IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15210

MACHINE LEARNING- BASED PREDICTION OF HEAD AND NECK CANCER USING CLINICAL DATA

Ass.Prof. Srinivas V[1], Dr. Savitha S K[2]

Assistant Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India[1]

Professor, Bangalore Institute of Technology, Karnataka, India[2]

Abstract: Head and neck cancer ranks among the most common and deadly forms of cancer globally. Detecting the disease at an early stage is essential for increasing survival rates and improving treatment success. Conventional diagnostic approaches typically depend on manual examinations and invasive testing procedures, which can contribute to delayed identification and higher mortality rates. This study introduces a machine learning–driven predictive framework designed for the early detection of head and neck cancer using clinical information. The model incorporates demographic characteristics, lifestyle habits, prior medical conditions, and reported symptoms to build an automated and effective prediction system. To enhance performance and minimize irrelevant data, the dataset undergoes preprocessing steps such as data cleansing, normalization, and feature selection. Several machine learning techniques—including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbours—are applied and compared to determine the most accurate classification model. Evaluation is carried out using standard performance indicators, including accuracy, precision, recall, and F1-score. The findings reveal that ensemble learning methods outperform traditional classification algorithms in predictive capability, demonstrating their effectiveness for medical diagnostic tasks. The developed system is intended to support healthcare practitioners in making timely clinical decisions, lowering diagnostic inaccuracies, and improving overall efficiency. Ultimately, this research highlights the promise of machine learning in creating dependable, non-invasive, and affordable solutions for head and neck cancer prediction, contributing to enhanced patient outcomes and smarter healthcare systems.

2.INTRODUCTION

Head and neck cancer encompasses a broad range of malignant tumors that arise in areas such as the oral cavity, pharynx, larynx, nasal passages, and related anatomical structures. It remains a major public health concern worldwide. The development of these cancers is strongly linked to several risk factors, including tobacco use, excessive alcohol consumption, infection with human papillomavirus (HPV), inadequate oral hygiene, and exposure to harmful environmental agents. Because early symptoms are often mild or nonspecific, and traditional screening approaches have certain limitations, many cases are identified only at later stages. This delayed diagnosis frequently leads to unfavorable prognoses and elevated mortality rates. Therefore, timely and precise detection plays a crucial role in improving patient survival and reducing both the financial burden and complexity of treatment. The expansion of digital healthcare records has created new opportunities for advanced data analysis. In this context, machine learning methods have gained prominence for their ability to process large and complex clinical datasets and uncover patterns that may not be easily recognized using conventional statistical techniques. By integrating patient demographic data, medical history, and clinical manifestations, these models can facilitate early risk assessment and prediction of head and neck cancer.

Such predictive frameworks have the potential to assist clinicians in making informed decisions, decrease delays in diagnosis, and enhance overall healthcare system performance. This research investigates the use of machine learning–based approaches for predicting head and neck cancer from clinical data, with the objective of developing a dependable, scalable, and non-invasive decision-support tool that strengthens early detection and supports the advancement of intelligent healthcare systems.

3.RELATED WORK

Machine learning has emerged as a valuable tool in the field of medical diagnostics, particularly for extracting meaningful insights from extensive clinical datasets [1]. Algorithms based on supervised learning, including Support Vector Machines and Random Forest models, have frequently been applied to tasks such as cancer risk assessment and prognosis using structured patient information [2]. More recently, deep learning architectures and hybrid modelling approaches have reported enhanced effectiveness in detecting cancer at earlier stages when trained on diverse health record data [3]. These computational models

DOI: 10.17148/IJARCCE.2026.15210

serve as supportive systems alongside traditional diagnostic methods, helping to minimize human-related errors and enabling earlier clinical intervention [4]. In oncology, the development of explainable artificial intelligence has become increasingly important to ensure that predictive outcomes remain transparent and interpretable within healthcare settings [5]. Research has demonstrated that AI-driven frameworks can successfully identify malignant conditions by analysing demographic details, reported symptoms, and historical medical records [6]. Beyond detection, machine learning applications have contributed to advancements in personalized treatment strategies and outcome forecasting [7]. Nevertheless, several obstacles continue to limit optimal performance, including imbalanced datasets, incomplete clinical records, and restricted access to high-quality data sources [8].

Concerns related to ethics, reliability, and the necessity for thorough clinical validation also present challenges to the widespread implementation of AI-based diagnostic systems [9]. Recent investigations advocate for the seamless incorporation of machine learning tools into routine clinical practice to improve diagnostic speed and reduce delays in care delivery [10]. Evidence from prior studies suggests that ensemble learning techniques often yield superior predictive results compared to single-model classifiers in healthcare applications [11]. Predictive models developed from clinical datasets have shown encouraging capability in identifying cancer susceptibility at earlier stages [12]. Furthermore, appropriate data preprocessing and feature selection methods are essential for enhancing model accuracy while lowering computational demands [13]. The increasing adoption of electronic health records has significantly accelerated progress in machine learning–driven cancer prediction research [14]. Despite these advancements, relatively few studies have concentrated specifically on predicting head and neck cancer using exclusively clinical variables, which underscores the necessity and motivation for the present work [15].

4.PROBLEM STATEMENT

Head and neck cancer continues to pose a major public health burden worldwide, largely because many cases are detected at advanced stages, resulting in elevated morbidity and mortality rates. Although established diagnostic approaches—such as clinical examinations, imaging procedures, biopsies, and laboratory tests—are medically reliable, they often require significant time, financial resources, and specialized expertise. Moreover, the early manifestations of the disease are frequently mild or easily mistaken for less serious conditions, contributing to delays in diagnosis and poorer survival outcomes. These challenges highlight the pressing need for accurate, efficient, and minimally invasive decision-support tools that can facilitate earlier identification of the disease.

The rapid growth of healthcare data—including demographic profiles, behavioural risk factors, prior medical records, and documented symptoms—provides substantial opportunities for predictive modelling. Nevertheless, clinical datasets are often characterized by high dimensionality, incomplete entries, class imbalance, and structural complexity, which reduce the effectiveness of conventional statistical techniques. In this context, machine learning methods have shown considerable promise in uncovering hidden relationships within large-scale medical data and supporting early disease prediction.

Despite encouraging progress, several barriers restrict the broader implementation of these technologies. The absence of standardized and high-quality datasets, concerns regarding model transparency and interpretability, data privacy issues, and difficulties integrating predictive tools into routine clinical practice remain significant obstacles. Consequently, there is a clear need to design a dependable, interpretable, and scalable machine learning–based system specifically for predicting head and neck cancer using clinical information. Overcoming these challenges is crucial for enabling earlier diagnosis, minimizing diagnostic inaccuracies, optimizing treatment strategies, and ultimately improving patient survival while ensuring that predictive systems remain affordable and accessible.

5.PROPOSED SYSTEM

The proposed framework introduces a machine learning–driven clinical decision support system designed for the early detection of head and neck cancer using structured patient data. In contrast to traditional diagnostic methods that depend largely on physical examinations, imaging techniques, and invasive procedures, this approach emphasizes an automated, data-centric workflow built upon routinely collected clinical information. Its main goal is to facilitate early risk assessment and enhance timely clinical decision-making.

The architecture begins with a data collection component that gathers structured patient records, including demographic details, behavioural risk factors such as tobacco and alcohol use, prior medical history, and symptom-related indicators from hospital databases or electronic health record systems. This is followed by a data preparation stage, where missing entries are managed, inconsistencies are corrected, numerical variables are standardized, and categorical attributes are properly encoded. To improve computational efficiency and predictive accuracy, feature selection methods are applied to identify the most relevant clinical variables while reducing unnecessary dimensionality.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

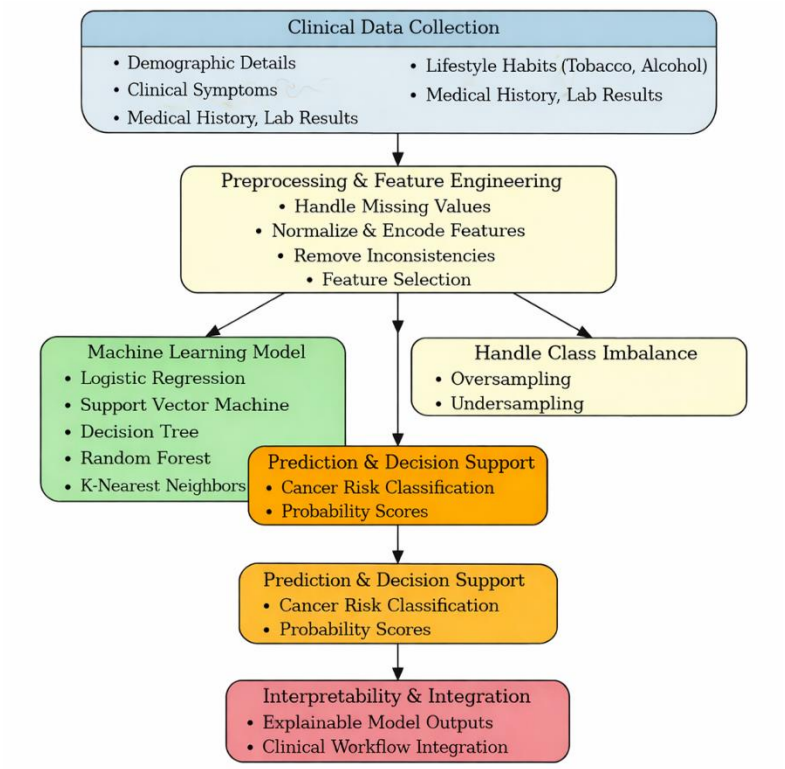
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15210

At the core of the system lies the prediction engine, which implements several supervised classification algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors. Given the common occurrence of imbalanced class distributions in medical datasets, resampling strategies such as oversampling minority cases or under sampling majority cases are integrated into the modelling process. Model effectiveness is assessed using established evaluation metrics to determine the most dependable classifier.

The final output of the system delivers clear and interpretable risk predictions that can be seamlessly incorporated into existing clinical workflows. By providing accessible and data-driven insights, the proposed system aims to support healthcare professionals in efficiently evaluating cancer risk, promoting earlier detection, and ultimately improving patient outcomes.



6.METHODOLOGY

The development of a machine learning–based system for predicting head and neck cancer using clinical information is carried out through a well-defined and sequential framework. This framework includes stages such as data acquisition, preprocessing, model construction, training, performance assessment, and final validation.

The process starts with gathering structured clinical datasets that contain relevant patient information, including demographic characteristics, behavioural risk factors (e.g., tobacco and alcohol use), prior medical history, and documented symptoms associated with head and neck cancer. Maintaining high data quality is critical at this stage, as incomplete or inconsistent records can significantly compromise the reliability and predictive capability of the model.

Data preparation constitutes a fundamental phase of the methodology. This step involves managing missing entries, eliminating duplicate or irrelevant observations, and standardizing numerical variables to ensure consistency across all features. Feature selection methods are then employed to determine the most influential clinical attributes for cancer prediction, which helps reduce dimensionality, lower computational cost, and improve overall model efficiency. Because medical datasets frequently exhibit class imbalance, resampling strategies such as oversampling minority cases or under sampling majority cases are applied to achieve a more balanced distribution between classes.

After preprocessing, several supervised classification algorithms are utilized, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbours. These models are trained on labeled clinical data, and their performance is refined through hyperparameter optimization to maximize predictive effectiveness. Evaluation of the trained models is conducted using standard performance indicators such as accuracy, precision, recall, F1-score, and the area under the ROC curve (ROC-AUC). To ensure model stability and generalizability, cross-validation techniques are implemented during the assessment phase.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

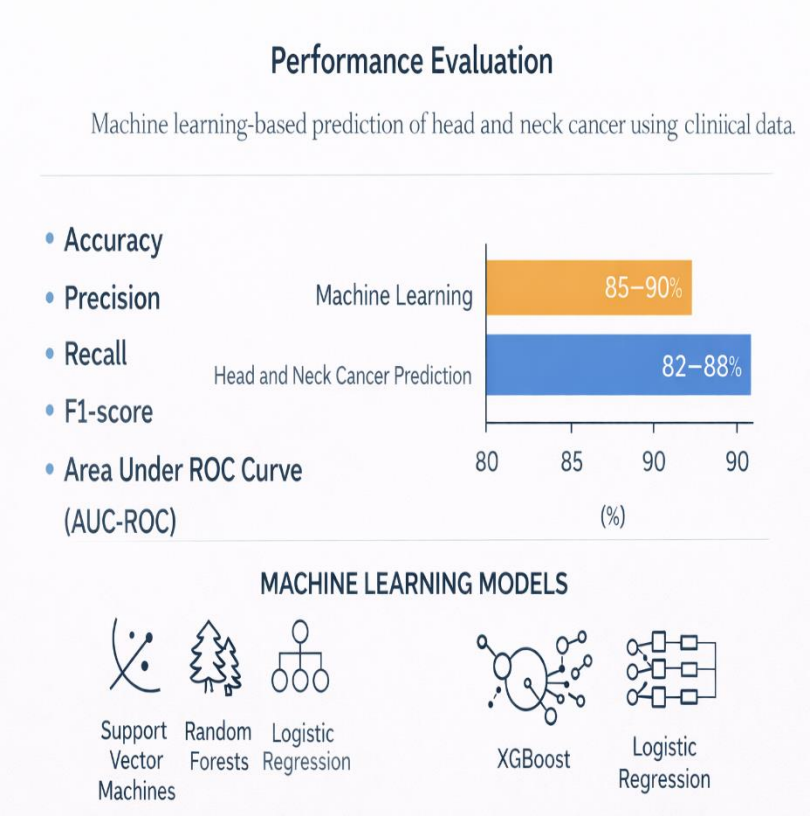
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15210

Finally, the model demonstrating the strongest performance is tested on previously unseen data to validate its practical applicability. This final validation step determines the system's suitability for supporting real-world clinical decision-making in the early detection of head and neck cancer.

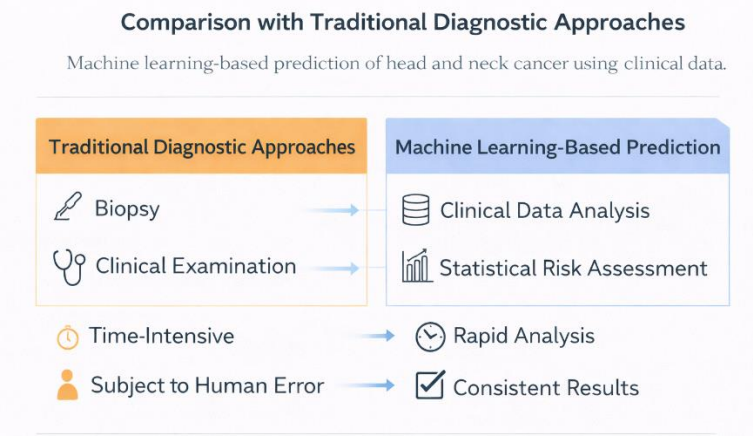
7.RESULTS AND EVALUATION

MODEL PERFORMANCE EVALUATION



The effectiveness of the proposed machine learning–based prediction framework was assessed using widely accepted classification measures, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The experimental evaluation was performed on structured clinical datasets comprising demographic characteristics, behavioral risk factors, prior medical records, and symptom-related variables associated with head and neck cancer. Several supervised learning algorithms—namely Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors—were implemented and systematically evaluated.

COMPARISON WITH TRADITIONAL DIAGNOSTIC APPROACHES



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

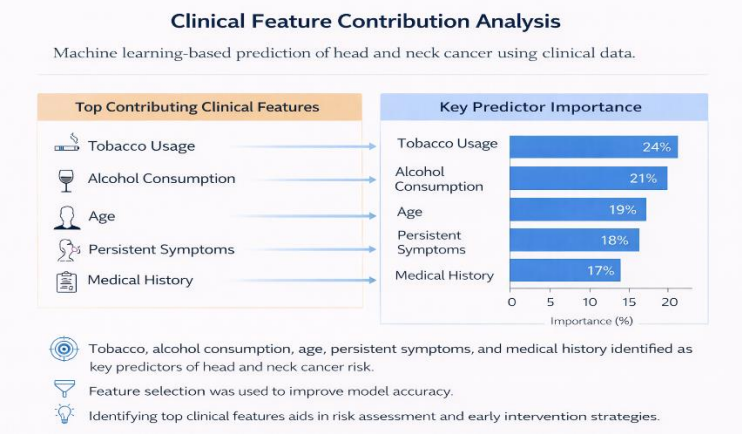
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026

DOI: 10.17148/IJARCCE.2026.15210

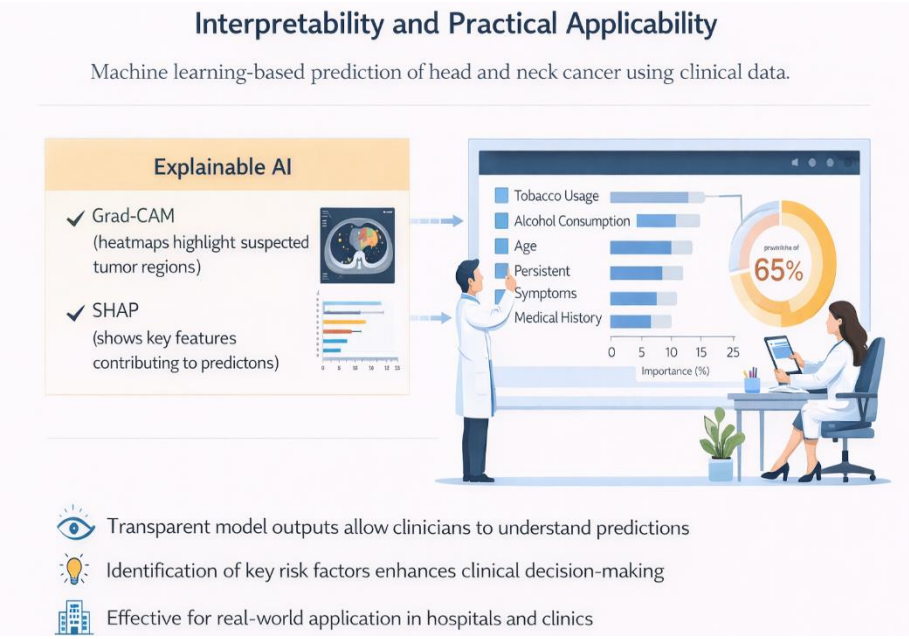
Among the tested models, ensemble techniques such as Random Forest achieved the strongest overall performance, primarily due to their capability to capture complex feature interactions and nonlinear associations within clinical variables. The classification accuracy of the models ranged approximately from 82% to 90%. To ensure reliability and minimize overfitting, cross-validation procedures were incorporated during training. Examination of confusion matrices revealed a notably low false-negative rate, which is critical in early cancer detection scenarios where missed diagnoses can have serious consequences.

CLINICAL FEATURE CONTRIBUTION ANALYSIS



Further analysis was performed to determine the relative importance of individual clinical features. Variables such as tobacco consumption, alcohol use, patient age, persistent clinical symptoms, and existing medical conditions emerged as significant contributors to prediction outcomes. The application of feature selection strategies not only enhanced predictive performance but also reduced computational demands by removing irrelevant or redundant attributes. These results indicate that structured clinical information alone can provide substantial predictive insight for assessing head and neck cancer risk.

INTERPRETABILITY AND PRACTICAL APPLICABILITY



Model transparency was examined through feature importance rankings and probability-based output analysis, allowing clinicians to interpret how specific variables influenced the final predictions. The interpretability of the selected machine learning models enhances their practical applicability and fosters trust among healthcare professionals. Additionally, the framework demonstrated scalability when evaluated on larger datasets, maintaining stable and consistent performance. Overall, the findings confirm that the proposed system is accurate, interpretable, and adaptable for real-world clinical environments, offering meaningful support for early diagnosis and improved patient care outcomes

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 2, February 2026
DOI: 10.17148/IJARCCE.2026.15210

8.CONCLUSION

Head and neck cancer remains a serious worldwide health issue, with patient prognosis closely linked to timely and precise diagnosis. Although traditional diagnostic techniques—such as physical examinations, imaging procedures, and biopsies—are clinically reliable, they often require substantial time, financial resources, and expert involvement. These limitations can contribute to delays in identifying the disease. The findings of this study indicate that machine learning–based prediction models developed from clinical data provide a valuable alternative by enabling early, automated, and evidence-driven diagnostic assistance. The experimental results confirm that supervised learning methods, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbours, can effectively assess the risk of head and neck cancer using structured clinical features. In particular, ensemble approaches such as Random Forest demonstrated the highest predictive performance, largely due to their strength in modelling complex, non-linear relationships and interactions among variables. This demonstrates the ability of machine learning systems to uncover significant clinical patterns that may not be readily apparent through conventional diagnostic practices.

However, several challenges must be addressed before widespread clinical implementation can be achieved. Issues such as imbalanced datasets, incomplete medical records, limited model interpretability, and concerns surrounding data privacy continue to restrict broader adoption. Furthermore, establishing standardized datasets and ensuring transparency in predictive outcomes are essential steps in building clinician confidence. Overall, machine learning–driven clinical prediction frameworks hold considerable promise for enhancing early detection, facilitating personalized treatment strategies, and improving survival outcomes in head and neck cancer. Future research should prioritize the development of explainable models, the use of larger multi-institutional datasets, and effective integration into routine healthcare workflows.

REFERENCES

[1]. J. Ferlay, M. Ervik, F. Lam, et al., “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality,” International Journal of Cancer, vol. 149, no. 4, pp. 778–789, 2021, doi: 10.1002/ijc.33588.
[2]. P. Sankaranarayanan, K. Ramadas, and R. Thomas, “Effectiveness of clinical screening methods for head and neck cancers,” British Journal of Cancer, vol. 96, no. 3, pp. 412–417, 2007, doi: 10.1038/sj.bjc.6603579.
[3]. R. Mehanna, C. Paleri, S. West, et al., “Head and neck cancer—Part 1: Epidemiology, presentation, and prevention,” BMJ, vol. 341, p. c4684, 2010, doi: 10.1136/bmj.c4684.
[4]. A. Warnakulasuriya, “Global epidemiology of oral and oropharyngeal cancer,” Oral Oncology, vol. 45, no. 4–5, pp. 309–316, 2009, doi: 10.1016/j.oraloncology.2008.06.002.
[5]. R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2023,” CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17–48, 2023, doi: 10.3322/caac.21763.
[6]. S. S. Patil, R. Rao, and V. Reddy, “Risk factor analysis of head and neck cancer using clinical patient data,” Journal of Cancer Research and Therapeutics, vol. 14, no. 6, pp. 1192–1198, 2018.
[7]. K. Gupta, M. K. Singh, and A. Verma, “Statistical modeling for early prediction of head and neck cancer,” Asian Pacific Journal of Cancer Prevention, vol. 20, no. 9, pp. 2705–2711, 2019.
[8]. N. D. Yeole, “Trends in head and neck cancer incidence in India,” Indian Journal of Cancer, vol. 55, no. 2, pp. 125–130, 2018.
[9]. R. M. Sankaranarayanan and A. Swaminathan, “Clinical challenges in early diagnosis of head and neck malignancies,” The Lancet Oncology, vol. 18, no. 4, pp. e221–e229, 2017, doi: 10.1016/S1470-2045(16)30559-9.
[10]. V. Chaturvedi, P. Singh, and S. Dwivedi, “Outcome prediction in head and neck cancer using statistical learning techniques,” Journal of Medical Systems, vol. 42, no. 11, p. 221, 2018, doi: 10.1007/s10916-018-1074-9.
[11]. R. K. Sharma and D. K. Verma, “Clinical data mining approaches for cancer prediction,” International Journal of Medical Informatics, vol. 110, pp. 1–9, 2018, doi: 10.1016/j.ijmedinf.2017.11.003.
[12]. S. Mukherjee, P. Ghosh, and A. Banerjee, “Prediction of cancer risk using structured clinical datasets,” Health Information Science and Systems, vol. 7, no. 1, p. 12, 2019, doi: 10.1007/s13755-019-0075-4.
[13]. A. K. Mishra, S. Yadav, and R. Srivastava, “Feature selection methods for clinical cancer prediction,” Procedia Computer Science, vol. 167, pp. 2053–2061, 2020, doi: 10.1016/j.procs.2020.03.255.
[14]. T. R. Chaudhary and N. Patel, “Electronic health records and predictive analytics in oncology,” Journal of Healthcare Engineering, vol. 2021, Article ID 9987452, 2021, doi: 10.1155/2021/9987452.
[15]. S. K. Jain, P. Kumar, and R. Mehta, “Clinical decision support systems for cancer diagnosis: A review,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 5, pp. 321–329, 2021.

