horizontal line

**Project Summary**

|  |  |
| --- | --- |
| Batch details | PGP-DSE OCT’24 |
| Team members | Niranjan Panda, Mudipudi Venkata Sai Tarun, Revanth K S, Vaishnavi Jaiswal, Anlie Mary Anil |
| Domain of Project | Healthcare: Oral Cancer Dataset |
| Proposed project title | Oral Cancer Prediction Based on Clinical and Lifestyle Factors |
| Group Number | 4 |
| Team Leader | Niranjan Panda |
| Mentor Name | MS Vidhya Kannaiah |

Date: 31-March-2025

Signature of the Mentor Signature of the Team Leader

# 

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| SL NO | Topic | Page No |
| 1 | Overview and Executive Summary about Dataset | 1 |
| 2 | Business problem statement | 2 |
| 3 | Topic survey in Brief | 2 |
| 4 | Critical assessment of topic survey | 3 |
| 5 | Methodology | 4 |
| 6 | References | 6 |

**Project Details**

# Overview

Oral cancer is a significant global health concern, with multiple risk factors contributing to its onset, progression, and treatment outcomes. The goal of this project is to leverage machine learning techniques to predict oral cancer diagnosis based on various demographic, lifestyle, and medical attributes. Using historical patient data, we aim to develop a predictive model that can assist in early detection, thereby improving patient outcomes and optimizing healthcare strategies.

# Executive Summary about the Dataset

The Oral Cancer Prediction Dataset consists of patient records with various demographic, lifestyle, clinical, and genetic factors to assess the likelihood of oral cancer. The dataset is designed to aid in early detection by leveraging machine learning techniques for risk assessment and prediction.

The dataset comprises a significant number of patient records, each containing multiple numeric and categorical features that capture risk factors associated with oral cancer. These features are classified into the following categories:

* Demographic Features: Age, Gender, Ethnicity, and Family History.
* Lifestyle Factors: Tobacco Use (Smoking, Chewing), Alcohol Consumption, Betel Quid Chewing, Dietary Habits, and Oral Hygiene Practices.
* Clinical Features: Presence of Oral Lesions, Tumour Stage, HPV Infection, Blood Sugar Levels, and Other Medical Conditions.

Analysis of Numeric Features: The dataset contains several numeric attributes, which provide crucial insights into patient health. Some of the key numeric features include:

* Age – A critical factor in assessing cancer risk.
* Duration of Tobacco and Alcohol Use – Measures the long-term impact of harmful substances.
* Frequency of Betel Quid Chewing – Associated with a high risk of oral cancer.
* Oral Lesion Size and Severity – Important for evaluating cancer progression.
* Body Mass Index (BMI) – May indicate lifestyle and dietary habits that contribute to cancer risk.

A statistical analysis of these numeric features includes:

* Histogram and KDE Plots to analyze data distribution and detect skewness.
* Boxplots to identify and handle outliers in attributes like lesion size and frequency of tobacco use.
* Correlation Analysis to evaluate the relationship between numeric features and oral cancer occurrence.

**Business problem statement**

Oral cancer remains a critical public health issue due to late-stage diagnosis, leading to increased mortality rates and treatment costs. The traditional diagnostic approach, which relies on biopsies and expert evaluations, is time-consuming and often unavailable in low-resource settings. Additionally, early symptoms of oral cancer, such as persistent mouth sores, red or white patches, and difficulty swallowing, are often misdiagnosed or overlooked, further delaying treatment.

The lack of access to specialized healthcare professionals and diagnostic equipment in many regions contributes to the late detection of the disease. Treatment costs for oral cancer, including surgery, radiation therapy, and chemotherapy, are significantly higher when diagnosed at an advanced stage, imposing a financial burden on patients and healthcare systems. Furthermore, survival rates decrease dramatically with delayed diagnosis. Thus, an AI-driven diagnostic tool that utilizes predictive modelling based on risk factors and early symptoms can be a game-changer in improving early detection and reducing the healthcare burden. By leveraging this dataset, the project aims to develop a scalable, cost-effective solution that enables timely and accurate diagnosis, facilitating early intervention and better patient outcomes.

**Topic survey in Brief**

Machine learning and deep learning have transformed medical diagnostics, particularly in cancer detection. Various studies have demonstrated the effectiveness of convolutional neural networks (CNNs) and machine learning algorithms such as support vector machines (SVMs) and random forests in classifying cancerous lesions. Advances in AI have shown significant improvements in medical imaging analysis, where models trained on large datasets can detect patterns indicative of cancer with high accuracy.

Several research efforts have focused on feature extraction techniques for identifying oral cancer symptoms from histopathological images and patient medical records. Transfer learning approaches, where pre-trained deep learning models are fine-tuned for specific medical datasets, have proven useful in scenarios where labeled data is limited. Moreover, combining clinical, lifestyle, and demographic data with medical imaging has improved predictive accuracy, making AI-driven diagnostic systems more robust. However, challenges such as data scarcity, variability in symptoms across populations, and the need for interpretable AI models still need to be addressed for widespread adoption.

**Critical Assessment of Topic Survey**

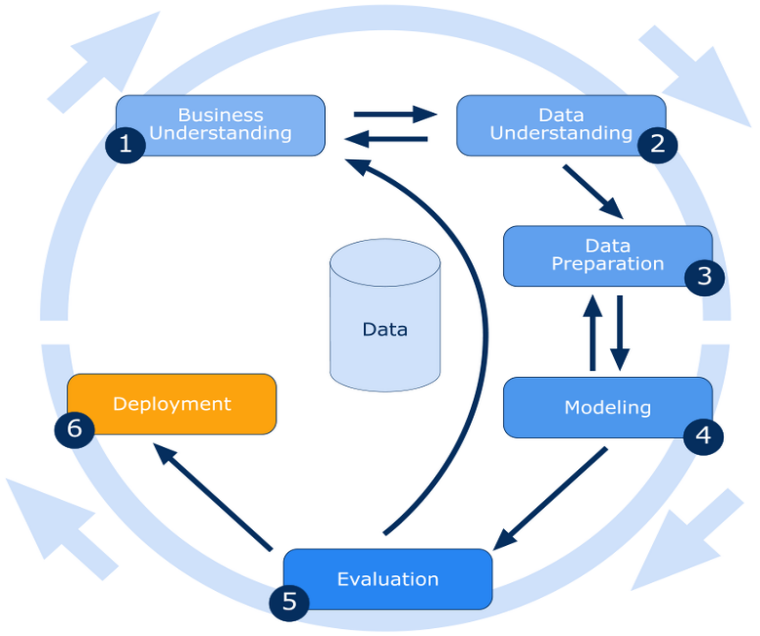
Despite progress in AI-driven cancer detection, several challenges remain. Data imbalance, variations in clinical imaging techniques, and the lack of standardized diagnostic criteria hinder model generalization. The dataset includes a mix of categorical and numerical variables, requiring robust feature engineering. Furthermore, since oral cancer risk factors vary by geographic region, socioeconomic status, and genetic predisposition, the AI models must be trained on diverse datasets to ensure unbiased predictions.

Another major challenge is the black-box nature of deep learning models. Many AI-driven diagnostic systems struggle with interpretability, making it difficult for medical professionals to trust and validate AI-generated predictions. Explainable AI (XAI) techniques, such as SHAP and Grad-CAM (Gradient-weighted Class Activation Mapping), are being explored to provide transparent insights into model decisions.

Ethical considerations, including patient data privacy, informed consent, and adherence to regulatory standards, are also critical factors in deploying AI models in healthcare. Regulatory bodies such as the FDA and WHO have emphasized the need for rigorous validation before implementing AI-based diagnostics in clinical settings. Our project aims to address these challenges by integrating explainable AI techniques, ensuring fairness in model predictions, and complying with ethical guidelines to make AI-driven oral cancer detection both reliable and accessible.

# Methodology

In this project, we aim to predict the likelihood of oral cancer in patients based on demographic, lifestyle, and clinical attributes. By leveraging machine learning techniques, we seek to develop an early detection system that assists healthcare professionals in diagnosis and risk assessment.



Understanding the problem:

* + Prediction of oral cancer presence based on attributes related to medical, lifestyle, and genetic factors.
  + Impact of each attribute in determining the likelihood of oral cancer occurrence.
* Data Collection and Understanding:
  + Collecting relevant features as per the problem statement, including feature belonging medical, lifestyle, and genetic factors.
  + Features including Patient Demographics (such as Age, Gender, Family History, Ethnicity etc.), Clinical Features (Oral Lesions, Tumor Stage, HPV Infection), and Lifestyle Factors (such as Tobacco Use, Alcohol Consumption, Betel Quid Chewing, Poor Oral Hygiene, etc.).
  + Integrating data into a single comprehensive dataset.
  + Examine and understand the data to uncover patterns and insights.
* Data Preparation and Preprocessing:
  + Handling Missing Values: Filling missing values using statistical imputation (mean, median, mode)
  + Removing Duplicates: Identifying and eliminating redundant records to prevent bias.
  + Outlier Detection and Treatment: Using statistical methods such as Interquartile Range (IQR) and Boxplots to identify and handle outliers.
  + Data Type Treatment: Ensuring numerical and categorical variables are correctly classified.
  + Data Cleaning: Checking for inconsistencies, anomalies, and erroneous data points. Removing irrelevant columns.
  + Data Transformation: Scaling numerical features using Standard Scaling (Z-score normalization) or Min-Max Scaling for consistency. Converting categorical variables into numerical representations using One-Hot Encoding or Label Encoding.
  + Feature Selection: Evaluating relationships between features and the target variable using Chi-square test, ANOVA, and correlation analysis. Applying Recursive Feature Elimination (RFE) to retain the most significant features.
* Model Building and training:
  + Model building refers to the process of using the data to train a machine learning model and to make predictions based on unseen data.
  + Model Training: Training different models to learn patterns in the data, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Gradient Boosting Techniques (AdaBoost, XGBoost, LightGBM)
  + Use of Ensemble Learning to combine multiple models for improved accuracy.
  + Hyperparameter Tuning using Grid Search and Random Search to optimize model performance.
* Model Evaluation:
  + Evaluating the performance of a model by applying various metrics to measure its accuracy, precision, and effectiveness.
  + Evaluating how well the model performs by applying metrics such as accuracy, precision, recall, Confusion Matrix and F1 score, ROC-AUC.

# References:

1. <https://www.who.int/news-room/fact-sheets/detail/oral-health>
2. <https://www.sciencedirect.com/science/article/abs/pii/S1386505621001830>
3. <https://onlinelibrary.wiley.com/doi/abs/10.1111/jop.13157>
4. <https://www.mdpi.com/2072-6694/16/3/617>
5. <https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.34277>
6. <https://www.sciencedirect.com/science/article/abs/pii/S1368837516302135>
7. <https://ieeexplore.ieee.org/abstract/document/5986713>

|  |  |
| --- | --- |
| Original owner of data | Ankush Panday |
| Data set information | https://www.kaggle.com/datasets/ankushpanday2/oral-cancer-prediction-dataset/data |
| Any past relevant articles using the dataset | - |
| Reference | Kaggle |
| Link to web page | [Oral Cancer prediction Dataset](https://www.kaggle.com/datasets/ankushpanday2/oral-cancer-prediction-dataset/data) |