PREDICTION OF CANCER DIAGNOSIS USING LOGISTIC REGRESSION AND DECISION TREE CLASSIFIERS

M.Devendiran , Dr.S.Sasikala

Research Scholar, Thanthai Periyar Government Arts and Science College, Trichy
Associate Professor and Research Supervisor, P.G. & Research Department of Statistics,
Thanthai Periyar Government Arts and Science College, Trichy
(Affiliated to Bharathidasan University)

Abstract

Cancer remains a leading cause of morbidity and mortality worldwide, necessitating improved diagnostic methods for early detection and prevention. This study investigates the application of machine learning techniques, specifically Logistic Regression (LR) and Decision Tree (DT) classifiers, for predicting cancer diagnosis using the Wisconsin dataset. The dataset, comprising 1,500 patient records with demographic, lifestyle, and clinical attributes, was selected due to its rich feature set and reliability. Logistic regression analysis revealed significant predictors of cancer, including age, gender, body mass index (BMI), smoking status, genetic risk, alcohol intake, and prior cancer history, while physical activity showed a protective effect. Model performance was evaluated using confusion matrices, accuracy, precision, recall, F1-score, and ROC-AUC metrics. Results indicated that LR achieved strong predictive accuracy (optimal threshold 0.4–0.5, AUC = 0.91), balancing sensitivity and specificity effectively. The DT model provided interpretable decision rules, identifying cancer history, genetic risk, BMI, and age as key discriminators. Both approaches highlight the potential of regression-based and tree-based models in supporting early cancer diagnosis. These findings demonstrate the utility of predictive modeling in clinical decision-making and underscore the importance of integrating lifestyle and genetic risk factors into diagnostic tools.

Keywords : Machine learning techniques, Logistic Regression, Decision Tree, Body Mass Index

Introduction :

Cancer is a broad term used to describe a group of diseases that involve the uncontrolled growth and spread of abnormal cells in the body. Normally, the body's cells grow, divide, and die in a controlled manner. However, in cancer, this process is disrupted, causing cells to grow uncontrollably, forming tumors (masses of tissue) or spreading to other parts of the body through the bloodstream or lymphatic system. Cancer can occur in virtually any part of the body and is typically named based on where it originates. For example, cancer that begins in the lungs is called lung cancer, while cancer that begins in the skin is called skin cancer. There are over 100 different types of cancer, with some of the most common being breast cancer, lung cancer, prostate cancer, and colorectal cancer.

The causes of cancer are diverse and may include genetic mutations, biological science factors (such as tobacco smoke, radiation, or exposure to chemicals), or lifestyle choices (like diet, physical activity, and alcohol consumption). In many cases, a combination of these factors leads to the development of cancer. The disease can be diagnosed using various methods, including medical imaging (like CT scans, MRIs),blood tests and biopsies, where a sample of tissue is taken for examination. Treatment options depend on the type, stage and location of cancer as well as the patient's overall health. These options include surgery, radiation therapy, chemotherapy, immunotherapy and targeted therapies aimed at specific genetic mutations or proteins involved in the cancer process. Early detection and advances in treatment have significantly improved the prognosis for many types of cancer, though research is ongoing to find more effective and less invasive treatments.Preventive measures, such as lifestyle changes and regular screenings, are also essential in reducing.

Previous work :

Numerous researchers have explored disease prediction using various regression methodologies. This article highlights several studies to illustrate current trends in predictive modeling through regression analysis. In [1], the author conducted a survey on disease outbreak prediction using multiple regression techniques. Paper [2] compared different machine learning methods, including the use of linear regression for predicting breast cancer rates using the UCI dataset, achieving a notable accuracy of 99.06%. In [3], the authors applied logistic regression with CUDA-based parallel programming to predict breast cancer, yielding results closely aligned with real-world cases. In [4], a similar approach was taken, reinforcing the effectiveness of logistic regression with computational support to enhance performance. Study [5] focused on developing a logistic regression-based clinical prediction model, emphasizing its economic benefits and relevance to real-time healthcare data. In [6], the authors discussed the importance of regression techniques in healthcare analytics and emphasized how these models can improve patient care when integrated with electronic health records and clinical systems. Similarly, [7] examined factors influencing breast cancer survivability using various machine learning models, including logistic regression. Paper [8] employed a multiple linear regression model for heart disease prediction and reported satisfactory accuracy. Lastly, in [9], researchers conducted a comparative study between decision trees and logistic regression to predict breast cancer survival rates, analyzing factors that influence patient outcomes. Collectively, these studies underscore the effectiveness of regression techniques in medical prediction and decision-making. Marian Constantin et al.,[10] discussed the major genetic alterations involved in appendix tumors and their roles in promoting cell growth and survival, invasiveness, angiogenesis, and resistance to growth-inhibitory signals. Jin-Hee Kwon et al. [11] analyzed 96,174 cancer patients and found a 2.3% incidence of multiple primary cancers (MPCs). Breast cancer was the most common first malignancy, while lung cancer was the most frequent second. The study highlighted decreasing intervals between cancers and identified lymphoma as having the highest risk for MPCs, offering important insights for clinical management. Malhotra et al. [12] discussed tobacco products are the major contributors for various cancers and other diseases. In India, tobacco-related cancers (TRCs) contribute nearly half of the total cancers in males and one-fifth in females. Wang et.al.,[14] studies have indicated that an imbalance in the structural and signaling properties of β-catenin leads to various cancers, such as cervical cancer.

Summary of review work

The choice of the Wisconsin cancer dataset was primarily motivated by the limited availability of real-time cancer datasets, which guided the objective of this article and the selection of data. This dataset includes information from 1,500 cancer patients and is notable for its extensive number of features, making it a valuable resource for research. Additionally, regression methods—widely recognized for their effectiveness in similar predictive tasks—were selected for analysis. Given these factors, the Wisconsin cancer dataset was deemed the most suitable for the current study. This paper is structured to present insightful research findings derived from the available data and relevant literature, contributing meaningfully to the field.

Methods and models:

The present section discusses the regression technique, Wisconsin data set description, the algorithm used in the prediction and methodology of the work. The methodology aims to analyse the most helpful feature in prediction of Cancer. It may help to visualize general trend in selecting appropriate model. The objective is to predict of cancer with the help of python. The focus is on using Logistic Regression, Decision Tree Classifier (DT).

Logistic Regression technique:

It is used to predict categorical outcomes, unlike linear regression that predicts a continuous outcome. In the simplest case there are two outcomes, which is called binomial. A basic machine learning approach that is frequently used for binary classification tasks is called logistic regression. Binary Logistic Regression (Logit Model) is used when the target variable has two categories (binary classification). The model uses a logistic (sigmoid) function to predict the probability that an observation belongs to the positive class (usually represented as 1). In logistic regression, odds and probabilities are closely related but represent different concepts.

Logistic Function:

The probability is modeled using the logistic function, which is given by: $P(y=1 \square X)= \frac{1}{1+e^{-z}}$ where is a linear combination of the features (i.e., Y = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n X_n$ and e is the base of the natural logarithm. This function outputs values between 0 and 1, representing the predicted probability of the event occurring.

Decision Tree :

In machine learning, decision trees are one of the most intuitive and widely-used algorithms for classification and regression tasks. It is the model of decision-making processes in a tree-like structure where each internal node represents a feature (attribute), each branch represents a decision rule, and each leaf node represents the outcome. Decision trees are easy to visualize and interpret, making them an excellent tool for explaining the logic behind predictions.

A decision tree is a supervised learning algorithm that can be used for both classification and regression tasks. It works by recursively partitioning the data into subsets based on feature values, making decisions at each node to maximize a specific criterion (e.g., information gain or Gini index).

The process of creating a decision tree involves are Selecting the Best Attribute, Repeating the Process and Dataset to check mathematically if any pure split or not we use entropy or Gini impurity. Entropy is a concept borrowed from information theory and is commonly used as a measure of uncertainty or disorder in a set of data. In the context of decision trees, entropy is often employed as a criterion to decide how to split data points at each node, aiming to create subsets that are more homogeneous with respect to the target variable.

Impurity Measures: Gini impurity is a measure of the impurity or disorder in a set of elements, commonly used in decision tree algorithms, especially for classification tasks. It quantifies the likelihood of misclassification of a randomly chosen element in the dataset.

$$\text{Entrophy}(S) = -\sum_{i=1}^{c} p_i \log_2 (p_i) \qquad \text{Gini}(S) = 1 - \sum_{i=1}^{c} pi^2$$

Confusion Matrix:

A confusion matrix is a powerful tool used in statistics and machine learning to evaluate the performance of classification models, especially when simply looking at accuracy isn't enough. It provides a detailed breakdown of correct and incorrect predictions made by a classification algorithm, allowing for a more nuanced understanding of how a model performs. A confusion matrix is a table, typically 2x2 for binary classification problems (two possible outcomes) or N x N for multiclass classification problems (N possible outcomes). The rows represent the actual classes or true values of the target variable. The columns represent the predicted classes or values predicted by the model. For a binary classification problem, the confusion matrix typically has four components: True Positive (TP): Instances where the actual class is positive, and the model correctly predicted it as positive. True Negative (TN): Instances where the actual class is negative, and the model correctly predicted it as negative**.** False Positive (FP): Instances where the actual class is negative, but the model incorrectly predicted it as positive (a Type I error). False Negative (FN): Instances where the actual class is positive, but the model incorrectly predicted it as negative (a Type II error).

Diagonal Elements represent the correct predictions made by the model for each class. Off-Diagonal Elements represent the misclassifications, or the instances where the model "confused" one class for another.

Accuracy measures the overall correctness of the model

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision (Positive Predictive Value): Proportion of Correctly Predicted positives out of all predicted positives

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall (Sensitivity or True Positive Rate): Proportion of Correctly Predicted positives out of all actual positives

$$\text{Precision} = \frac{TP}{TP+FN}$$

F1 Score: Harmonic Mean of precision and recall. A good balance between the two

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recal}$$

ROC Curve (Receiver Operating Characteristic): A plot of True positive Rate (Recall) vs. False Positive Rate(FPR) at various thresholds.

$$\text{FPR} = \frac{FP}{PP+TN}$$

AUC (Area Under the ROC Curve): Ranges from 0 to 1. Closer to 1 means better model performance

Specificity(True Negative Root) : Proportion of correctly predicted negatives out of all actual negatives

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Table-1 : Data set description :

S.No.	Attribute name	Attribute type
1.	Age	integer, patient's age
2.	Gender	integer, patient's gender (0 – Male, 1 – Female)
3.	BMI	float
4.	Smoking	integer – dichotomous (0 – No, 1 – Yes)
5.	Genetic risk	nominal (0-Low, 1-Medium, 2-High)
6.	Physical-activity (MET)*	float
7.	Alcohol intake (ABV)**	float
8.	Cancer history	integer – dichotomous (0 – No, 1 – Yes)
9.	Diagnosis	integer– dichotomous (0 – Absent, 1 – Present)

** MET = Metabolic Equivalent of Task; ** ABV = Alcohol By Volume*

Table -2 : Binary Logistic Regression model estimates

	Coefficient	standard error	z	P>|z|	[0.025	0.975]
constant	-10.1177	0.632	-16.011	0.000	-11.356	-8.879
Age	0.0523	0.005	10.412	0.000	0.042	0.062
Gender	2.0841	0.177	11.804	0.000	1.738	2.430
BMI	0.1172	0.012	9.891	0.000	0.094	0.140
Smoking	1.9645	0.185	10.	0.000	1.603	2.326
Genetic Risk	1.5572	0.125	12.456	0.000	1.312	1.802
Physical Activity	-0.2431	0.029	-8.292	0.000	-0.301	-0.186
Alcohol Intake	0.6242	0.061	10.173	0.000	0.504	0.745
Cancer History	4.2624	0.297	14.346	0.000	3.680	4.845

The binary logistic regression model identified several significant predictors of the outcome, all with p-values < 0.001(Table-2). Age, gender, BMI, smoking status,genetic risk, alcohol intake, and cancer history were positively associated with increased odds of the outcome, while physical activity showed a protective effect. Specifically, each additional year of age increased the odds by approximately 5.4%, and each unit increase in BMI raised

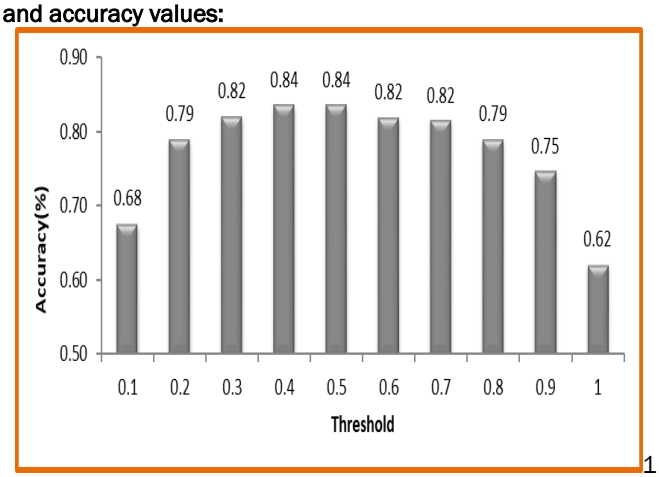
the odds by about 12.4%. Being male was associated with over 8 times higher odds compared to f emales. Smoking and genetic risk significantly raised the odds by about 7 and 4.7 times, respectively.Alcohol intake nearly doubled the odds, and individuals with a history of cancerhad a dramatically higher risk approximately 71 times greater. Conversely, physical activity was linked to a 21.6% reduction in odds.These findings underscore the importance of lifestyle and genetic factors in influencing the likelihood of the outcome.

Table -3 : Confusion matrix for the cut-off values:

Cut-off value	Classification	Precision	Recall	F1-score	Support	Accuracy
0.1	0 1	0.96 0.54	0.50 0.96	0.66 0.69	326 199	0.68
0.2	0 1	0.95 0.66	0.70 0.93	0.81 0.77	326 199	0.79
0.3	0 1	0.92 0.71	0.78 0.88	0.84 0.79	326 199	0.82
0.4	0 1	0.88 0.78	0.86 0.80	0.87 0.79	326 199	0.84
0.5	0 1	0.85 0.82	0.90 0.73	0.87 0.77	326 199	0.84
0.6	0 1	0.80 0.86	0.94 0.63	0.87 0.72	326 199	0.82
0.7	0 1	0.78 0.92	0.97 0.56	0.87 0.70	326 199	0.82
0.8	0 1	0.75 0.97	0.99 0.46	0.85 0.63	326 199	0.79
0.9	0 1	0.71 1. 00	1.00 0.33	0.83 0.50	326 199	0.75

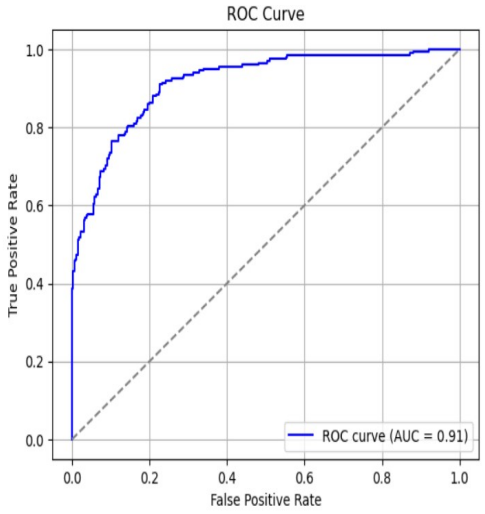
The classification matrix across different cut-off values reveal how threshold selection impacts the model's performance. At lower cut-off values (0.1 to 0.3), the model demonstrates very high recall, indicating it correctly identifies most of the actual positive cases. For instance, at a cut-off of 0.1, recall reaches 0.96, but precision remains lower at 0.50,reflecting many false positives. As the threshold increases to 0.3, precision improves to 0.71 while maintaining a strong recall of 0.88, yielding a balanced F1-score of 0.84 and accuracy of 0.82. The optimal trade-off appears around a cut-off of 0.4 to 0.5, where precision and recall are both high, and the F1-score and accuracy peak around 0.87 and 0.84 respectively. Beyond this range, particularly at thresholds of 0.7 and above, precision remains relatively high but recall drops significantly, resulting in a decline in both F1-score and overall accuracy. For example, at a 0.9 cut-off, recall is perfect (1.00), but precision drops to 0.33, and accuracy falls to 0.75. These trends indicate that while higher thresholds reduce false positives, they increase false negatives. Therefore, a threshold of 0.4 to 0.5 offers the best balance between precision and recall, making it ideal for maintaining overall model performance.

Fig. 1 : Threshold and accuracy values:



1

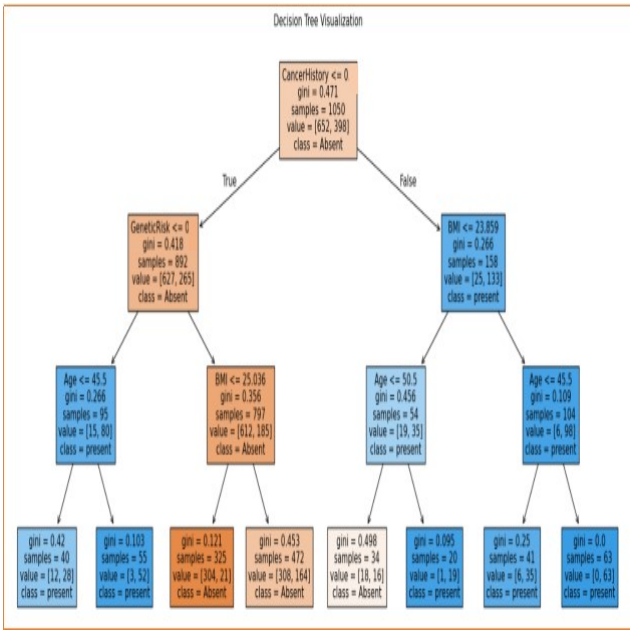
Fig. 2 : ROC curve:



AUC Score: 0.91

The binary logistic regression model is statistically significant and moderately well-fitted (Pseudo $R^2 = 0.47$). All predictors are highly significant ($p < 0.001$), contributing meaningfully to the outcome. Age, BMI, smoking, alcohol intake, and genetic risk significantly increase disease odds. Physical activity reduces disease risk, acting as a protective factor. Cancer history is the strongest predictor, increasing odds by over 70 times. Model performance varies with the decision threshold; accuracy peaks around 0.4–0.5. Overall, a threshold around 0.4–0.5 is optimal, balancing both public health sensitivity and diagnostic precision.The AUC score of 0.91 indicates excellent model performance and strong ability to distinguish between the two classes.

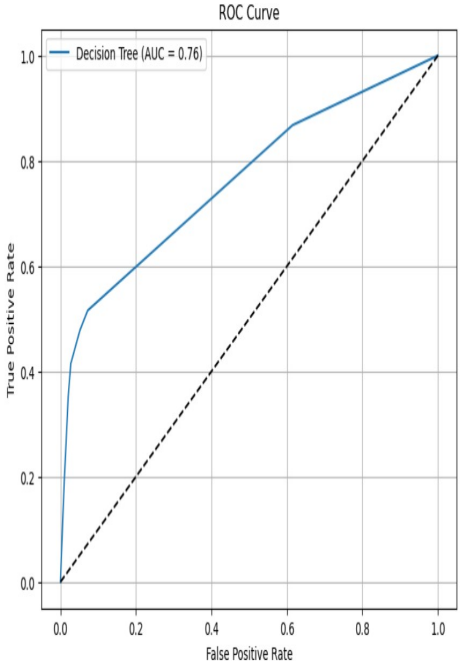
Fig.3 - Decision Tree Visualization To Classify Cancer Patients with Diagnosis



The decision tree visualization provides insight into the hierarchical structure used by the model to classify instances as either "present" or "absent" for a certain condition. The root node splits on **Cancer History**, where individuals with no cancer history (CancerHistory ≤ 0) proceed left, while those with a history proceed right. Among those with no cancer history, the next key feature is **Genetic Risk**. Individuals with low or no genetic risk (GeneticRisk ≤ 0) are further divided based on **Age**, indicating that older individuals (Age > 45.5) are more likely to have the condition present. On the other hand, those with higher genetic risk are split based on **BMI**, where a BMI ≤ 25.036 is more strongly associated with the "Absent" class.

For individuals with a history of cancer (right branch), **BMI** again plays a significant role. Those with a BMI ≤ 23.859 are more likely to have the condition "present". Subsequent splits on **Age** refine the classification further. Notably, terminal nodes (leaves) with very low Gini impurity, such as one with BMI > 23.859 and Age > 45.5 (value = [0, 63]), suggest pure classification, indicating high confidence in predictions. Overall, cancer history, genetic risk, BMI, and age are key discriminators in this tree.

Fig. 3 : ROC Curve using decision tree



Conclusion :

This study demonstrates the effectiveness of machine learning models, particularly Logistic Regression and Decision Tree classifiers, in predicting cancer diagnosis using demographic, lifestyle, and genetic factors from the Wisconsin dataset. The logistic regression model revealed that age, gender, BMI, smoking, alcohol intake, genetic predisposition, and cancer history significantly increased cancer risk, while physical activity acted as a protective factor. The model achieved high predictive accuracy, with an optimal decision threshold of 0.4–0.5 and an AUC score of 0.91, indicating excellent discriminatory power. Similarly, the decision tree model provided interpretable rules that highlighted cancer history, genetic risk, BMI, and age as primary determinants, enabling straightforward clinical interpretation. Together, these findings underscore the value of combining statistical modeling with interpretable algorithms to enhance cancer risk assessment. Importantly, the study highlights that both lifestyle choices and genetic susceptibility are critical in influencing cancer outcomes, emphasizing the need for preventive interventions alongside diagnostic tools. While promising, future research should focus on integrating larger and more diverse real-world datasets to improve model generalizability and robustness. Overall, predictive modeling holds significant potential in supporting early detection, clinical decision-making, and personalized healthcare strategies for cancer management.

References :

[1] J.P. Nayak, B.D. Parameshachari, K.S. Soyjaudah, R. Banu, A.C. than, denfication of PCB faults using image processing, in 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), IE, 2017, pp. 1-4.

[2] H. Liang, M. Loo, R. Wang, P. La, W. La, I. Long, Big data in health care applications and challenges, Data Inf. Manage. 2 (3) (2018) 175-197

[3] J Sivapriya, V Aravind Kumar, S Siddarth Sai, S Sriram, Breast cancer predictio using machine learning, Int. J. Recent Technol. Eng. 8 (4) (2019).

[4] P. Subramani, P. BD, Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post COVID-19 patients, Pers. Ubiquitour Comput. (2021) 1-14.

[5] A. Peretti, F. Aments, Breast cancer prediction by logistic regression with CUDA parallel programming support, Breast Cancer Curr. Res. 1 (111) (2016) 2

[6] N.T. Le, J.W. Wang, D.H. Le, C.C. Wang, T.N. Nguyen, Fingerprint enhancement based on tensor of wavelet subbands for classification, IEEE Access 8 (2020) 6602-6615.

[7] Reddy Prasad, Pidaparthi Anjali, S. Adil, N. Deepa, Heart disease prediction using logistic regression algorithm using machine learning, Int. J. Eng. Adv. Technol. 8(35) (2019).

[8] Ch. Shravya, K. Pravalika, Shaik Subhani, Prediction of breast cancer using supervised machine learning techniques, Int. J. Innov. Technol. Exploring Eng. 8 (6) (2019).

[9] K. polaraju, D.Durga Prasad, Prediction of heart disease using multiple linear regression model, Int. J. Eng. Dev. Res. 5 (4) (2017) 1419-1425.

[10] M.C. Mariani, F. Biney, O.K. Tweneboah, Analyzing medical data by using statistical learning models, Mathematics 9 (9) (2021) 968.

[11] Constantin M, Mătanie C, Petrescu L, Bolocan A, Andronic O, Bleotu C, Mitache MM, Tudorache S, Vrancianu CO. Landscape of Genetic Mutations in Appendiceal Cancers. Cancers (Basel). 2023 Jul 12;15(14):3591. doi: 10.3390/cancers15143591. PMID: 37509254; PMCID: PMC10377024.

[12] Kwon JH, Kim H, Lee JK, Hong YJ, Kang HJ, Jang YJ. Incidence and Characteristics of Multiple Primary Cancers: A 20-Year Retrospective Study of a Single Cancer Center in Korea. Cancers (Basel). 2024 Jun 26;16(13):2346. doi: 10.3390/cancers16132346. PMID: 39001408; PMCID: PMC11240339

[13] Malhotra RK, Manoharan N, Nair O, Deo S, Rath GK. Trends in Lung Cancer Incidence in Delhi, India 1988-2012: Age-Period-Cohort and Joinpoint Analyses. Asian Pac J Cancer Prev. 2018 Jun 25;19(6):1647-1654. doi: 10.22034/APJCP.2018.19.6.1647. PMID: 29937537; PMCID: PMC6103588.

[14] Wang B, Li X, Liu L, Wang M. β-Catenin: oncogenic role and therapeutic target in cervical cancer. Biol Res. 2020 Aug 5;53(1):33. doi: 10.1186/s40659-020-00301-7. PMID: 32758292; PMCID: PMC7405349.