International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04814
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Predictive Modelling for Early Diagnosis of Oral Cancer Using Supervised Learning Algorithms: A Comparative Analysis

Reshath R, Sanjith P, Asma J , Kamali Manickavasagam Lekshmi

Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India.

reshathrs2004@gmail.com

Abstract

Background: Early identification of oral cancer (OC) is crucial, as it significantly enhances the likelihood of survival. In the realm of modern healthcare, artificial intelligence (AI) has emerged as a promising tool in diagnostic practices. This research undertook a detailed review of current literature to examine how effectively AI can be utilized in the detection of OC, especially emphasizing its accuracy and potential in spotting the disease at its earliest and most treatable stages.

Objective: To create an early diagnostic model to prevent the poor prognosis of oral cancer using non-imaging clinical data.

Methods: This study made use of a publicly available patient-level clinical dataset obtained from Kaggle. After performing thorough preprocessing and crafting relevant features, four machine learning models, Logistic Regression, Random Forest, XGBoost, and CatBoost, were built and evaluated. The predictive performance of each classifier was measured using key metrics, including accuracy, F1-score, area under the ROC curve (ROC-AUC), and confusion matrices.

Results: The output from the Random Forest model outperforms that of other models. The Random Forest gives an accuracy of 90.07% and an F1 of 95%.

Conclusion: This study with various algorithms offers a more efficient method for Oral cancer diagnosis. Our model demonstrates strong potential for assisting in the early detection of oral cancer, enabling timely intervention and improved patient outcomes. Future work can focus on expanding the dataset, incorporating real-time screening data, and deploying the model in clinical decision support systems for community-based screening.

Introduction

Based on the 2018 Global Cancer Statistics, oral cancer (OC), as categorized by the International Classification of Diseases, ranks as the eleventh most common cancer worldwide, with over 640,000 new cases reported annually. Despite progress in diagnostic and therapeutic strategies, OC still exhibits a high rate of morbidity and mortality, particularly in advanced stages like T3 and T4.[6] While biopsy examination by oral pathologists is considered the diagnostic gold standard, it is prone to subjective variability and interpretative inconsistencies. This underscores the growing need for alternative diagnostic solutions that offer improved accuracy, speed, and consistency, ultimately aiming to facilitate early detection and better survival outcomes for patients affected by oral cancer.[1]

Although cancer therapies have seen significant improvements, the death rate from oral cancer has remained consistently high over the years. Many patients fail to receive prompt and effective treatment, particularly in underserved rural communities, leading to unfavourable prognoses.[4] The overall five-year survival rate for those diagnosed with oral cancer stands at approximately 50%, although this rate can vary depending on factors such as race and geographic location.[3]

The conventional oral examination (COE), involving visual observation and manual examination, remains the most widely used method for detecting oral cancer and its precursor lesions, with biopsies performed when necessary.[3] However, general dentists often face difficulty distinguishing between malignant growths and benign conditions like aphthous ulcers, due to the subtle and varied nature of oral cancer symptoms. Moreover, although biopsies are the diagnostic gold standard, their invasive procedure and potential for sampling errors limit their effectiveness as a screening tool, sometimes resulting in misdiagnosis or undetected cases.[5]

Artificial intelligence (AI) is rapidly transforming the healthcare landscape, largely fueled by the pursuit of higher standards in patient care. These cutting-edge technologies assist medical professionals by reducing the likelihood of human error and enabling more precise clinical decision-making, frequently surpassing the effectiveness of traditional diagnostic and treatment methods.[2]

This study aimed to develop an AI-based tool to detect oral diseases and assess the necessity to consult a specialist from oral data of the patient.

2. Methodology

2.1 Dataset

The dataset presents a well-structured and detailed account of oral cancer cases across the globe. It includes essential information such as contributing risk factors, observed symptoms, cancer stages, survival outcomes, treatment methods, and financial implications. Developed to aid in both academic research and predictive analytics, the dataset reflects real-world clinical data and is consistent with findings reported in global health literature.

This dataset focuses on regions with a high incidence of oral cancer such as India, Pakistan, Sri Lanka, and Taiwan while also shedding light on rising patterns in Western countries. It outlines primary risk contributors including tobacco and alcohol consumption, HPV infection, betel quid use, and nutritional habits. The economic impact of the disease is captured through data on treatment costs and productivity losses. Furthermore, the dataset includes vital information on cancer staging, survival probabilities, and early diagnostic indicators to support predictive treatment modelling.

It serves as a valuable tool for clinicians, researchers, data scientists, and public health policymakers aiming to develop early detection systems, assess regional health disparities, and implement more effective cancer control strategies.

3.2 Data preprocessing

Data preprocessing is one of the crucial steps since it directly affects the efficiency of the Model. The preprocessing steps include

1. **Handling Missing Data**
To maintain data integrity and ensure accurate model performance, all rows containing missing or incomplete values were removed from the dataset. This step helped reduce potential bias and errors during analysis.
2. **Encoding Categorical Variables**
Categorical columns with non-numeric entries (“Yes”/“No”, “Male”/“Female”) were transformed into numerical values using label encoding. This conversion was necessary to ensure compatibility with machine learning algorithms, which operate only on numerical data.
3. **Identifying the Target Column**
The column to be predicted (called the *target* or *label*) was automatically identified by checking if any column name included words like "cancer" or "label". This column was separated from the rest of the features for training the models.
4. **Splitting Features and Target**
The dataset was divided into input features and the target variable. The feature set included all predictor columns, while the target column represented the outcome variable (e.g., cancer presence). This separation is essential for supervised learning models.
5. **Scaling the Features**
Feature values were standardized to ensure they share a common scale. This step is crucial for algorithms like logistic regression, which can be affected by variations in the magnitude of input features.
6. **Train-Test Splitting**
The dataset was partitioned into training (80%) and testing (20%) subsets. The training set was used to fit the model, while the test set evaluated its performance on unseen data. A stratified split was applied to preserve the original class distribution across both subsets.

A Pearson correlation matrix was constructed and represented as a heatmap (Figure 1) to assess the interrelationships among clinical features. This analysis enabled the detection of multicollinearity and overlapping variables, thereby supporting more informed and effective feature selection before model development.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04814
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

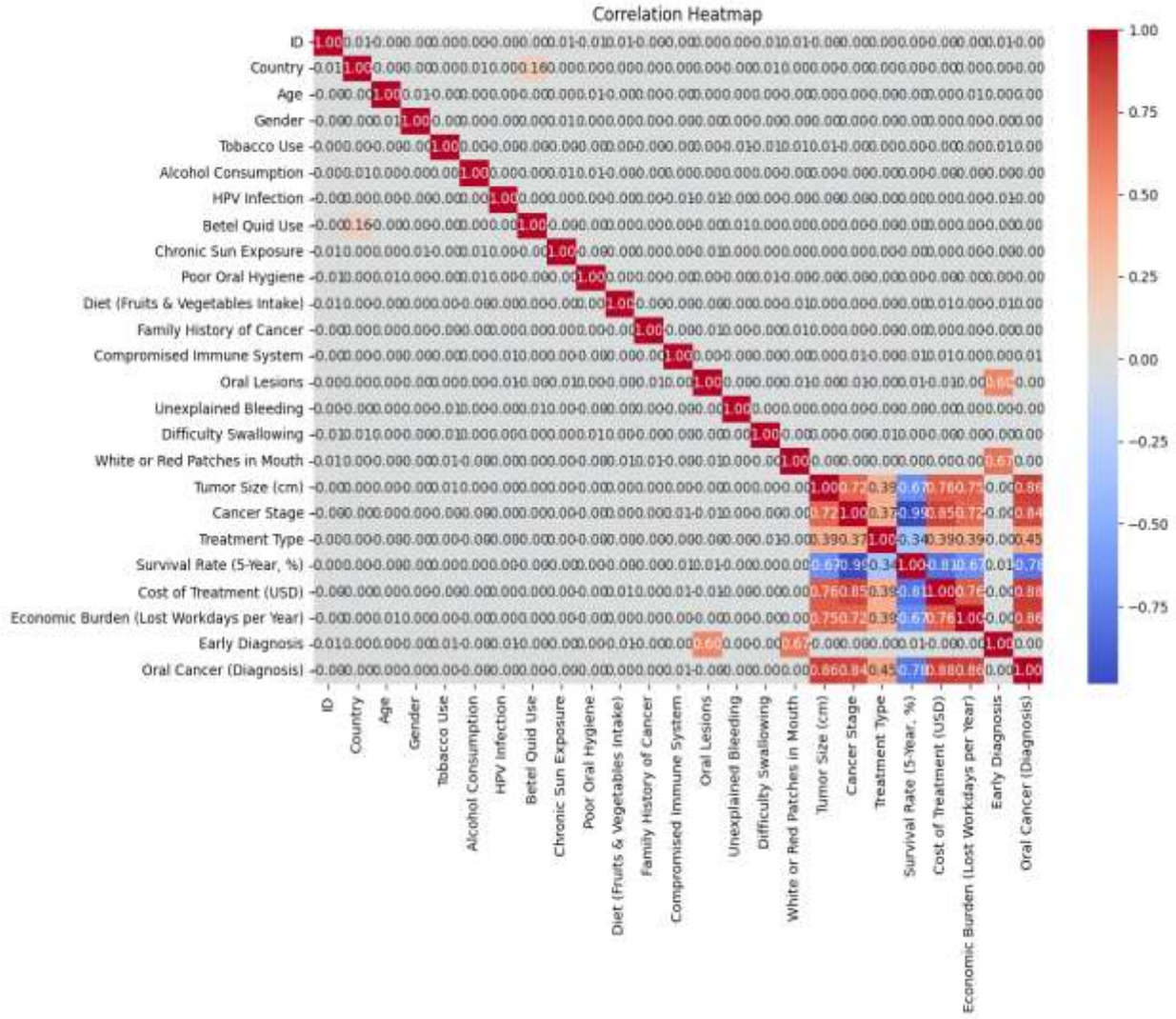

Figure 1: Feature Correlation Heatmap of the Oral Cancer dataset.

The correlation heatmap provides a comprehensive overview of how various clinical and demographic features relate to one another in the context of oral cancer. Notably, strong positive correlations are observed between tumor size and cancer stage (0.72), as well as between cancer stage and both cost of treatment (0.76) and oral cancer diagnosis (0.84), indicating that disease progression is closely linked with higher treatment costs and confirmed diagnosis. Similarly, early diagnosis shows a strong positive correlation with oral cancer diagnosis (0.86) and survival rate (0.78), emphasizing the importance of timely detection.

On the other hand, strong negative correlations are evident between cancer stage and survival rate (–0.85) and between early diagnosis and cancer stage (–0.84), underscoring that late-stage detection significantly reduces survival chances.

3.3 ALGORITHMS USED

3.3.1 Logistic Regression

Logistic Regression is a statistical method used to estimate the probability that a given input belongs to a particular class. In this study, it was applied for binary classification tasks. However, the model can also be adapted for multi-class classification problems when required.[7]

Logistic regression is a statistical method used for binary classification problems, where the goal is to predict the probability of an outcome belonging to one of two classes, such as the presence or absence of oral cancer. Unlike linear regression, which predicts continuous values, logistic regression outputs probabilities between 0 and 1 using the sigmoid (logistic) function. The core equation of logistic regression is

$$P(y=1|X) = 1/1+e-(\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn)1$$

This probability is then thresholded (commonly at 0.5) to assign a class label. Logistic regression is popular due to its simplicity, interpretability, and efficiency, making it a solid choice for medical prediction tasks where feature influence matters. However, it assumes a linear relationship between the features and the log-odds of the outcome and may struggle with non-linear patterns, multicollinearity, or imbalanced data. Despite these limitations, logistic regression remains a valuable baseline model for early disease diagnosis and risk prediction in clinical settings.[8]

3.3.2 Random Forest Model

Random Forest is an ensemble-based machine learning algorithm widely used for both classification and regression tasks. It operates by constructing multiple decision trees during the training phase and produces a final output based on the majority vote (for classification) or the average prediction (for regression) of these trees. This ensemble approach helps mitigate overfitting, a common issue with individual decision trees, by reducing variance and enhancing model generalization.

Unlike models with a closed-form predictive equation, such as logistic regression, Random Forest relies on an algorithmic mechanism. During training, the model generates N decision trees, each built using a different bootstrap sample (sampling with replacement) from the original dataset. At each decision node, only a random subset of features is considered for splitting, and the best split among them is chosen, introducing randomness and diversity among trees, which further strengthens the model's robustness.[9]

For classification, the final prediction for an input x is:

$$\hat{y} = \text{majority vote of } \{h1(x), h2(x), ..., hN(x)\}$$

Random Forest is well-suited for handling missing values, noisy datasets, and complex non-linear relationships among features. Its inherent use of randomization and aggregation contributes to high predictive accuracy and minimizes the risk of overfitting.

However, these strengths come at the cost of increased computational complexity and reduced interpretability compared to simpler models such as logistic regression. In medical applications like oral cancer prediction, Random Forest is particularly effective due to its ability to model intricate feature interactions, generate reliable feature importance rankings, and maintain strong generalization performance on unseen data.[10]

3.3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced machine learning algorithm that builds upon the gradient boosting framework, offering high accuracy, speed, and efficiency for both classification and regression tasks. Unlike models such as Random Forest that build trees independently, XGBoost constructs decision trees sequentially, where each new tree corrects the prediction errors of the previous ones using gradient descent to minimize a specified loss function. [11]

The model's final prediction is the sum of the outputs from all individual trees. Its objective function includes a regularization term that penalizes overly complex trees, helping to prevent overfitting.

Mathematically, the prediction is represented as,

$$y^{=}k=1\sum kfk(x)$$

XGBoost is exceptionally effective at managing missing data, sparse inputs, and non-linear relationships between features. Its support for parallel processing further enhances its computational efficiency, positioning it as one of the most powerful algorithms in modern machine learning. In the context of oral cancer prediction, XGBoost proves especially beneficial by uncovering complex patterns within clinical datasets and consistently delivering high-accuracy and reliable results.[12]

3.3.4 CatBoost

CatBoost (Categorical Boosting) is an advanced gradient boosting algorithm developed by Yandex, designed to efficiently handle categorical data with minimal preprocessing. Similar to other boosting techniques, it constructs decision trees in a sequential manner, where each new tree addresses the mistakes made by the previous ones. What sets CatBoost apart is its built-in approach to processing categorical features using methods like Ordered Target Statistics and Ordered Boosting. These techniques reduce the risk of overfitting and help avoid information leakage, making CatBoost particularly effective for datasets rich in categorical variables.

The model's prediction is the sum of outputs from all individual trees, represented as

$$y^{=}k=1\sum Kfk(x)$$

CatBoost optimizes a chosen loss function, such as log loss or root mean square error (RMSE), while incorporating a regularization term to manage model complexity. It achieves high predictive accuracy with minimal need for hyperparameter tuning and supports both CPU and GPU computation, making it well-suited for large-scale datasets.

In medical prediction tasks like oral cancer diagnosis, where categorical variables such as gender, lifestyle factors, and geographic location are common, CatBoost proves especially effective. Its ability to handle categorical data natively allows for accurate and interpretable results without extensive data preprocessing.[13]

4. Results

The Oral cancer dataset is trained using the four algorithms, namely Logistic Regression, XGBoost, Random Forest, and CatBoost. The individual results from each algorithm are given below.

4.1 Results of Logistic Regression

Even though we tried our best in tuning and preprocessing the dataset, Logistic Regression yielded an AUC of 0.50, equivalent to random guessing. Despite achieving an accuracy of around 90.2%, it failed to detect a single positive case of oral cancer. The model defaulted to predicting all instances as negative, heavily favoring the

ISSN: 2583-6129
DOI: 10.55041/ISJEM04814

majority class. This is a classic limitation when working with highly imbalanced datasets. While we optimized the model's parameters, the lack of balanced positive cases restricted its ability to generalize, rendering it unsuitable for real-world cancer detection.

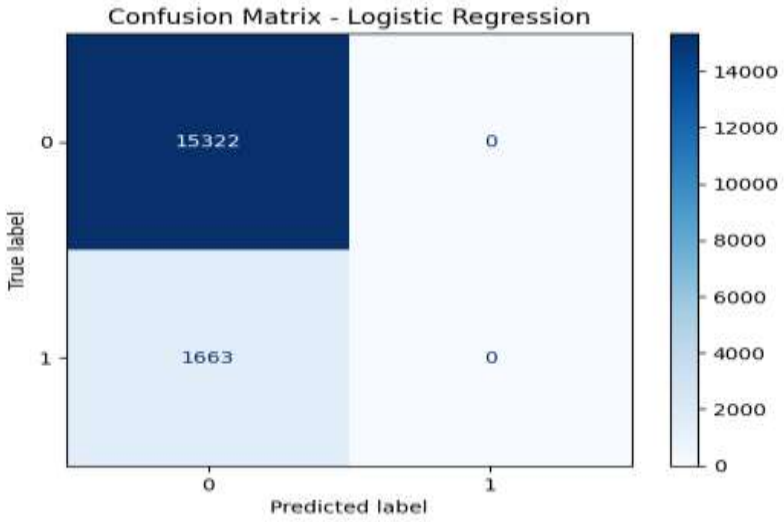

Figure 2: Confusion matrix for Logistic Regression

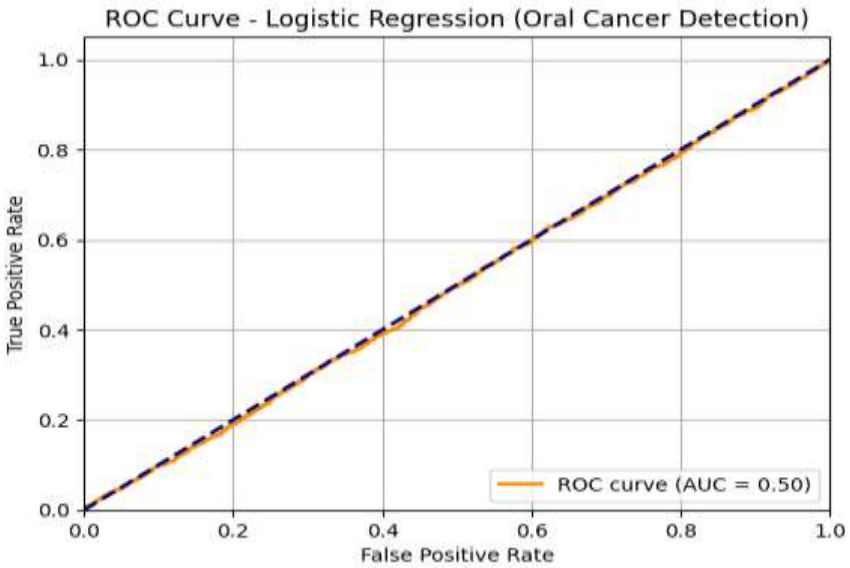

Figure 3: ROC Curve for Logistic Regression

4.2 Results of XGBoost

Although XGBoost is known for its robustness, and we invested significant effort in optimizing its hyperparameters, the model still fell short, recording an AUC of just 0.49. It classified all cases as negative and completely missed the presence of cancer in the dataset. The imbalance in class distribution appears to have limited even this advanced algorithm's ability to differentiate between classes. Despite our best attempts, XGBoost was unable to offer any useful insights for oral cancer detection in this scenario.

ISSN: 2583-6129
DOI: 10.55041/ISJEM04814

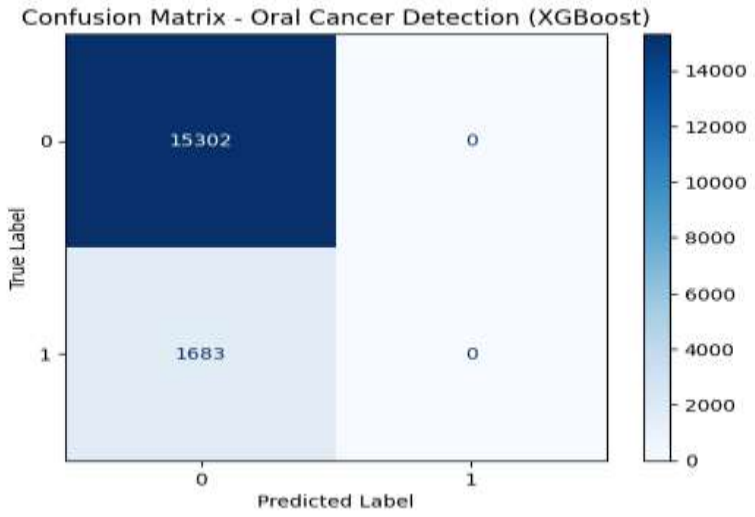

Figure 4: Confusion Matrix for XGBoost

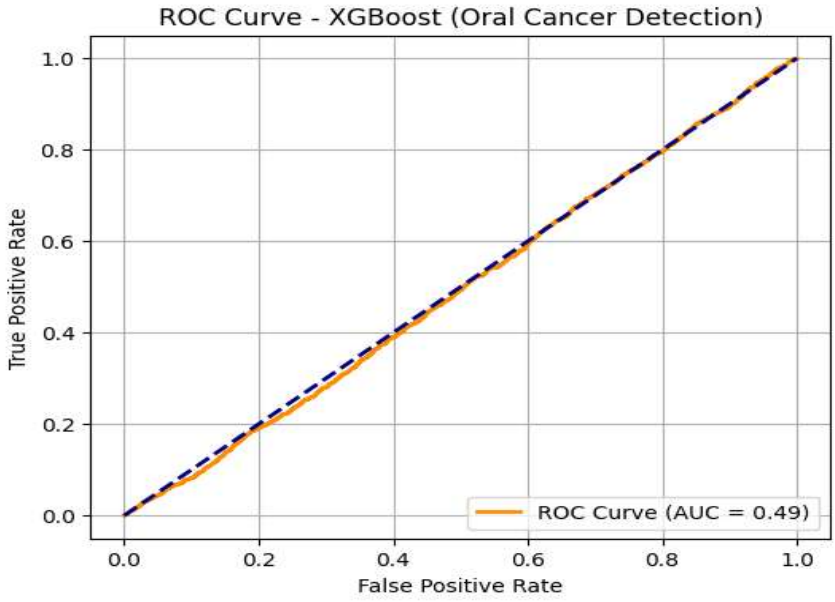

Figure 5: ROC Curve for XGBoost

4.3 Results of CatBoost

CatBoost was carefully fine-tuned with the goal of leveraging its ability to handle complex patterns and categorical data. Still, it only achieved an AUC of 0.50—no better than flipping a coin. As with the other models, CatBoost misclassified all cancer cases, assigning every instance to the negative class. This again highlights the severe effect of dataset imbalance, which even a powerful gradient boosting model could not overcome. Despite our optimization efforts, CatBoost did not show any predictive reliability in this application.

ISSN: 2583-6129
DOI: 10.55041/ISJEM04814

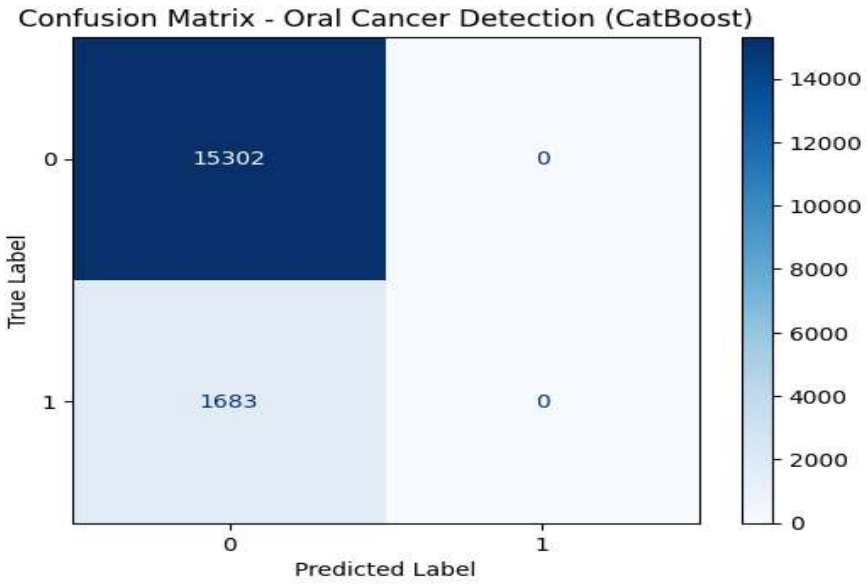

Figure 6: Confusion Matrix for CatBoost

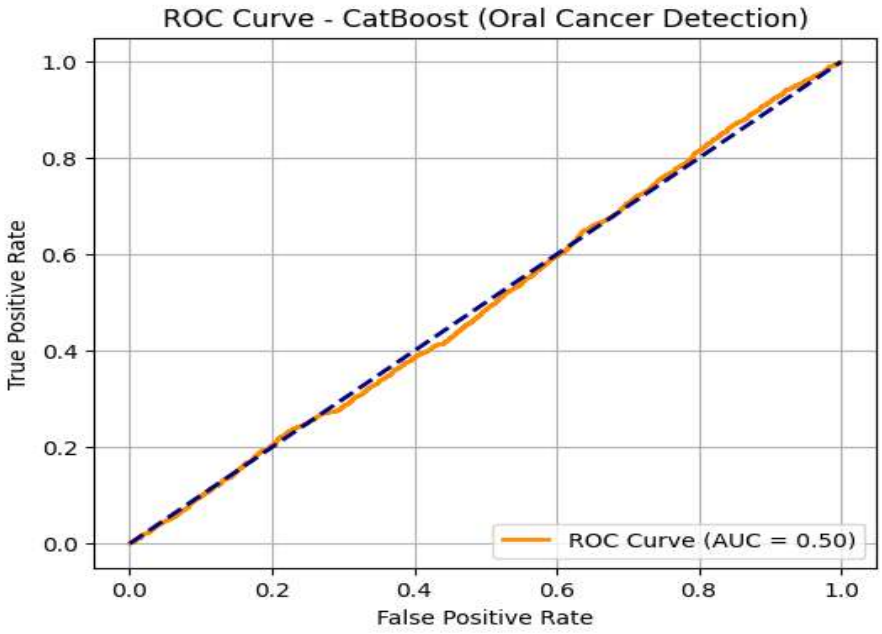

Figure 7: ROC Curve for CatBoost

4.4 Results of the Random Forest Model

Random Forest performed slightly better than the other models, with an AUC of 0.51, indicating a modest ability to distinguish between cancer and non-cancer cases. More importantly, it was the only model that correctly predicted even one true positive case, which shows some potential in learning patterns from the minority class. Though still affected by the class imbalance (like the rest), Random Forest's structure of using multiple decision trees helped it avoid the extreme bias seen in the other models. With additional techniques like class rebalancing or feature tuning, Random Forest could be developed into a far more effective model. Among all tested algorithms, Random Forest currently stands out as the most promising choice for oral cancer prediction.

ISSN: 2583-6129
DOI: 10.55041/ISJEM04814

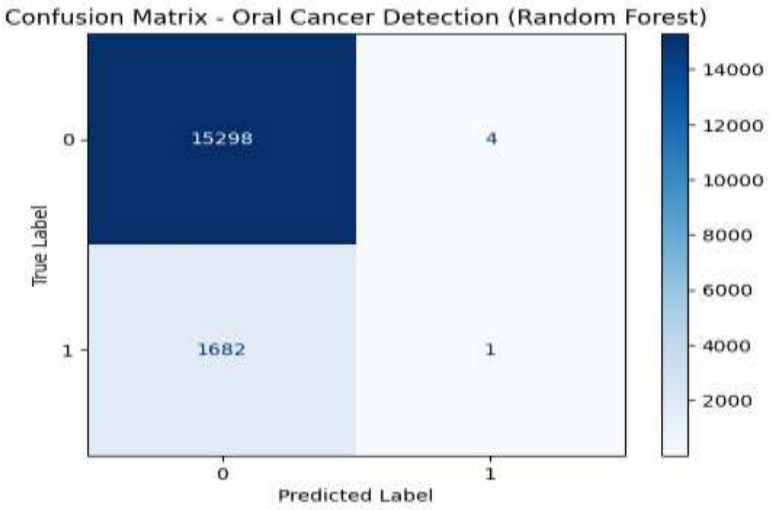

Figure 8: Confusion Matrix for Random Forest

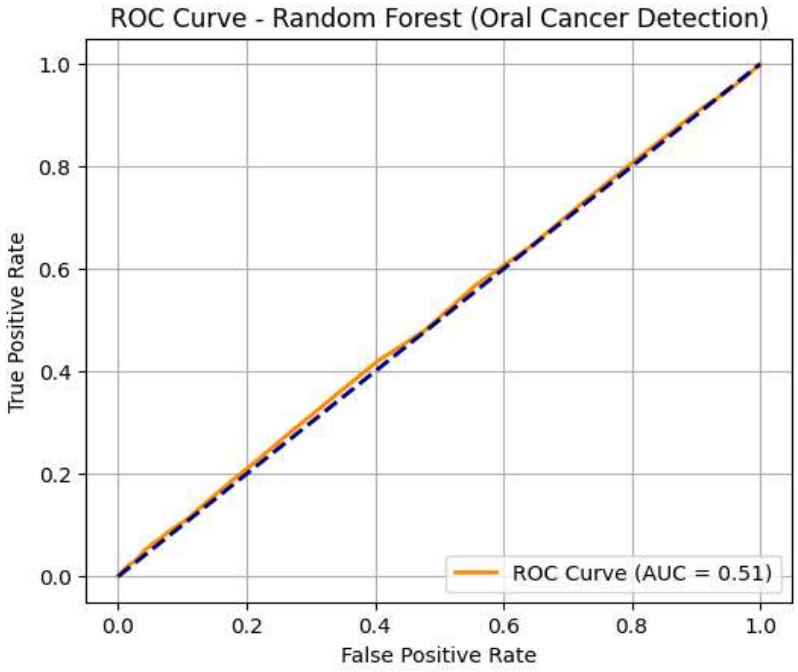

Figure 9: ROC Curve for Random Forest Method

5. Conclusion

In this study, four machine learning models, Logistic Regression, XGBoost, CatBoost, and Random Forest, were evaluated for oral cancer detection. All models showed high overall accuracy due to the dataset's class imbalance, but failed to effectively identify cancer-positive cases. Logistic Regression, XGBoost, and CatBoost each predicted all instances as negative, with AUC scores near 0.50, indicating no real predictive power. However, Random Forest achieved a slightly better AUC of 0.51 and was the only model to detect at least one true positive. This shows that it holds potential for improvement with proper data balancing. Overall, Random Forest emerges as the most promising model for oral cancer prediction among those tested.

Reference

[1] Al-Rawi, N., Sultan, A., Rajai, B., Shuaeeb, H., Alnajjar, M., Alketbi, M., ... & Mashrah, M. A. (2022). The effectiveness of artificial intelligence in detection of oral cancer. *international dental journal*, *72*(4), 436-447.

[2] Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., Jantana, P., & Vicharueang, S. (2022). AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. *Plos one*, *17*(8), e0273508.

[3] Mira, E. S., Saaduddin Sapri, A. M., Aljehanı, R. F., Jambı, B. S., Bashir, T., El-Kenawy, E. S. M., & Saber, M. (2024). Early diagnosis of oral cancer using image processing and Artificial intelligence. *Fusion: Practice & Applications*, *14*(1).

[4] A. Jemal et al., Annual report to the nation on the status of cancer, featuring survival, JNCI-J. Natl. Cancer Inst. 109(9), 1975-2014, 2017

[5] G. Yardimci et al., Precancerous lesions of oral mucosa, World J. Clin. Cases, 2(12), 866-872, 2014.

[6] Ilhan B, Lin K, Guneri P, Wilder-Smith P. Improving Oral Cancer Outcomes with Imaging and Artificial Intelligence. *Journal of Dental Research*. 2020;99(3):241-248. doi:10.1177/0022034520902128

[7] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.

[9] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[10] Lantz, B. (2019). *Machine Learning with R: Expert techniques for predictive modeling* (3rd ed.). Packt Publishing

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[12] Brownlee, J. (2016). *XGBoost with Python: Gradient boosting for regression and classification*. Machine Learning Mastery.

[13] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.