ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

Analysis of the SEER Database: An ML Based Decision Support System (DSS) to Treat the Cancer Patients Under Head and Neck Category

[1] Hrishikesh Kumbhar, [2] Urvi Pandit, [3] Manav Tanna, [4] Bhargav Bodhankar

[1] [2] [3] [4] Department of Information Technology (IT), Vivekanand Education Society's Institute of Technology, Mumbai, India

Corresponding Author Email: [1] 2020.hrishikesh.kumbhar@ves.ac.in, [2] 2020.urvi.pandit@ves.ac.in, [3] 2020.manav.tanna@ves.ac.in, [4] 2020.bhargav.bodhankar@ves.ac.in

***Abstract*— Oral cancer is a formidable adversary, impacting millions of lives globally. The work done so far has employed AI and ML techniques for prognosis, drug discovery and treatment of cancer, but needs more attention on this category of cancer. This research aims at discovering various implications of different attributes such as histology type, treatment and months from diagnosis, on survival.**

***Keywords*—Oral Cancer, Head and Neck Cancer, Machine Learning, Healthcare.**

I. INTRODUCTION

Oral cancer stands as a daunting opponent in the realm of public health, posing a significant challenge globally. Its intricate interplay of genetic predispositions, lifestyle factors, and environmental influences renders it a complex puzzle for researchers and clinicians alike. Despite advancements in medical science, understanding the nuanced dynamics of oral cancer incidence and progression remains elusive.

Much research has been done in an attempt to illuminate effects of different treatment types like radiation, chemotherapy and even the drugs. The researchers have extensively utilized the power of artificial intelligence and machine learning to fasten the process of prognosis and analyzing most effective treatment and drugs combination for ensuring higher and better survival probabilities. The already done research speaks vividly about lung, liver, breast, blood and thyroid gland cancer. Hence creating an immense need of improved understanding in the head and neck category of cancer, which is so widely spread one and prevalent.

This research delves into the web of various factors influencing the survival rates of oral cancer. It seeks to unravel hidden dependencies between site of cancer, gender, histology type to name a few, and chances of survival. The research focuses majorly on extending the survival months and also takes into consideration the recurrence probability. It purely is a decision support system that depends on a dataset spanning more than 25 years, hence the results and inferences are seasoned through many experiences.

II. LITERATURE REVIEW

Machine learning applications in cancer prognosis and prediction [3], authored by Konstantina Kourou et al., explores the application of machine learning (ML) techniques in cancer prediction and prognosis. The study emphasizes the integration of mixed data and underscores the importance of external validation for predictive models. The potential of ML methods to enhance accuracy in cancer prediction is highlighted, with a specific focus on the necessity for larger datasets and rigorous validation, particularly in the context of gene expression profiling in clinical practice. Findings indicate that ML techniques, when coupled with feature selection and classification algorithms, hold promise for inference in the cancer domain. However, challenges such as the need for better statistical analysis of datasets, larger and more valid cancer databases, and thorough validation of ML models before clinical implementation are identified. The paper suggests future research directions to overcome limitations, improve prediction accuracy, and explore the potential of ML techniques in personalized medicine.

Artificial intelligence in cancer target identification and drug discovery[1], authored by Yujie You et al., discusses the application of artificial intelligence (AI) in cancer target identification and drug discovery. The study explores AI algorithms, including network-based and machine learning-based approaches, to analyze complex biological networks and identify potential anticancer targets. Emphasizing the importance of integrating multiomics data, the paper addresses challenges in the field. AI, particularly network-based and machine learning-based algorithms, demonstrates promise in identifying novel anticancer targets and discovering drugs. However, challenges such as integrating heterogeneous data, ensuring interpretability of AI models, and addressing data bias are identified. The study suggests future directions involving the development of efficient feature selection applications, improvement in druggability prediction, advancement of drug property prediction models, and utilization of AI for clinical trial

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

design.

Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis[2], authored by Moustafa Mourad et al., focuses on the application of machine learning algorithms, specifically artificial neural networks (ANNs), to predict the prognosis of thyroid cancer patients. The study utilizes a database obtained from the U.S. SEER-18 database, concentrating on papillary carcinoma and follicular carcinoma subtypes. The research demonstrates the ability of ANNs to accurately predict patient outcomes and distinguish between different prognosis categories. The study also discusses the potential of incorporating additional data sources, such as genomics or proteomics studies, to further enhance the algorithms. Findings highlight the relevance of extrathyroidal spread and the impact of various clinical variables on prognosis. The results show high accuracy in predicting outcomes, emphasizing the limitations of the TNM staging system and suggesting the inclusion of medically important features in future patient databases.

III. IMPLEMENTATION

A. SEER Dataset

The Surveillance, Epidemiology, and End Results (SEER) program is a comprehensive and authoritative source of cancer statistics in the United States. Developed and maintained by the National Cancer Institute (NCI), SEER collects and disseminates data on cancer incidence, prevalence, survival, and mortality. The database encompasses a wide range of demographic, clinical, and tumor-specific information, making it a valuable resource for researchers, healthcare professionals, and policymakers. The SEER registries collect data on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment, and they follow up with patients for vital status. The SEER dataset selected was Incidence-SEER Research Data, 18 Registries, Nov 2020 Sub (2000-2018) and is thoroughly explored in order to find relevant attributes. Total 1,13,893 rows and 80 columns were selected.

B. Data Cleaning

Through the removal of columns with substantial missing values, a refined dataset was obtained. Subsequently, to improve data completeness, the most frequently occurring values (modes) for each column were identified and utilized to impute any remaining missing values. This comprehensive cleaning process resulted in a streamlined dataset, consisting of 46 attributes, after the removal of redundant attributes and non-useful columns from the initial set of 80 attributes.

C. Important Attribute Selection

The Random Forest algorithm, renowned for its efficacy in ensemble learning, emerges as a versatile tool applicable to both classification and regression tasks. Features that exert a more substantial influence on the model's predictive accuracy and effectiveness are accorded higher importance. By discerning and prioritizing these influential features, Random Forest aids in streamlining the dataset, emphasizing those elements that contribute most significantly to the model's overall efficacy. The basic idea behind using Random Forest for feature selection is to measure the importance of each feature in the dataset.

D. Stratified Sampling

Given the extensive number of columns in the dataset and approximately 1,16,000 rows, conducting analyses on the entire dataset posed a challenge and resulted in biased outcomes. To mitigate this, a strategic approach involving stratified sampling was employed. Stratified sampling is instrumental in ensuring a balanced representation of various classes or groups within the dataset. In this context, a subset of the data comprising 20,000 rows was created using the stratified sampling method. This involved the random selection of rows from the entire dataset, contributing to a more representative and unbiased sample for subsequent analyses. The adoption of stratified sampling served as a pivotal step to address the challenges posed by the extensive dataset, facilitating more robust and reliable insights.

E. Model Training and Testing

In working with the stratified dataset, our approach involved allocating 80\% of the data for model training, reserving the remaining 20\% for testing purposes. To ensure a comprehensive evaluation across various scenarios, we explored a distinctive technique by selecting contiguous rows for training and testing. This involved dividing the entire dataset into five equal parts. In a series of five iterations, each part was designated for testing, while the rest were utilized for training. This systematic process allowed us to assess model accuracy across diverse combinations, providing a thorough evaluation of performance under varied conditions.

IV. RESULTS

A. The correlation matrix

II. From the above correlation matrix, we found that some attributes were showing the degree at which they are correlated with each other. There were no significant attributes found between 0.8 to 1 correlation. CS site specific factor 3,4,5,6 and CS lymph nodes (2004-15), Survival months and year of follow up recode, CS mets at dx and CS mets at eval lie between correlation value 0.4 to 0.6. Rest of the attributes lie between correlation values 0 to 0.2.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

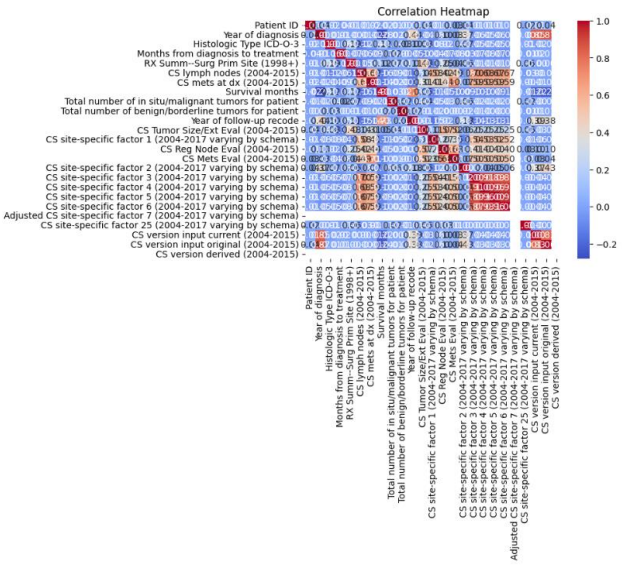

Fig. 1. Correlation Matrix

B. Important attributes for different outcomes

1. Survival Months

Year of diagnosis, CS Schema - AJCC 6th Edition, Median household income inflation adj to 2021, Derived AJCC T, 6th ed (2004-2015), Rural-Urban Continuum Code, RX Summ--Surg Prim Site (1998+), Months from diagnosis to treatment, CS site-specific factor 1 (2004-2017 varying by schema), CS version input original (2004-2015), Site recode - rare tumors, Histologic Type ICD-O-3, ICD-O-3 Hist/behav, malignant, Year of follow-up recode, COD to site recode, COD to site rec KM..

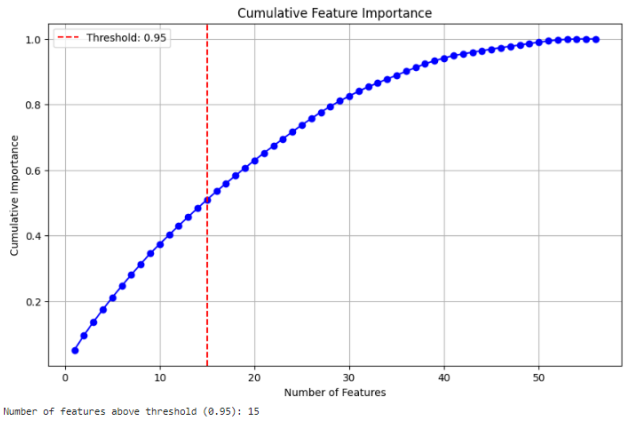

Fig. 2. Feature Importance for Survival Months

2. Vital status recode (study cutoff used)

SEER cause-specific death classification, SEER other cause of death classification, Year of diagnosis, Survival months, Rural-Urban Continuum Code, Median household income inflation adj to 2021, RX Summ--Surg Prim Site (1998+), Combined Summary Stage (2004+), Survival months flag, SEER historic stage A (1973-2015), ICD-O-3 Hist/behav, malignant, Histologic Type ICD-O-3, SEER Combined Summary Stage 2000 (2004-2017), RX Summ--Systemic/Sur Seq (2007+), CS site-specific factor 1 (2004-2017 varying by schema), Reason no cancer-directed surgery, Sequence number, Total number of in situ/malignant tumors for patient, CS site-specific factor 3 (2004-2017 varying by schema), Site recode - rare tumors, CS Schema - AJCC 6th Edition, Histology recode - broad groupings, CS site-specific factor 2 (2004-2017 varying by schema), RX Summ--Surg/Rad Seq, Months from diagnosis to treatment.

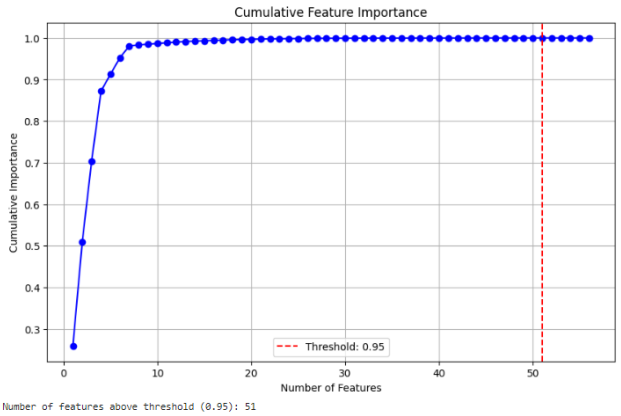

Fig. 3. Feature Importance of Vital status recode

Similarly, important attributes were found for features like histology type and therapy type are as follows:

3. Histology Type

ICD-O-3 Hist/behav, malignant, Histology recode - broad groupings, Site recode - rare tumors, CS Schema - AJCC 6th Edition, Survival months, Year of diagnosis, RX Summ--Surg Prim Site (1998+), Median household income inflation adj to 2021, Rural-Urban Continuum Code, Reason no cancer-directed surgery, Months from diagnosis to treatment, Laterality, SEER historic stage A (1973-2015), Vital status recode (study cutoff used), CS site-specific factor 1 (2004-2017 varying by schema), Race recode (White, Black, Other), RX Summ--Surg/Rad Seq, Total number of in situ/malignant tumors for patient, Sequence number, Sex, Survival months flag, Combined Summary Stage (2004+).

4. Therapy Type

Survival months, Year of diagnosis, CS Schema - AJCC 6th Edition, Months from diagnosis to treatment, Median household income inflation adj to 2021, Rural-Urban Continuum Code, Reason no cancer-directed surgery, SEER historic stage A (1973-2015), RX Summ--Surg/Rad Seq, Site recode - rare tumors, ICD-O-3 Hist/behav, malignant, Histologic Type ICD-O-3.

C. ROC curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a classification model across different threshold settings. It is widely used in binary and multiclass classification problems

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

to assess the trade-off between sensitivity (true positive rate) and specificity (true negative rate).

Good Discriminatory Power: All three classes (0, 1, and 2) have ROC AUC values greater than 0.5, which is the baseline for random guessing. Class 2 Performs Slightly Better: Class 2 has the highest ROC AUC value (0.94), suggesting that the model performs particularly well in distinguishing Class 2 from the rest. Class 0 and Class 1 Perform Well: Both Class 0 and Class 1 have ROC AUC values above 0.9, indicating that the model is also effective at distinguishing these classes from the others. However, Class 0 has a slightly higher ROC AUC value than Class 1.

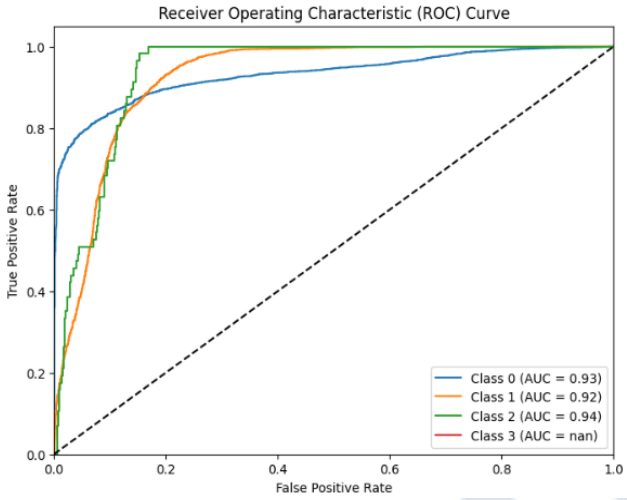

Fig. 4. ROC Curve

D. Performance Evaluation of Classifiers

We used Decision Trees, Random Forest and Logistic Regression for training the model and got the highest accuracy for Random Forest at 95.29%

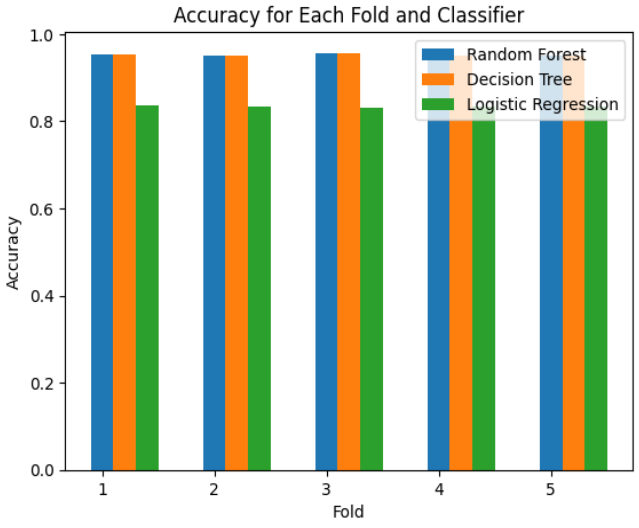

Fig. 5. Accuracy comparison for different algorithms and folds

Table I. Accuracies of Different Classifiers

No.	Classifier	Accuracy
1.	Random Forest	0.9529
2.	Decision Tree	0.9528
3.	Logistic Regression	0.8342

V. EDA

The below figure shows distribution of data between Male and Female based on Survival Status.

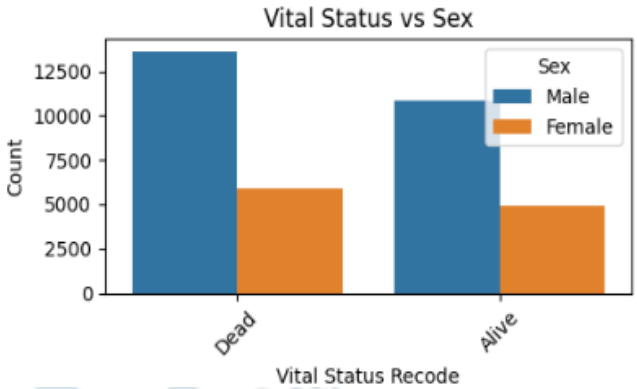

Fig. 6. Vital Status vs Sex

The below figure shows distribution of patients with different laterality and its impact on survival status.

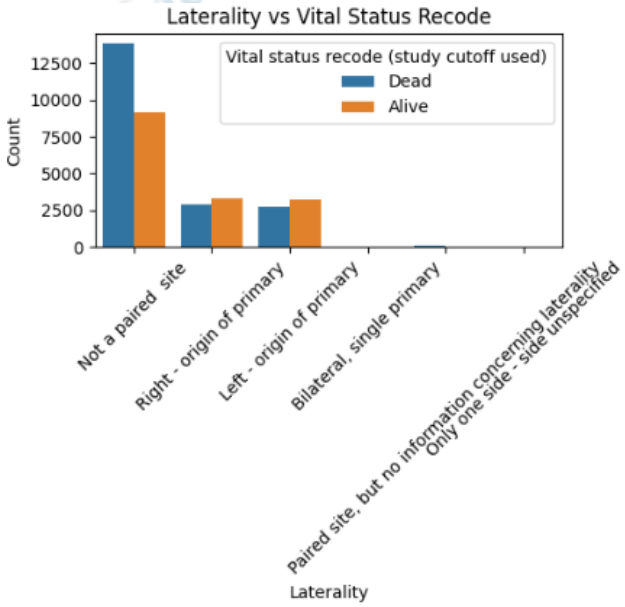

Fig. 7. Laterality vs Vital Status Recode

VI. CONCLUSION

Through rigorous analysis of SEER data, we have gained a comprehensive understanding of the dominant attributes associated with oral cancer. Exploring specific indicators within dominant attributes enhances our understanding and

aids in early detection methods, revealing hidden insights within the dataset. After analyzing the data, it's evident that the Random Forest and Decision Tree classifiers exhibit remarkably high accuracies, with Random Forest achieving 95.29% and Decision Tree reaching 95.28%. On the other hand, the Logistic Regression classifier shows a noticeably lower accuracy of 83.42%. These results suggest that tree-based classifiers, particularly Random Forest and Decision Tree, are well-suited for this dataset compared to Logistic Regression.

VII. ACKNOWLEDGMENT

We extend our sincere gratitude to two esteemed experts whose guidance significantly contributed to this research

1. Dr. Arjun Singh, MDS (Oral and Maxillofacial Surgery), MFDS RCPS (Glasgow), Assistant Professor, Surgeon-Scientist "E", Department of Head and Neck Oncology, Advanced Centre for Treatment, Research and Education of Cancer (ACTREC), Navi Mumbai, Tata Memorial Centre, Mumbai.
2. Prof. Pankaj Chaturvedi, Surgeon, Department of Head Neck Surgery, Deputy Director, Centre for Cancer Epidemiology, Tata Memorial Centre,Mumbai, Department of Atomic Energy, Government of India, Director, International Federation of Head and Neck Oncologic Societies

REFERENCES

[1] You, Yujie, Xin Lai, Yi Pan, Huiru Zheng, Julio Vera, Suran Liu, Senyi Deng, and Le Zhang. “Artificial intelligence in cancer target identification and drug discovery.” . Signal Transduction and Targeted Therapy 7, no. 1 (2022): 156.

[2] Mourad, Moustafa, Sami Moubayed, Aaron Dezube, Youssef Mourad, Kyle Park, Albertina Torreblanca-Zanca, José S. Torrecilla, John C. Cancilla, and Jiwu Wang. “Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis.”. Scientific reports 10, no. 1 (2020): 5176.

[3] Themis, P., P. Konstantinos, V. Michalis, and I. Dimitrios. “Machine learning applications in cancer prognosis and prediction.” . Computational and Structural Biotechnology Journal (2015).

[4] Fatapour, Yasaman, Arash Abiri, Edward C. Kuan, and James P. Brody. 2023. "Development of a Machine Learning Model to Predict Recurrence of Oral Tongue Squamous Cell Carcinoma" Cancers 15, no. 10: 2769. https://doi.org/10.3390/cancers15102769

[5] Warren JL, Mariotto A, Melbert D, Schrag D, Doria-Rose P, Penson D, Yabroff KR. Sensitivity of Medicare Claims to Identify Cancer Recurrence in Elderly Colorectal and Breast Cancer Patients. Med Care.2016 Aug;54(8):e47-54. doi: 10.1097/MLR.0000000000000058. PMID: 24374419; PMCID: PMC4072852.

[6] Umesh D R, & Ramachandra, B. (2015). Association rule mining based predicting breast cancer recurrence on SEER breast cancer data. 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). doi:10.1109/erect.2015.7499044

[7] Yang, J., Guo, K., Zhang, A. et al. Survival analysis of age-related oral squamous cell carcinoma: a population study based on SEER. Eur J Med Res 28, 413 (2023). https://doi.org/10.1186/s40001-023-01345-7

[8] A. Farrag, Z. M. Fadlullah, M. M. Fouda and N. S. Almalki, "Survival-Based Treatment Planning Using Stage-Specific Machine Learning Models," in IEEE Access, vol. 11, pp. 134404-134420, 2023, doi: 10.1109/ACCESS.2023.3337117.

[9] Baxi SS, Pinheiro LC, Patil SM, Pfister DG, Oeffinger KC, Elkin EB. Causes of death in long-term survivors of head and neck cancer. Cancer. 2014 May 15;120(10):1507-13. doi: 10.1002/cncr.28588. Epub 2014 Feb 22. PMID: 24863390; PMCID: PMC4101810.

[10] Ma Z, Yang S, Yang Y, Luo J, Zhou Y, Yang H. Development and validation of prediction models for the prognosis of colon cancer with lung metastases: a population-based cohort study. Front Endocrinol (Lausanne). 2023 Jul 31;14:1073360. doi: 10.3389/fendo.2023.1073360. PMID: 37583430; PMCID: PMC10424923.