offer invaluable support to doctors, aiding them in assessing patient survivability and determining appropriate treatments for those who have survived. Additionally, our model aims to assist doctors in diagnosing patients based on the historical data sourced from the SEER dataset.

2. REQUIREMENT GATHERING AND ANALYSIS

We utilized data sourced from SEER (Surveillance, Epidemiology, and End Results), a program conducted by the National Cancer Institute, offering comprehensive epidemiological insights into cancer incidence and survival rates in the United States. Our dataset encompassed cancer-related information pertaining to the head and neck region, spanning from the program's start until November 2022. To gain profound insights, we conducted an in-depth analysis, employing advanced techniques. Initially, a correlation matrix was applied, elucidating significant relationships within the dataset and highlighting dominant attributes influencing outcomes. Subsequently, we employed feature importance methodologies, enabling us to identify and prioritize the top 10 most influential attributes in our dataset. Hence this helped us is predicting and classifying the data

VIII. UML DIAGRAMS

1. FLOWCHART DIAGRAM

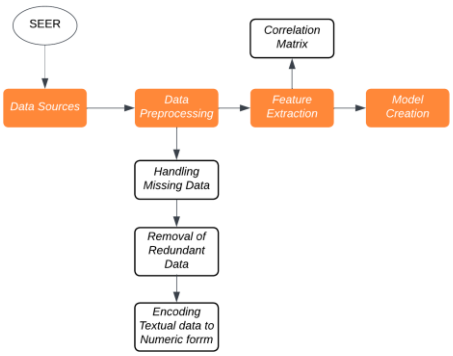

Figure 1-General Workflow

IX. RESULTS AND DISCUSSION

1. RESULTS

In this study, the focus was on exploring the survivability of patients and the year of death. Six classification models were employed to assess the survivability of patients: *Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, K-Nearest Neighbour Classifier, Naive Bayes Classifier, and XGBoost Classifier* [11]. Upon evaluating these classifiers, it was evident that the **XGBoost** model outperformed the others, boasting an impressive accuracy rate of **80%**. Notably, its precision, recall, and F1-scores for both the classes **0 (Patient did not survive)** and **1 (Patient survived)** were well-balanced, making it a reliable and robust choice. The Random Forest Classifier also displayed a commendable performance, achieving an accuracy of 79.07 percent and demonstrating good precision and recall for both classes. Similarly, the Gradient Boosting Classifier showcased strong results, with an accuracy rate of 79.53 percent and well-balanced precision and recall scores. The Decision Tree Classifier yielded decent results with an accuracy of 73.29 percent. Although it had slightly lower precision and recall values compared to the top performers, it remained a viable option for classification tasks. On the other hand, the K-Nearest Neighbours Classifier and Naive Bayes Classifier exhibited lower accuracies at 71.45 percent and 51.86 percent, respectively. These lower accuracy rates indicated the limitations of these classifiers for the given dataset. These findings underscore the critical importance of selecting an appropriate classifier tailored to specific requirements.

In this study, the XGBoost model emerged as the most effective choice, highlighting the significance of employing advanced machine learning techniques in healthcare contexts.

Table 1- Comparative Scoring of Classification Models

Classifier	Precision (0/1)	Recall (0/1)	F1-score (0/1)
Decision Tree	68%/77%	68%/77%	68%/77%
Random Forest	76%/81%	73%/83%	74%/82%
Gradient Boosting	75%/82%	75%/83%	75%/83%
K-Nearest Negihbors	65%/76%	67%/75%	66%/75%
Naïve Bayes	46%/83%	94%/22%	62%/35%
XGBoost	75%/83%	76%/82%	76%/83%

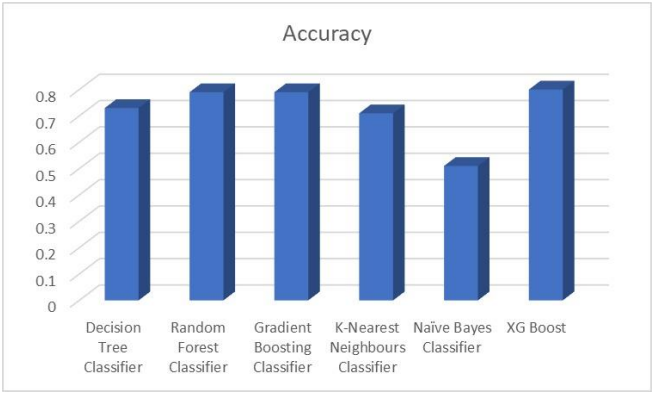

Figure 2 – Visual Representation of Models' Accuracy

For predicting the year of death, 9 regression models were tested. So, the Linear Regression model demonstrated a mean squared error of 18.68 and an R-squared (R2) score of 0.70, indicating a relatively good fit to the data. Similarly, the Random Forest Regression model exhibited a mean squared error of 18.28 and an R-squared score of 0.70, reflecting its strong predictive ability. The Gradient Boosting Regression model outperformed others with a mean squared error of 17.35 and an R-squared score of 0.72, indicating its superior accuracy in predicting outcomes. In contrast, the Support Vector Regression model showed a significantly higher mean squared error of 60.68 and a very low R-squared score of 0.01, suggesting poor predictive capability for the given data. The Multi-layer Perceptron Regression model yielded a mean square error of 28.31 and an R-squared score of 0.54, indicating moderate predictive accuracy. Additionally, the Ridge Regression model displayed a mean squared error of 18.67 and an R-squared score of 0.70, showcasing a performance similar to Linear Regression. Lasso Regression, with a mean squared error of 22.15 and an R-squared score of 0.64, and AdaBoost Regression, with a mean

squared error of 26.54 and an R-squared score of 0.57, fell in between, indicating moderate predictive capabilities.

Table 2 – Comparative Scoring of Regression Models

Regression Model	Mean Squared Error	R-squared (R2) Score
Linear Regression	18.68	0.70
Decision Tree Regression	34.75	0.43
Random Forest Regression	18.28	0.70
Gradient Boosting Regression	17.35	0.72
Support Vector Regression	60.68	0.01
Multi-layer Perceptron Regression	28.31	0.54
Ridge Regression	18.67	0.70
Lasso Regression	22.15	0.64
AdaBoost Regression	26.54	0.57

In our search for productive cancer treatment, we went into studying the sequence of therapies, concentrating on features critical for success. Across three main outcomes - *chemotherapy, radiation sequence, and surgery* - we applied multiple models to test their relative prediction accuracy.

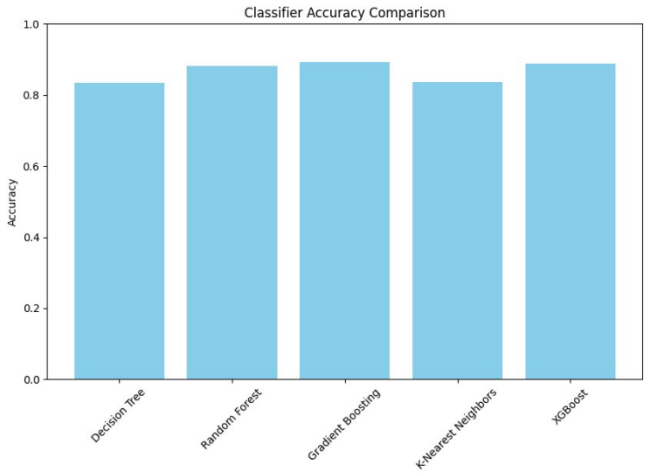

Figure 3 – Accuracy of Models for Chemotherapy

Among the models assessed for chemotherapy, gradient boosting and XGBoost emerged as the highest performers, with both reaching an accuracy of 84%. These findings underline the strength of ensemble approaches in capturing complicated patterns and boosting forecast accuracy. Random Forest showed great performance with an accuracy of 82%, suggesting its effectiveness in decreasing overfitting and enhancing generalization. Meanwhile, Decision Tree and K Nearest Neighbors (KNN) showed equal accuracies of 78%, showing modest performance compared to the ensemble approaches. These results give useful insights into the efficiency of various machine learning methodologies, offering help in picking the most suitable models for future predictive modeling jobs.

Table 3 – Scores for Chemotherapy as an Outcome

Classifier	Precision (0/1)	Recall (0/1)	F1-score (0/1)
Decision Tree	83%/70%	83%/70%	83%/70%
Random Forest	86%/76%	86%/76%	86%/76%
Gradient Boosting	88%/76%	86%/80%	87%/78%
K-Nearest Negihbors	83%/69%	82%/71%	83%/70%
XGBoost	89%/75%	85%/80%	87%/78%

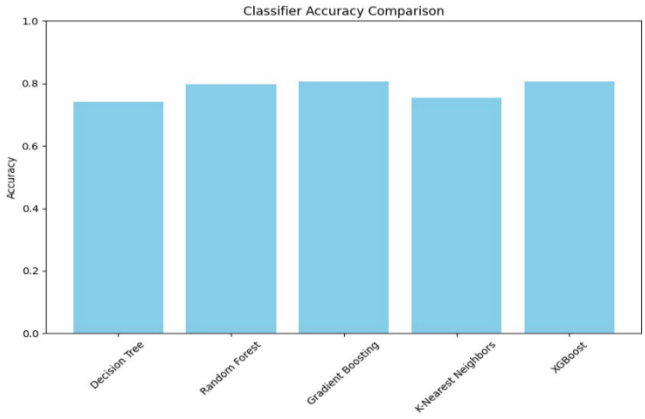

Figure 4 – Accuracy of Models for Surgery

Similar sets of models were examined to discover the greatest accuracy on whether the patient should have a surgery or not. Among these models, Decision Tree, Gradient Boosting, and XGBoost all displayed high accuracy rates of 84%, suggesting their usefulness in correctly predicting the need for surgery. Random Forest also displayed good performance with an accuracy of 82%, displaying its abilities to generalize effectively and deliver solid predictions. In comparison, K Nearest Neighbors (KNN) demonstrated a significantly lower accuracy of 78%, showing possible limits in its capacity to determine surgical needs properly. These findings illustrate the usefulness of Decision Tree, Gradient Boosting, Random Forest, and XGBoost in surgical prediction tasks, delivering significant insights for enhancing treatment planning and patient care techniques.

Table 4- Scores for Surgery as an Outcome

Classifier	Precision (0/1)	Recall (0/1)	F1-score (0/1)
Decision Tree	74%/88%	74%/88%	74%/88%
Random Forest	82%/91%	80%/92%	81%/91%
Gradient Boosting	83%/92%	83%/92%	83%/92%

K-Nearest Negihbors	74%/88%	74%/88%	74%/88%
XGBoost	82%/92%	83%/91%	83%/92%

Next, we tackled one of the most critical components of our project: predicting the sequence of therapy needed by cancer patients. These sequences, grouped under RX Summ--Surg/Rad Seq, outline the order of therapies essential for optimal management. They comprise situations such as no radiation and/or cancer-directed surgery, radiation after surgery, radiation prior to surgery, radiation before and after surgery, intraoperative radiation with other radiation before or after surgery, and surgery both before and after radiation. To obtain precise predictions, we once again deployed multiple trained models, seeking to determine the most successful strategy.

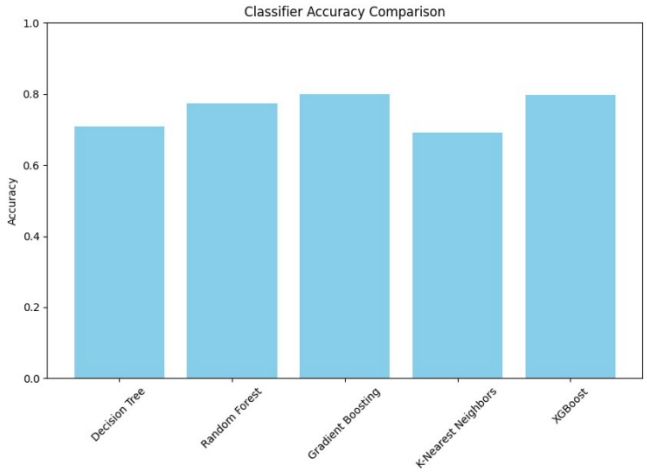

Figure 5 – Accuracy of models for Radiation Sequence

Among these models, Gradient Boosting and XGBoost achieved the greatest accuracy rates of 80%, indicating their usefulness in correctly forecasting the sequence of radiation treatments. Random Forest also exhibited great performance, with an accuracy of 77%, suggesting its capacity to generalize effectively and make solid predictions. However, Decision Tree and K Nearest Neighbors (KNN) revealed lesser accuracies of 71% and 69%, respectively, highlighting possible limits in their prediction powers for radiation sequencing. These findings underline the efficiency of Gradient Boosting, XGBoost, and Random Forest in radiation sequence prediction tasks, giving significant insights for enhancing treatment planning and patient care methods connected to radiation therapy [18].

*Table 5 – Scores for Radiation Sequence Testing with **XGBoost Model***

Radiation Sequence	**Precision**	**Recall**	**F1-score**
0 – No Radiation	0%	0%	0%
1 – After Surgery	82%	89%	85%
2 – Before Surgery	75%	67%	71%
3 – Before and After Surgery	0%	0%	0%
4 – Intraoperative Before and After Surgery	44%	7%	12%
5 – Surgery Before and After Radiation	0%	0%	0%

During this examination of radiation sequences, five separate phases are determined depending on the time of radiation delivery in relation to surgery. The first stage involves no radiation, resulting in a precision, recall, and F1-score of 0%, indicating that this category was not present. Following the surgical procedure, the second stage demonstrates a significant enhancement in accuracy, measuring at 82%, and in the ability to correctly identify relevant instances, measuring at 89%. This improvement results in an F1-score of 85%. In contrast, the third stage, prior to surgery, shows a significantly lower level of accuracy at 75% and completeness at 67%, leading to an F1-score of 71%. Nevertheless, the dataset does not include any instances of the latter two stages, which involve a mix of radiation and surgical schedules. As a result, the precision, recall, and F1-score for these stages are all 0%. This investigation highlights the significance of taking into account the timing of radiation treatments in relation to surgical operations in order to enhance accuracy, recollection, and overall performance in clinical settings.

2. DISCUSSION

In our research, we focused on extracting pertinent data related to head and neck cancers from the SEER dataset. From this extensive dataset, we identified key outcomes including Survivability of the Patient, Year of Death, as well as information about treatments such as Chemotherapy, Radiation, and Surgery. These specific outcomes are invaluable as they serve as crucial factors for making predictions and classifications related to cancer diagnoses and patient outcomes. To gain deeper insights and understand the relationships between these variables, we constructed a correlation matrix. This matrix allowed us to comprehensively study the interconnections among the attributes, enabling us to draw meaningful conclusions and make informed decisions based on the intricate relationships within the data. The correlation analysis provided valuable insights that are fundamental in enhancing our understanding of the complexities of head and neck cancers, ultimately giving us dominant attributes to focus on for creating an accurate model.

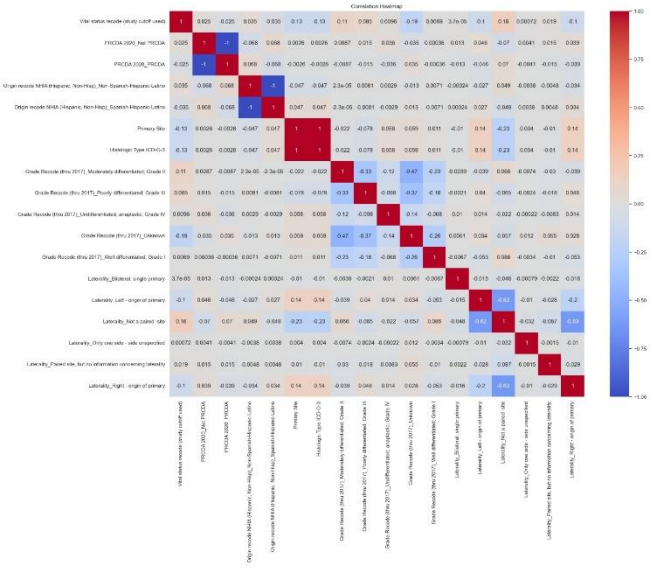

Figure 6 – Correlation Matrix generated

In determining the survivability of the patient, several key attributes played a significant role. They were primary site, histologic type ICD-O-3 coding, grade recode (up to 2017), laterality, SS seq indicating malignancy status, marital status at diagnosis, median household income adjusted for inflation up to 2021, record number recode, total number of in situ/malignant tumors for the patient, year of follow-up recode, race recode (categorized as W: *White*, B: *Black or African American*, AI: *American Indian or Alaska Native* , API: *Asian or Pacific Islander*), first malignant primary indicator, RX SummSurg/Rad Seq detailing surgery and radiation sequences, reasons for the absence of cancer-directed surgery, chemotherapy information, positive regional nodes, total regional nodes examined, year of diagnosis, age recoded with specific brackets (including individuals aged 90 and above), and broader site recode. These attributes collectively contributed significantly to developing accurate predictive models.

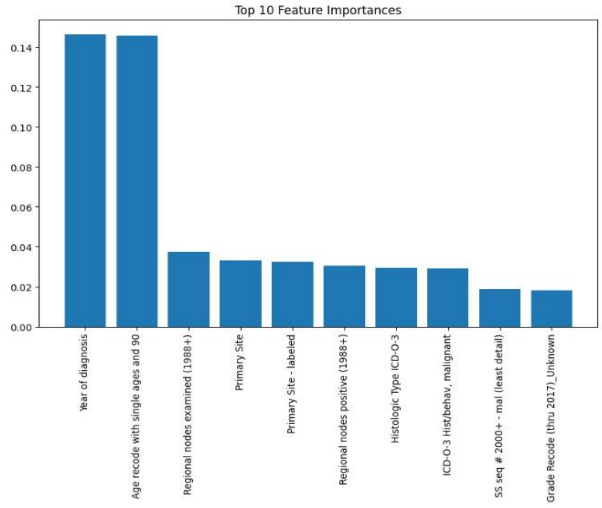

Figure 7 – Feature Importance Graph – Survivability

While it came to predict the Year of Death, the attributes Age recode with single ages and 90+, Year of diagnosis, TNM 7/CS v0204+ Schema recode, Race Asian, Pacific Islander, Black, Site recode ICD-O-3 2023 Revision, Primary Site, Histologic Type ICD-O-3, Grade Recode, Laterality, Site recode ICD-O-3/WHO 2008 (for SIRs), SEER historic stage, RX Summ, Reason no cancer-directed surgery, Radiation recode Combination of beam with implants or isotopes, Chemotherapy recode (yes, no/unk), Regional nodes, COD to site recode ICD-O-3 2023 Revision Expanded (1999+), Year of follow-up recode, Vital status, First malignant primary indicator, COD to site recode, COD to site recode, SEER cause-specific death classification, Survival months flag were found to be dominant

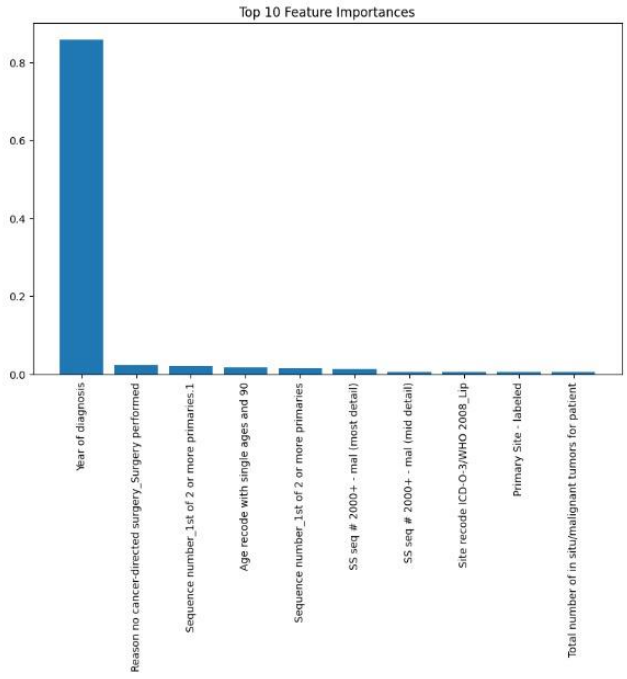

Figure 8 – Feature Importance Graph – Year of Death

The metric used to calculate feature importances in tree-based models like decision trees, random forests, and gradient boosting machines is typically Gini impurity or entropy (information gain). These metrics measure the impurity of a node in the decision tree. A node is considered pure (i.e., containing only one class) when all the data points in that node belong to the same class. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the set. Entropy, on the other hand, measures the amount of disorder or uncertainty in the set. Lower entropy means less disorder and higher information gain. In our analysis, we employed classical machine learning models, as well as neural networks, specifically utilizing an RNN (Recurrent Neural Network). The RNN model yielded a commendable performance with an R-Squared score of 0.7, indicating a strong correlation between the predicted and actual values. Furthermore, the test loss was recorded at 22.43 percent, signifying the accuracy of our neural network model in predicting the target variable.

In our search for productive cancer treatment, we went into studying the sequence of therapies, concentrating on features critical for success. Across three main outcomes—chemotherapy, radiation sequence, and surgery—we applied multiple models to test their relative prediction accuracy.

For Chemotheraphy, further study comprised grading all attributes (relatively on a scale of 0-1) using our model and picking the top 10 with their associated scores. Notably, "Site recode: rare tumors_5.1 Squamous cell carcinoma with variants of oropharynx" emerged as the most important characteristic with a score of 0.25, followed by "year of diagnosis" at 0.15, and "site recode: ICD-0-3/WHO

2008_Nasopharynx" at 0.14. Other major features were "Age recode with single ages and 90" (0.075), "Regional nodes positive (1988+)" (0.067), and "TNM 7/CS V0204+ Schema recode_Hypopharynx" (0.06). Additionally, "Regional nodes examined (1988+)" (0.05), "TNM 7/CS V0204+ Schema recode_Nasopharynx" (0.04), "Site recode ICD-0-3/WHO 2008_Lip" (0.02), and "Histologic Type ICD-0-3" (0.019) were discovered as relevant variables. These results give useful insights into the major parameters driving predictive performance, supporting informed decision-making about tumor prognosis and treatment methods.

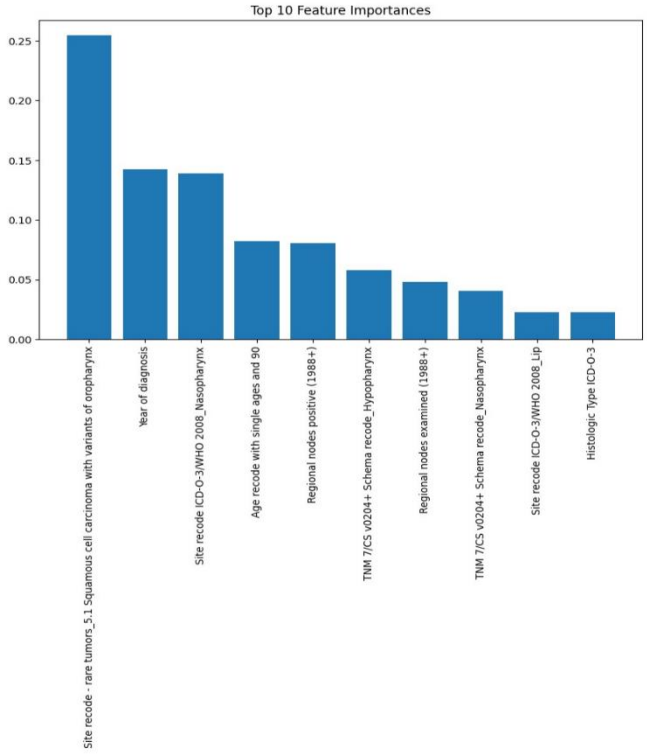

Figure 9 – Feature Importance Graph - Chemotherapy

The examination of critical characteristics for predictive modeling in cancer therapy showed various relevant parameters, each awarded a value on a scale from 0 to 1. Topping the list is "Regional nodes positive (1988+)" with a grade of 0.28, followed closely by "Site recode rare tumors_5.1 Squamous cell carcinoma with variants of oropharynx" at 0.27. These qualities show a major influence on treatment results, presumably suggesting the severity or spread of the illness. Following this, "Site recode ICD-0-3/WHO 2008_Nasopharynx" and "Regional nodes examined (1988+)" earned scores of 0.16 and 0.085, respectively, underlining the need of addressing both the location and level of nodal involvement in treatment planning. Further down the list, characteristics such as "TNM 7/CS V0204+ Schema recode_Nasopharynx" and "Year of diagnosis" got ratings of 0.04 and 0.025, respectively, demonstrating their comparatively smaller but still important contributions to prediction accuracy. Additionally, traits like "Primary Site-labeled_C01.9-Base of tongue, NOS" and "Histologic Type ICD-0-3" were awarded scores of 0.025 and 0.024, showing their potential usefulness in identifying tumor characteristics and directing treatment options. Lastly, "Grade Recode (thru 2017)_Unknown" and "Site recode ICD-0-3/WHO 2008_Hypopharynx" earned scores of 0.023 and 0.02, respectively, demonstrating their considerably smaller influence on predictive modeling results. These results give useful insights into the relative relevance of numerous traits in predicting cancer treatment outcomes, contributing in the creation of more effective and tailored treatment options.

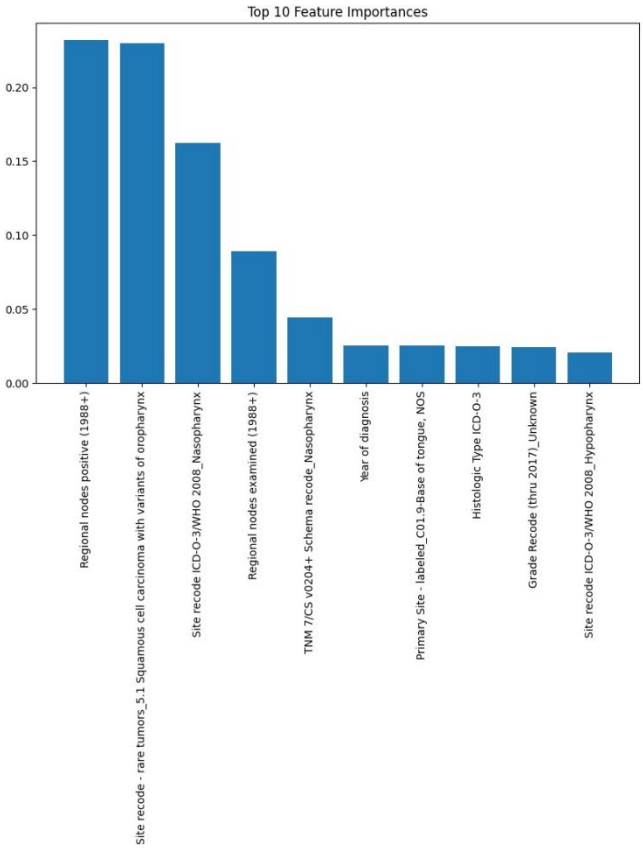

Figure 10 – Feature Importance Graph - Surgery

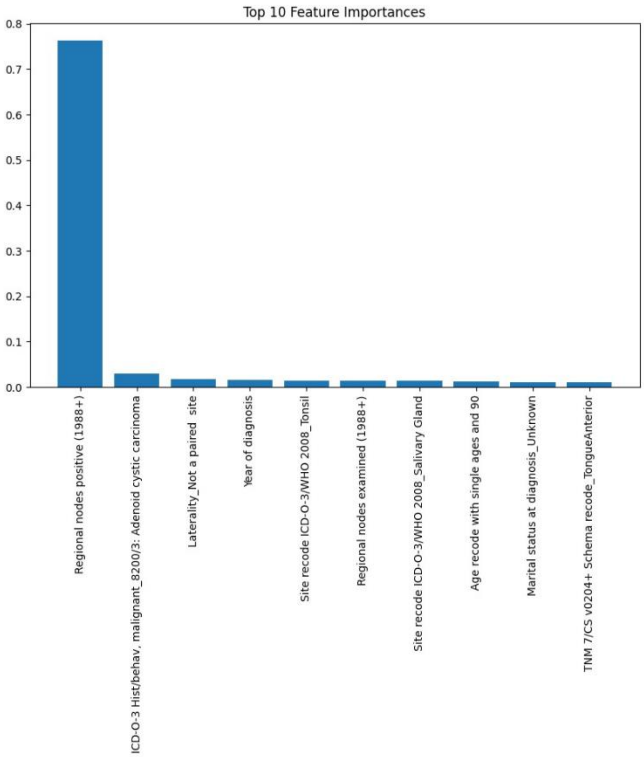

Figure 11- Feature Importance Graph – Radiation Sequence

Among the attributes considered for predictive modeling in cancer treatment, "Regional nodes positive (1988+)" stands out as the most influential, with a substantial importance score of 0.75, indicating its significant impact on treatment outcomes. Conversely, several other attributes, including "ICD-O-3 Hist/behav, malignant_8200/3: Adenoid cystic carcinoma," "Laterality_Not a paired site," "Year of diagnosis," "Site recode ICD-0-3/WHO 2008_Tonsil," "Regional nodes examined (1988+)," "Site recode ICD-0-3/WHO 2008_Salivary Gland," "Age recode with single ages and 90," "Marital status at diagnosis_Unknown," and "TNM 7/CS V0204+ Schema recode_TongueAnterior" all displayed importance scores below 0.1. While these attributes may contribute to the predictive model to some extent, their lesser importance underscores their relatively minor impact compared to "regional nodes positive (1988+)," suggesting that nodal involvement plays a critical role in determining treatment strategies and patient outcomes. These insights are invaluable for refining predictive models and optimizing treatment decisions in cancer care.

X. CONCLUSION

1. SUMMARY

In conclusion, our exploration of the SEER dataset has provided valuable insights into cancer-related outcomes. We identified a specific set of outcomes, including Survivability of the Patient, Year of Death, Chemotherapy, Radiation, and Surgery. To understand the relationships between these outcomes and the dataset's attributes, we conducted a comprehensive correlation analysis using a correlation matrix. Through this analysis, we determined the dominant attributes associated with each outcome. For example, when considering Survivability of the Patient, attributes such as Age Recode with Singles Ages 90+, Year of Diagnosis, and Vital Status Recode emerged as significant factors influencing the outcome. Similarly, for Year of Death prediction, attributes like Age Recode with Singles Ages 90+and Year of Diagnosis. These attributes helped us in creating models for classification as well as prediciton.

2. FUTURE SCOPE

In our exploration of the SEER dataset, we delved into two specific outcomes among the identified five. However, three outcomes remain unexplored. Our analysis revealed accuracies ranging from 60 percent to 80 percent for the outcomes investigated. To enhance the accuracy and robustness of our models, we propose the inclusion of additional data. This supplementary information could encompass clinical data and details about patients' habits, such as drinking and smoking patterns. By incorporating these factors, we anticipate the development of more comprehensive and accurate predictive models. The introduction of these additional variables is expected to refine our understanding of the intricate dynamics influencing cancer-related outcomes, thereby paving the way for more precise predictions and informed interventions in the realm of oncology.

XI. REFERENCES

1 Surveillance, Epidemiology, and End Results Program. (n.d.). SEER. https://seer.cancer.gov/

2 Comprehensive Cancer Information. (n.d.). National Cancer Institute. https://www.cancer.gov/

3 SEER Cancer Stat Facts. (n.d.). SEER. https://seer.cancer.gov/statfacts/

4 Walker JG, Licqurish S, Chiang PP, Pirotta M, Emery JD. Cancer risk assessment tools in primary care: a systematic review of randomized controlled trials. Ann Fam Med. 2015 Sep;13(5):480-9. doi: 10.1370/afm.1837. PMID: 26371271; PMCID: PMC4569458.

5 Kwak MS, Eun YG, Lee JW, Lee YC. Development of a machine learning model for the prediction of nodal metastasis in early T classification oral squamous cell carcinoma: SEER-based population study. Head Neck. 2021 Aug;43(8):2316-2324. doi: 10.1002/hed.26700. Epub 2021 Mar 31. PMID: 33792112.

6 Rasheed Omobolaji Alabi, Alhadi Almangush, Mohammed Elmusrati, Ilmo Leivo, Antti A. Mäkitie,
An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer,
International Journal of Medical Informatics,
Volume 168,
2022,
104896,
ISSN 1386-5056,
https://doi.org/10.1016/j.ijmedinf.2022.104896.
(https://www.sciencedirect.com/science/article/pii/S1386505622002106)

7 Miller KD, Siegel RL, Liu B, Zhu L, Zou J, Jemal A, Feuer EJ, Chen HS. Updated Methodology for Projecting U.S.- and State-Level Cancer Counts for the Current Calendar Year: Part II: Evaluation of Incidence and Mortality Projection Methods. Cancer Epidemiol Biomarkers Prev. 2021 Nov;30(11):1993-2000. doi: 10.1158/1055-9965.EPI-20-1780. Epub 2021 Aug 17. PMID: 34404684.

8 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022 Jan;72(1):7-33. doi: 10.3322/caac.21708. Epub 2022 Jan 12. PMID: 35020204.

9 Chang, SW., Abdul-Kareem, S., Merican, A.F. *et al.* Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics* **14**, 170 (2013). https://doi.org/10.1186/1471-2105-14-170

10 McCloskey, B. (2022, June 9). *Feature Selection for Data Science: Simple Approaches*. Medium. https://towardsdatascience.com/feature-selection-for-data-science-simple-approaches-e1f2527cb363

11 Xiong D, Zhang Z, Wang T, Wang X. A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences. Comput Struct Biotechnol J. 2021 May 24;19:3255-3268. doi: 10.1016/j.csbj.2021.05.038. PMID: 34141144; PMCID: PMC8192570.

12 Mourad M, Moubayed S, Dezube A, Mourad Y, Park K, Torreblanca-Zanca A, Torrecilla JS, Cancilla JC, Wang J. Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. Sci Rep. 2020 Mar 20;10(1):5176. doi: 10.1038/s41598-020-62023-w. PMID: 32198433; PMCID: PMC7083829.

13 Zhan X, Cheng J, Huang Z, Han Z, Helm B, Liu X, Zhang J, Wang TF, Ni D, Huang K. Correlation Analysis of Histopathology and Proteogenomics Data for Breast Cancer. Mol Cell Proteomics. 2019 Aug 9;18(8 suppl 1):S37-S51. doi: 10.1074/mcp.RA118.001232. Epub 2019 Jul 8. PMID: 31285282; PMCID: PMC6692775.

14 Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis,
Machine learning applications in cancer prognosis and prediction,
Computational and Structural Biotechnology Journal,
Volume 13,
2015,
Pages 8-17,
ISSN 2001-0370,
https://doi.org/10.1016/j.csbj.2014.11.005.
(https://www.sciencedirect.com/science/article/pii/S2001037014000464)

15 You, Y., Lai, X., Pan, Y. *et al.* Artificial intelligence in cancer target identification and drug discovery. *Sig Transduct Target Ther* **7**, 156 (2022). https://doi.org/10.1038/s41392-022-00994-0

16 Tran, K.A., Kondrashova, O., Bradley, A. *et al.* Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* **13**, 152 (2021). https://doi.org/10.1186/s13073-021-00968-x

17 Tsopra, R., Fernandez, X., Luchinat, C. *et al.* A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. *BMC Med Inform Decis Mak* **21**, 274 (2021). https://doi.org/10.1186/s12911-021-01634-3

18 Hansen CR, Crijns W, Hussein M, Rossi L, Gallego P, Verbakel W, Unkelbach J, Thwaites D, Heijmen B. Radiotherapy Treatment plannINg study Guidelines (RATING): A framework for setting up and reporting on scientific treatment planning studies. Radiother Oncol. 2020 Dec;153:67-78. doi: 10.1016/j.radonc.2020.09.033. Epub 2020 Sep 22. PMID: 32976873.

19 Zhang X, Liu G, Peng X. A Random Forest Model for Post-Treatment Survival Prediction in Patients with Non-Squamous Cell Carcinoma of the Head and Neck. J Clin Med. 2023 Jul 30;12(15):5015. doi: 10.3390/jcm12155015. PMID: 37568416; PMCID: PMC10419643.