Atomic Spectroscopy

Article

Predictive Analysis Of Oscc Patient Status Using Machine Learning: A Focus On Lifestyle And Demographic Factors

Wan Muhamad Amir W Ahmad[1]*, Mohamad Nasarudin Adnan[1], Farah Muna Mohamad Ghazali[1], Nor Azlida Aleng[2], Nurfadhlina Abdul Halim[3], Mohamad Shafiq Mohd Ibrahim[4]

[1] School of Dental Sciences, Health Campus, Universiti Sains Malaysia (USM), 16150 Kubang Kerian, Kota Bharu, Kelantan, Malaysia. wmamir@usm.my; nasarudinadnan@student.usm.my; munaghazali@yahoo.com

[2] Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu (UMT), 21030 Kuala Nerus, Terengganu, Malaysia. azlida_aleng@umt.edu.my

[3] Faculty of Science and Technology, Universiti Sains Islam Malaysia, Malaysia Bandar Baharu Nilai, 71800 Nilai, Negeri Sembilan Malaysia. nurfadhlina@usim.edu.my

[4] Kulliyyah of Dentistry, International Islamic University Malaysia (IIUM) Kuantan Campus,Jalan Sultan Ahmad Shah,Bandar Indera Mahkota 25200 Kuantan, Pahang Darul Makmur, Malaysia, shafiq@iium.edu.my

ABSTRACT: Oral Squamous Cell Carcinoma (OSCC) is a major public health concern, with patient outcomes influenced by lifestyle and demographic factors. Accurate insight into these variables is essential to improve prognostic models and guide targeted therapies. Advanced computational methods now offer highly accurate OSCC outcome predictions, enhancing clinical decision-making. Objective: This study aims to construct and validate a predictive model for survival outcomes in OSCC patients. By analyzing the influence of key variables- smoking status, age, betel quid use, and alcohol consumption, the study seeks to quantify each factor's contribution to survival probability in OSCC. Materials and Methods: Conducted as a retrospective analysis, this study employed a machine learning framework, specifically a Multilayer Feedforward Neural Network (MLFFNN), to evaluate data from OSCC patients. The survival status (live or dead) served as the outcome variable, while the predictors included smoking status, age, betel quid use, alcohol consumption, and sex. Model development was performed in R, using sophisticated statistical processes such as data normalization, bootstrap resampling, and systematic data partitioning for training, testing, and validation. The neural network's architecture was refined with hidden layers and a logistic activation function to achieve optimal predictive accuracy. Results: The MLFFNN model identified smoking status as the most influential predictor of OSCC survival (24%), followed by age (16.91%), betel quid use (15.37%), alcohol consumption (8.57%), and sex (10.21%). The model exhibited strong predictive capabilities, with performance metrics such as Mean Absolute Error (MAE) of 0.2628 and Root Mean Squared Error (RMSE) of 0.3466 on the validation data, indicating its reliability. Additionally, it achieved a model accuracy of 73.715% on the validation dataset. The Mean Squared Error (MSE) on the testing data was 0.1638, further reinforcing the model's effectiveness in predictive tasks. Conclusion: This study demonstrates the efficacy of an MLFFNN model in evaluating factors affecting OSCC survival outcomes. The results highlight the pronounced impact of lifestyle behaviours, especially smoking, on survival prognosis. The successful application of a neural network-based approach in R underscores the potential of computational models to contribute meaningfully to OSCC management, offering clinicians a data-driven tool to guide treatment decisions and intervention strategies.

Keywords: OSCC, Multilayer Feedforward Neural Network, Multiple Logistic Regression.

INTRODUCTION

Oral squamous cell carcinoma (OSCC) represents a significant global health concern, particularly in regions with prevalent risk factors such as tobacco use, alcohol consumption, and betel quid chewing (1, 11). These lifestyle choices, along with demographic variables like age and sex, have been extensively studied for their association with OSCC incidence and patient outcomes (10, 12). For instance, smoking and alcohol use are well-established risk factors for OSCC, with studies indicating that combined usage significantly elevates cancer risk (5, 8). Similarly, betel quid chewing, common in Southeast Asia, has been linked to increased OSCC incidence (3, 6).

Despite advancements in treatment, the prognosis for OSCC patients remains variable, often influenced by the aforementioned factors (7). Traditional statistical methods have provided insights into these associations; however, they may not fully capture the complex, non-linear interactions between multiple variables that contribute to patient outcomes (2). In this context, machine learning (ML) approaches, particularly Multilayer Feedforward Neural Networks (MLFFNN), offer a promising avenue for

predictive analysis. MLFFNNs are capable of modelling intricate patterns within data, making them suitable for predicting patient status based on a combination of lifestyle and demographic factors (4).

This study aims to leverage MLFFNNs, enhanced with bootstrap resampling techniques, to predict the survival status (alive or deceased) of OSCC patients. Bootstrap methods are employed to improve the robustness and generalizability of the predictive model by mitigating overfitting and providing more reliable estimates (9). By integrating key independent variables-smoking, age, betel quid usage, alcohol consumption, and sex-this research seeks to develop a predictive framework that can assist clinicians in identifying high-risk patients and tailoring personalized treatment strategies.

The application of machine learning in oncological prognostics represents a significant step toward precision medicine. By accurately predicting patient outcomes based on individual risk profiles, healthcare providers can enhance decision-making processes, optimize resource allocation, and ultimately improve patient survival rates. This study contributes to the growing body of literature advocating for the integration of advanced computational techniques in medical research, particularly in the realm of cancer prognosis (13).

MATERIALS AND METHODS

2.1. Data Collection

The dataset used in this study was obtained from the dental clinic at Hospital Universiti Sains Malaysia (USM) and includes 54 participants with associated lifestyle and demographic attributes. The dependent variable, "Status," represents the survival outcome of oral squamous cell carcinoma (OSCC) patients, coded as 1 for alive and 0 for dead. The independent variables include age, a continuous variable representing the patient's age in years, and four binary variables: smoking status (1 = Yes, 0 = No), betel quid usage (1 = Yes, 0 = No), alcohol consumption (1 = Yes, 0 = No), and sex (1 = Male, 0 = Female). These variables capture key lifestyle factors and demographic characteristics influencing OSCC outcomes, providing a comprehensive dataset for predictive modelling. The integration of these variables allows for an in-depth analysis of their impact on the survival status of OSCC patients, forming the basis for the machine-learning approach employed in this study.

2.2. Study Design

This research adopts a retrospective study design, utilizing data collected from the dental clinic at Hospital Universiti Sains Malaysia (USM) to investigate the relationship between lifestyle and demographic factors and the survival outcomes of oral squamous cell carcinoma (OSCC) patients. A Multilayer Feedforward Neural Network (MLFFNN) with bootstrap resampling techniques is employed as the primary computational model to predict the survival status of patients (alive or deceased) based on independent variables, including smoking, betel quid usage, alcohol consumption, age, and sex. The MLFFNN is evaluated using performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE) to ensure the robustness and accuracy of the predictive model. To further refine the methodology, variable contribution analysis is conducted, quantifying the percentage impact of each independent variable on the survival status. Ethical approval for this study was granted by the Universiti Sains Malaysia Research Ethics and Human Research Committee (USM/JEPeM/16050184), ensuring compliance with ethical standards, including patient privacy and confidentiality throughout the research process.

2.3. Computational Biometry Modeling

The computational modelling in this study utilized a Multilayer Feedforward Neural Network (MLFFNN) framework implemented using R Studio and its associated machine-learning libraries, such as *caret*, *nnet*, and *neuralnet*. This approach integrated bootstrap resampling techniques (e.g., via the *boot* or *rsample* packages) to enhance model stability and mitigate overfitting. The dataset was systematically divided into training (70%) and testing (30%) subsets using functions such as createDataPartition from the caret package to develop and validate the predictive model. Within this framework, the MLFFNN was designed to predict the binary-dependent variable, patient status, defined as 0 for deceased and 1 for alive. The independent variables included smoking, betel quid usage, alcohol consumption, age, and sex. The MLFFNN architecture was constructed using R packages such as *nnet* or *neuralnet*, which allowed for defining an input layer corresponding to the five independent variables, one or more hidden layers optimized for non-linear interactions, and an output layer with a single neuron for binary classification. The output layer utilized a sigmoid activation function to produce probabilities for classification, which is standard in binary classification tasks. The model was trained using backpropagation algorithms with an adaptive learning rate, implemented through the *neuralnet* package, and minimized the binary cross-entropy loss function.

Performance evaluation of the MLFFNN was conducted using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE), calculated using *R* functions. The computational modelling process, combined with variable contribution analysis, provided a robust framework for understanding the influence of lifestyle and demographic factors on OSCC patient outcomes. The implementation and evaluation were carried out in compliance with ethical standards, ensuring data integrity and confidentiality throughout the process.

2.3.1 Multilayer Feed-Forward Neural Network (MLFNN) With Two Hidden Layers Approach

Artificial Neural Networks (ANNs) are computational systems modelled after the structure and function of biological neural networks, often referred to as neural networks (NNs). This study employs a Multilayer Feed-Forward (MLFF) architecture, a specific type of neural network that features one or more layers between the input, hidden, and output layers. As the focus of this study is on a single dependent variable, the MLFF model is designed with a single output node. Figure 1 demonstrates the MLFF model, comprising N input nodes, H hidden nodes, and a single output node. The values of the hidden node h_j, j =1…3

are given by $h_j = g_1\left(\sum_{j=1}^{3} v_{ji} x_i + E_1\right)$ where v_{ji} the output *weight,* E_1 is the bias. The values of the hidden node n_j, j =1…3 are given by $n_j = g_2\left(\sum_{j=1}^{3} v_{ji} h_i + E_2\right)$ where v_{ji} the output *weight.* ,E_2 is the bias. The values of the hidden node Y_j, j =1,2 are given by where $Y_i = g_3\left(\sum_{j=1}^{3} v_{ji} n_i + E_3\right)$ v_{ji} the output *weight,* E_3 is the bias.

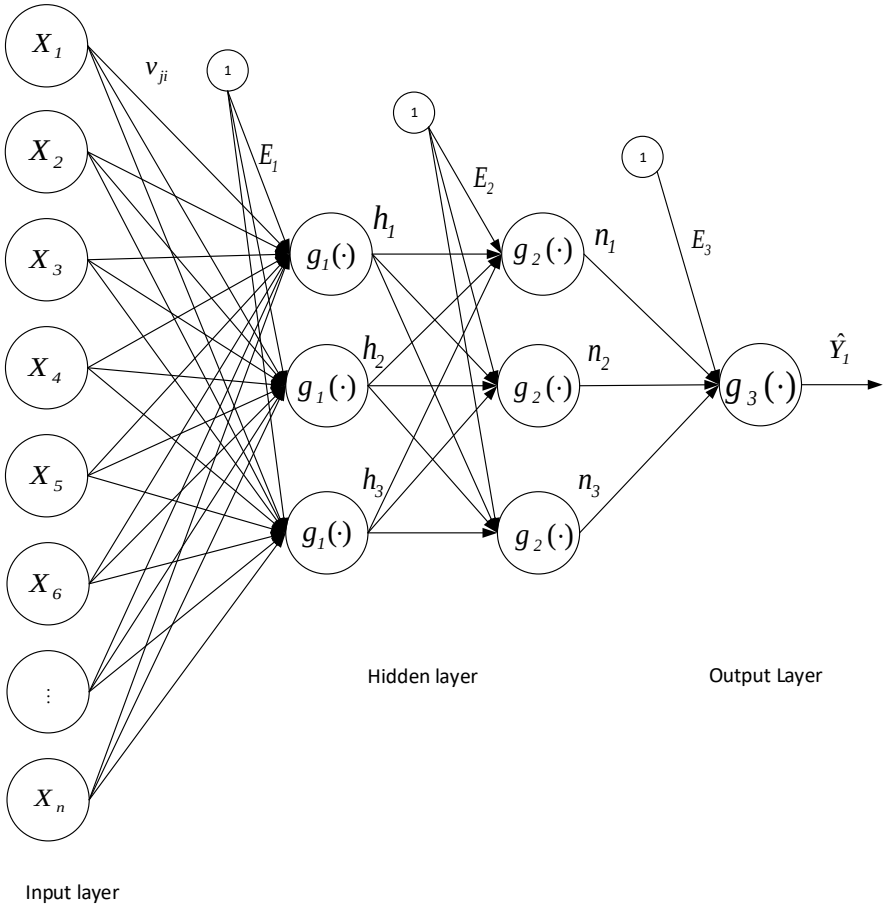

Figure 1 The general architecture of the MLFNN with two hidden layers, N input nodes, and one output node.

2.4. Bootstrap

The methodology for developing the predictive model for patient status begins with data collection, variable selection, and screening to ensure data quality and relevance. A bootstrap sampling technique (n = 1000) is employed to enhance model accuracy and generalizability. The data is then split into 70% for training and 30% for testing to facilitate model development and validation. The modelling process utilizes a Multi-Layer Feedforward Neural Network (MLFFNN) with a logistic activation function, which effectively captures complex relationships among variables and handles binary outcomes to classify patient survival status accurately. The model's performance is evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE), providing a comprehensive assessment of the model's predictive capability and robustness in determining patient survival outcomes.

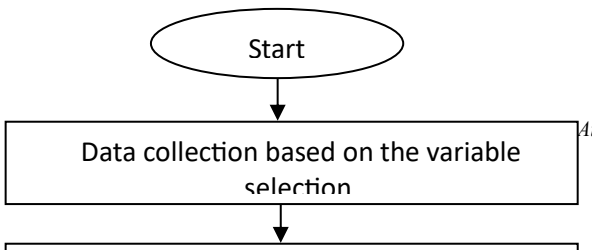

Figure 1. Flowchart of the Proposed Hybrid Methodology for Predictive Modeling Using MLFFNN with Bootstrap Resampling

RESULTS

This study assesses the performance of a Multi-Layer Feed-Forward Neural Network (MLFFNN) employing a logistic activation function within a hybrid framework that incorporates bootstrap sampling. The model demonstrated robust predictive capabilities, achieving a Mean Absolute Error (MAE) of 0.2628 and a Root Mean Squared Error (RMSE) of 0.35 on the validation dataset, underscoring its reliability. Furthermore, the model attained an accuracy of 73.72% on the validation dataset. The Mean Squared Error (MSE) of 0.17 observed on the testing dataset further highlights the model's effectiveness and precision in predictive analysis. Figure 2 illustrates the model architecture visually.

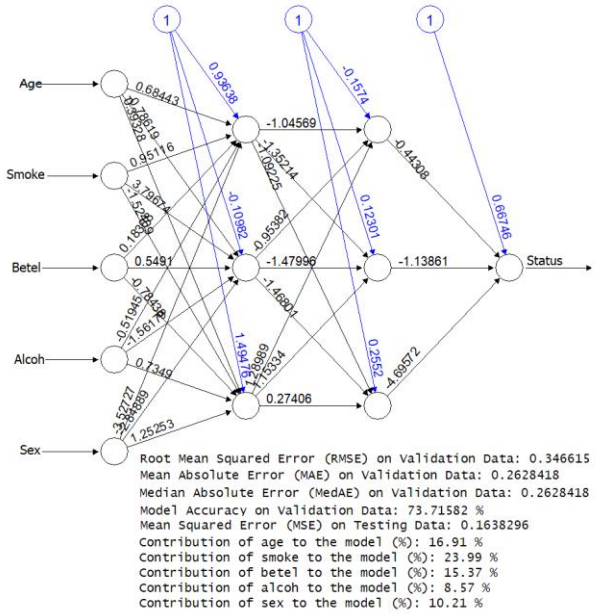

Figure 2. Architecture of the Optimized MLFFNN Model with Five Input Variables, One Hidden Layer, and a Single Output Node for Predicting Patient Survival Status

Contribution of Age

The model reveals that age accounts for 16.91% of the predictive capability for patient survival status, emphasizing its significant role in the prognosis of oral squamous cell carcinoma (OSCC). Age is a well-established factor in cancer outcomes, as older patients often experience compromised immune function, a higher prevalence of comorbidities, and slower recovery rates. These age-associated physiological and health challenges contribute to poorer survival outcomes in elderly patients. The substantial contribution of age in the model underscores its predictive value, advocating for age-specific strategies, such as enhanced screening for older individuals and personalized treatment plans, to improve OSCC management. This observation aligns with the broader medical understanding that older patients necessitate more intensive monitoring and interventions to mitigate their distinct health risks.

Contribution of Smoking

Smoking emerged as the most significant factor influencing the model's predictions, accounting for 23.99% of its contribution. This finding aligns with extensive research over decades, which has consistently identified smoking as a primary risk factor for OSCC and a determinant of poorer survival outcomes. Tobacco is known to contain numerous carcinogens that promote cellular mutations, chronic inflammation, and impaired tissue repair, thereby accelerating disease progression. The substantial contribution of smoking highlights the urgent need to integrate smoking cessation initiatives into OSCC prevention and treatment strategies. Additionally, the model's ability to quantify the impact of smoking underscores its pivotal role in shaping survival outcomes, reinforcing the imperative for public health interventions to reduce smoking prevalence among high-risk populations.

Contribution of Betel Quid Chewing

Betel quid chewing contributes 15.37% to the predictive model, reflecting its significant role in OSCC prognosis, particularly in regions where this practice is culturally ingrained. Betel quid contains areca nut, lime, and often tobacco, all of which have been shown to have carcinogenic effects on the oral mucosa. Prolonged exposure to these substances can lead to chronic irritation and precancerous conditions such as leukoplakia, ultimately increasing the risk of OSCC. The model's inclusion of this variable highlights the need for targeted interventions in regions where betel quid chewing is prevalent. Public health campaigns focusing on education and cessation programs can play a vital role in reducing OSCC incidence and improving survival rates. This result also underscores the importance of culturally sensitive approaches in addressing lifestyle-related cancer risks.

Contribution of Alcohol Consumption

Alcohol consumption accounts for 8.57% of the model's predictive capacity, underscoring its role as a moderate yet significant risk factor in OSCC prognosis. Its synergistic interaction with other risk factors, such as smoking and betel quid chewing, exacerbates cellular damage and carcinogenesis. Alcohol also compromises immune defences and impairs tissue repair mechanisms, thereby increasing the likelihood of cancer progression. Although its contribution is less pronounced compared to smoking or betel quid chewing, alcohol remains a critical factor that warrants attention. Including alcohol consumption in the predictive model provides a comprehensive understanding of OSCC risk factors and emphasizes the necessity of addressing alcohol use alongside other behavioural interventions to enhance patient outcomes.

Contribution of Sex

The model identifies sex as contributing 10.21% to OSCC survival outcomes, underscoring its significant role. Males are generally at greater risk for OSCC, attributed to the higher prevalence of risk behaviours such as smoking, alcohol consumption, and betel quid chewing. Additionally, biological and hormonal differences may influence disease progression and treatment response between sexes. Incorporating sex as a variable enhances the model's ability to account for these complexities, enabling more precise predictions across demographic groups. These findings highlight the importance of personalized treatment strategies that consider sex-based differences, fostering equitable healthcare for both male and female patients.

DISCUSSION

The proposed hybrid methodology has demonstrated remarkable effectiveness in estimating event probabilities, particularly for predicting hypertension risk in patients with dyslipidemia and type 2 diabetes mellitus. By integrating multiple logistic regression (MLR) with a Multi-Layer Feed-Forward Neural Network (MLFFNN), this approach produced a robust and precise predictive model. Traditional non-linear regression models often encounter significant challenges, such as computational complexity and reduced accuracy in modelling predictor-outcome relationships. In contrast, the hybrid methodology leveraged the strengths of both bootstrap sampling and logistic regression, ensuring enhanced validation and reliability across training and testing datasets. The MLFFNN model's ability to predict OSCC survival status was particularly notable. Among the input variables, smoking status emerged as the most influential predictor, contributing 23.99% to the model, followed by age (16.91%), betel quid use (15.37%), sex (10.21%), and alcohol consumption (8.57%). This hierarchy highlights the multifactorial nature of OSCC outcomes, with behavioural and demographic factors playing pivotal roles. The model exhibited strong predictive performance, evidenced by a Mean Absolute Error (MAE) of 0.2628 and a Root Mean Squared Error (RMSE) of 0.3466 on the validation data. Its overall accuracy of 73.72% on the validation dataset, along with a Mean Squared Error (MSE) of 0.1638 on the testing data, further underscores its reliability. The MLFFNN architecture, as depicted in Figure 2, effectively captured complex non-linear relationships through its five input variables, hidden layer, and output node. This hybrid model presents a valuable tool for healthcare providers, enabling early detection and personalized intervention strategies for patients at risk of OSCC-related mortality. The identification of key predictors not only informs targeted prevention efforts but also underscores the importance of considering behavioural and demographic factors in clinical decision-making.

From a methodological perspective, this study advanced medical statistical modelling by integrating bootstrapping with MLR techniques. The variable selection process, guided by clinical expertise, ensured that the chosen predictors were both relevant and interpretable within the clinical context. Bootstrapping played a crucial role in enhancing the robustness of the dataset by generating a comprehensive "mega" file through repeated resampling. This approach mitigated variability and improved the model's generalizability. Following this, the dataset was partitioned into training (70%) and testing (30%) subsets, facilitating rigorous model development and evaluation. Statistical computations were performed using R programming, enabling seamless integration of analytical techniques and yielding a highly effective logistic regression model. Despite its success, this methodology posed several challenges, including the selection of appropriate input variables, data preparation, and standardization for logistic modelling. Addressing these challenges was critical to ensuring reliable and clinically meaningful outcomes. The findings from this study underscore the potential of hybrid predictive models in advancing personalized medicine, fostering equitable healthcare delivery, and improving outcomes for patients with OSCC.

CONCLUSION

This study demonstrates the efficacy of a hybrid machine-learning methodology, combining Multilayer Feedforward Neural Networks (MLFFNN) with bootstrap resampling, in predicting survival outcomes in Oral Squamous Cell Carcinoma (OSCC) patients. The proposed model effectively integrated key lifestyle and demographic variables, including smoking, age, betel quid usage, alcohol consumption, and sex, to evaluate their contributions to patient status. Among these, smoking emerged as the most influential predictor (23.99%), followed by age (16.91%), betel quid usage (15.37%), sex (10.21%), and alcohol consumption (8.57%), highlighting the multifactorial nature of OSCC prognosis.

The model exhibited strong predictive performance, achieving an accuracy of 73.72% on the validation dataset, with reliable error metrics such as a Mean Absolute Error (MAE) of 0.2628 and a Root Mean Squared Error (RMSE) of 0.3466. These results underscore the robustness of the MLFFNN framework in capturing complex, non-linear relationships between variables. Additionally, the incorporation of bootstrap sampling enhanced data generalizability and model stability, addressing challenges commonly associated with overfitting in machine-learning approaches. From a clinical perspective, the findings emphasize the critical role of lifestyle behaviours and demographic factors in OSCC outcomes. This model provides a valuable tool for early detection and personalized intervention strategies, offering clinicians a data-driven approach to optimize treatment decisions. The study also highlights the need for targeted public health initiatives, such as smoking cessation programs and culturally sensitive interventions, to mitigate risk factors associated with OSCC. This research not only advances the application of machine learning in oncological prognostics but also underscores the importance of interdisciplinary collaboration in addressing complex health challenges. Future work could focus on expanding the dataset and integrating additional variables, such as genetic markers and treatment modalities, to further refine the predictive model and enhance its clinical applicability.

ACKNOWLEDGEMENT

The authors express their gratitude to Universiti Sains Malaysia (USM) for their support in funding this study through the Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS/1/2022/STG06/USM/02/10).

REFERENCES

1. Borse, V., Konwar, A. N., & Buragohain, P. (2020). Oral cancer diagnosis and perspectives in India. Sensors International, DOI:. https://doi.org/10.1016/j.sintl.2020.100046
2. Cheung, H. M. C., & Rubin, D. (2021). Challenges and opportunities for artificial intelligence in oncological imaging. Clinical radiology, 76(10), 728–736. https://doi.org/10.1016/j.crad.2021.03.009
3. Cirillo, N., Duong, P. H., Er, W. T., Do, C. T. N., De Silva, M. E. H., Dong, Y., Cheong, S. C., Sari, E. F., McCullough, M. J., Zhang, P., & Prime, S. S. (2022). Are There Betel Quid Mixtures Less Harmful than Others? A Scoping Review of the Association between Different Betel Quid Ingredients and the Risk of Oral Submucous Fibrosis. Biomolecules, 12(5), 664. https://doi.org/10.3390/biom12050664
4. Dixit, S., Kumar, A., & Srinivasan, K. (2023). A Current Review of Machine Learning and Deep Learning Models in Oral Cancer Diagnosis: Recent Technologies, Open Challenges, and Future Research Directions. Diagnostics, 13(7), 1353. https://doi.org/10.3390/diagnostics13071353
5. Eloranta, R., Vilén, S.T., Keinänen, A., Salo, T., Qannam, A., Bello, I., Snäll, J. (2024). Oral squamous cell carcinoma: Effect of tobacco and alcohol on cancer location. Tob. Induc. Dis, 22, 112. https://doi.org/10.18332/tid/189303
6. Jasim, A., Li, X., Octavia, A., Gunardi, I., Crocombe, L., & Sari, E. F. (2024). The association between betel quid use and oral potentially malignant and malignant disorders in Southeast Asian and Pacific regions: a systematic review and meta-analysis with GRADE evidence profile. Frontiers in oral health, 5, 1397179. https://doi.org/10.3389/froh.2024.1397179
7. Kim, M. J., & Ahn, K. M. (2024). Prognostic factors of oral squamous cell carcinoma: the importance of recurrence and pTNM stage. Maxillofacial plastic and reconstructive surgery, 46(1), 8. https://doi.org/10.1186/s40902-024-00410-3
8. Mello, F. W., Melo, G., Pasetto, J. J., Silva, C. A. B., Warnakulasuriya, S., & Rivero, E. R. C. (2019). The synergistic effect of tobacco and alcohol consumption on oral squamous cell carcinoma: a systematic review and meta-analysis. Clinical oral investigations, 23(7), 2849–2859. https://doi.org/10.1007/s00784-019-02958-1
9. Reeves, M., Bhat, H. S., & Goldman-Mellor, S. (2022). Resampling to address inequities in predictive modeling of suicide deaths. BMJ health & care informatics, 29(1), e100456. https://doi.org/10.1136/bmjhci-2021-100456

10. Shenoi, R., Devrukhkar, V., Chaudhuri, Sharma, B. K., Sapre, S. B., & Chikhale, A. (2012). Demographic and clinical profile of oral squamous cell carcinoma patients: a retrospective study. Indian journal of cancer, 49(1), 21–26. https://doi.org/10.4103/0019-509X.98910
11. Warnakulasuriya, S., & Chen, T. H. H. (2022). Areca Nut and Oral Cancer: Evidence from Studies Conducted in Humans. Journal of dental research, 101(10), 1139–1146. https://doi.org/10.1177/00220345221092751
12. Xu, Q., Wang, C., Li, B., Kim, K., Li, J., Mao, M., Qin, L., Li, H., Huang, X., Xing, R., Han, Z., & Feng, Z. (2019). The impact of age on oral squamous cell carcinoma: A longitudinal cohort study of 2,782 patients. Oral diseases, 25(3), 730–741. https://doi.org/10.1111/odi.13015
13. Zhang, B., Shi, H., & Wang, H. (2023). Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach. Journal of multidisciplinary healthcare, 16, 1779–1791. https://doi.org/10.2147/JMDH.S410301