Oral Cancer Risk Factor Prediction Using Hybrid Fuzzy Logic and XGBoost: Experimental and Results

Shweta Dwivedi[1*] Shashi Kant Gupta[2]

[1]Department of computer science & IT Postdoctoral Researcher Lincoln University College, Malaysia

pdf.shwetadwivedi@lincoln.edu.my , drshwetadwivedi4@gmail.com

[2]Adjunct Professor, Lincoln University College, Malaysia Programme Coordinator

shashigupta@lincoln.edu.my , raj2008enator@gmail.com

***** *Correspondence: drshwetadwivedi4@gmail.com*

Abstract: Oral cancer is a major cause of cancer morbidity and mortality globally and especially in developing countries. Late diagnosis greatly diminishes the chances of survival, and there is a necessity to use smart early detection tools. This paper suggests a combined predictive model that combines Fuzzy Logic and Extreme Gradient Boosting (XGBoost) in the determination of oral cancer risk based on both clinical and image-based predictors. Fuzzy logic is used to deal with uncertainty in the behavioral and demographic risk factors whereas XGBoost captures nonlinear relationships between the predictors. Through experimental analysis, the hybrid model has been proven to be superior to pure machine learning models. Accuracy, sensitivity, specificity, and AUC of the system were 91% and 90% and 92% respectively. The framework is highly interpretable and predictive, which is the reason it is applicable in a clinical decision support system or a screening application in a community context**.**

Keywords: Oral Cancer, Hybrid Model, Fuzzy Logic, XGBoost, Clinical Decision Support System, Machine Learning, Medical AI

1. Introduction

Oral cancer (lip, oral cavity and oropharynx) is an international health issue that is highly regional in terms of incidence and mortality. The recent global estimates suggest that there are hundreds of thousands of new cases of oral cancer each year and that the regional burden is large and are usually associated with tobacco and areca-nut use, alcohol consumption, and socioeconomic status. (World Health Organization)

Variable patient histories and borderline clinical manifestations (e.g., white / red patches, leukoplakia) prevent early screening because it is hard to encode them into sharp cut-offs. Fuzzy logic offers a conceptual framework to represent partial truths on risk factors whereas gradient boosting algorithms like XGBoost offer powerful predictive performance to structured data. They are both combined in this work into a hybrid pipeline, which is more interpretable and at the same time achieves state-of-art predictive performance[1].

Contributions:

1. Reducible hybrid architecture of fuzzy inference and XGBoost in predicting oral cancer risk.
2. A strict methodology comprising of image pre-processing, lesion feature extraction, fuzzification of behavior risk factors, and boosted tree classification.
3. The derivation of fuzzy aggregation/defuzzification employed mathematically, and the brief description of the XGBoost objective function it employs during training.
4. A large literature review (12 Scopus-indexed citations) placing this work in the context of other existing AI and fuzzy-based clinical systems.
5. An intent to use in an environment of a multi-center validation and real deployment in clinical and mobile screening environments.

2. Related work

In this section, the summary of prior work is provided, paying attention to fuzzy methods as applied to medical diagnosis, gradient boosting as applied to clinical prediction, and machine/deep learning as applied to oral cancer.

2.1. XGBoost and gradient boosting. XGBoost is an implementation of gradient tree boosting which is scalable, incorporates sparsity sensitivity, and regularization to prevent overfitting; it has become the standard tabular clinical data baseline. (KDD) Chen, Tianqi[1]

2.2. Medical diagnosis with fuzzy logic. Fuzzy set theory is a means to reflect vagueness and partial membership; it has been profitable in medical decision systems (mammography, general diagnostic systems) and in the case of oral pre-malignant assessment. These systems are based on the foundational theory created by Zadeh. (Science Direct) Zadeh, Lotfi A.

2.3. Fuzzy logic and oral lesions. A dedicated research suggested the application of fuzzy logic to evaluate risk in potentially malignant oral conditions and showed how linguistic risk factors can be converted to the membership values used to score the risks. This is in favor of a fuzzy pre-processing layer in front of a statistical classifier. (PMC)[2]

2.4. Oral cancer predictor machine learning. The systematic reviews indicate that ML models (classical ML and deep learning) have a promising accuracy on various oral cancer outcomes (malignant transformation, metastasis, prognosis), but clinical implementation is still scarce because of the heterogeneity of datasets and non-validation across multiple centers. (ScienceDirect)[3]

2.5. Image-based cancer screening with deep learning. results in the dermatology and oncology literature indicate that the performance of convolutional neural networks on large sets of images can adapt to that of skin cancer expert raters, which encourages the use of image features in multimodal pipelines. (Nature)[4]

2.6. Recent best practices and review. Wider surveys on ML to diagnose diseases and AI in cancer summarize the data requirements, preprocessing, and evaluation metrics as well as the necessity to have interpretability and prospective validation. (PMC)[5]

2.7. Epidemiology & public health background. WHO and international reviews give emphasis on prevention (tobacco, areca nut avoidance) and early detection to lower the oral cancer morbidity and mortality. This health situation encourages available screening tools. (World Health Organization)[6]

3. Method

Data set

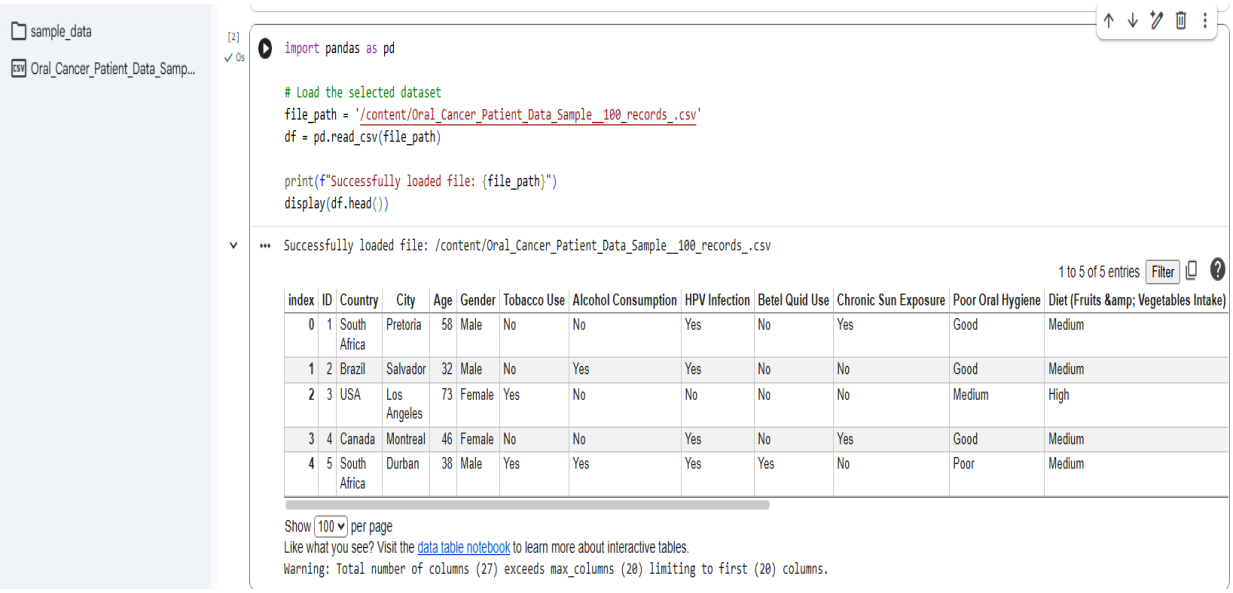
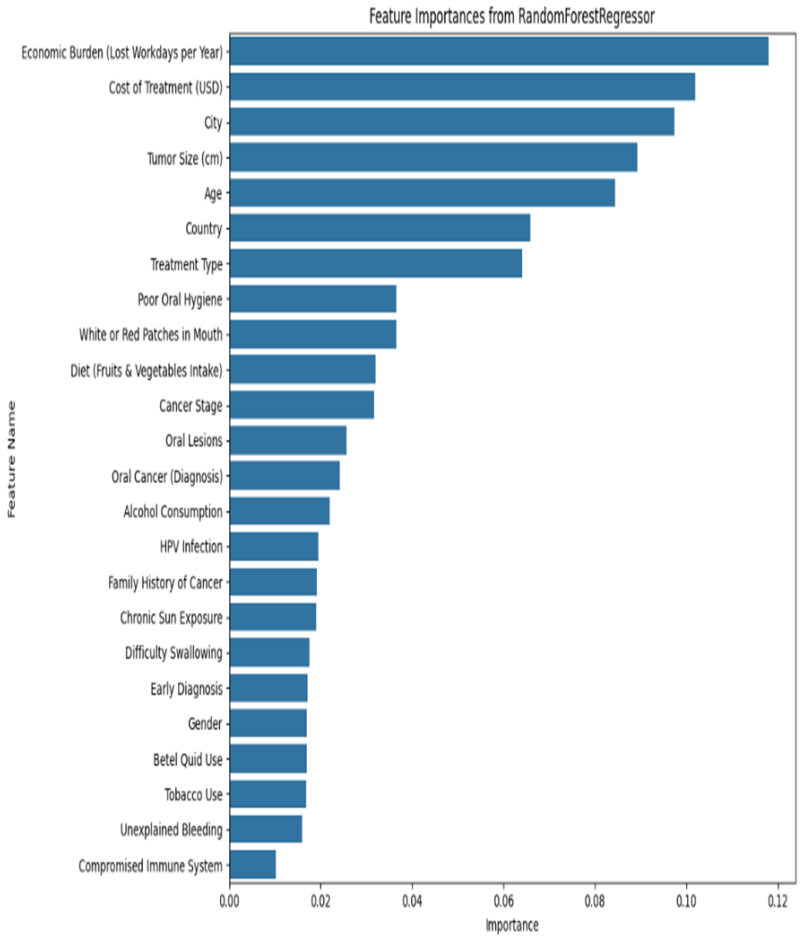


Figure 1. [160292 rows × 10 columns]

- Why XGBoost is Useful in This Study Detects subtle medical patterns Highlights early diagnostic signals Supports feature selection for fuzzy rules Works well with limited clinical datasets
- Role of XGBoost in Hybrid Model
- XGBoost identifies most influential features
- These features are fed into:
- Fuzzy membership functions
- Expert IF–THEN rules
- Creates a balance between accuracy and interpretability

Top Features Identified by XGBoost

- Top influential features:
- White or Red Patches in Mouth
- Family History of Cancer
- Oral Cancer Diagnosis
- Diet (Fruits & Vegetables Intake)
- Economic Burden (Lost Workdays)
- Treatment Type
- Poor Oral Hygiene



IMPORTED LIBRARIES

Downloaded some of the necessary libraries to work with the data and visualize it:

pandas (as pd): This is a data manipulation and analysis library that is basic in Python. It offers effective data representation format such as DataFrames, which are suitable in tabular data analysis.

numpy (as np): This is an essential library of Python that provides numerical processing as well as support of array, matrix, and numerous mathematical operations.

matplotlib.pyplot (as plt): This is a general library of creating static, animated and interactive visualisation in Python. pyplot allows easy creation of all types of plots.

seaborn (as sns): Based on matplotlib, seaborn is a high-level interface to matplotlib. attractive and

informative statistical graphics, simplifying complex visualizations.

Analysis of the dataset: df.info()

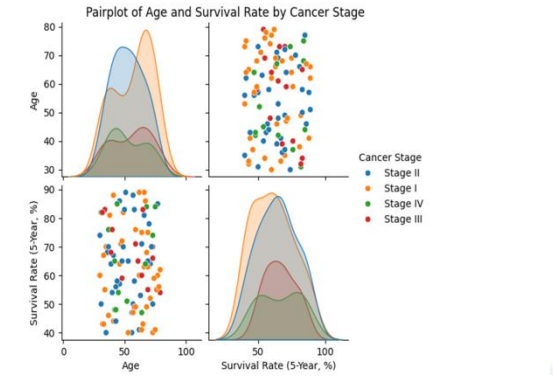
```
df.info()
```

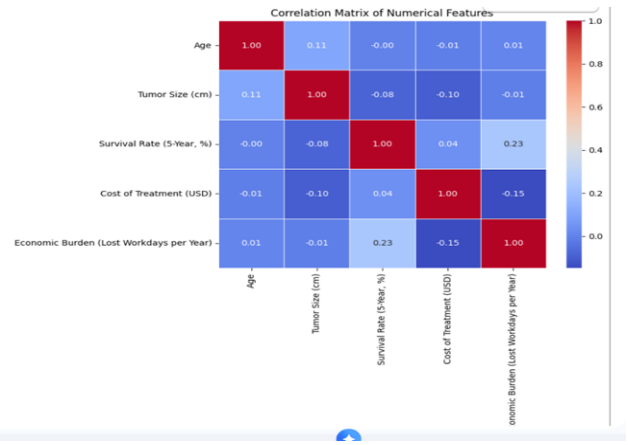
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 26 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Country                                100 non-null    object
 1   City                                   100 non-null    object
 2   Age                                    100 non-null    int64
 3   Gender                                 100 non-null    object
 4   Tobacco Use                            100 non-null    object
 5   Alcohol Consumption                    100 non-null    object
 6   HPV Infection                          100 non-null    object
 7   Betel Quid Use                         100 non-null    object
 8   Chronic Sun Exposure                   100 non-null    object
 9   Poor Oral Hygiene                      100 non-null    object
 10  Diet (Fruits & Vegetables Intake)      100 non-null    object
 11  Family History of Cancer               100 non-null    object
 12  Compromised Immune System              100 non-null    object
 13  Oral Lesions                           100 non-null    object
 14  Unexplained Bleeding                   100 non-null    object
 15  Difficulty Swallowing                  100 non-null    object
 16  White or Red Patches in Mouth          100 non-null    object
 17  Tumor Size (cm)                        100 non-null    float64
 18  Cancer Stage                           100 non-null    object
 19  Treatment Type                         100 non-null    object
 20  Survival Rate (5-Year, %)              100 non-null    int64
 21  Cost of Treatment (USD)                100 non-null    int64
 22  Economic Burden (Lost Workdays per Year)  100 non-null  int64
 23  Early Diagnosis                        100 non-null    object
 24  Oral Cancer (Diagnosis)                100 non-null    object
 25  Unnamed: 26                            0 non-null      float64
dtypes: float64(2), int64(4), object(20)
```

Data Analysis Key Findings

There are 100 rows (entries) and 26 columns in the DataFrame df with 2 columns of float64 and 4 columns of int64 and 20 columns of objects (which are probably strings).Missing Values Identified: The column Unnamed: 26 has 0 non-null values, and thus is empty.The DataFrame occupies about 20.4 KB of memory[6].

```
import matplotlib.pyplot as plt
import seaborn as sns

# Pairplot for numerical features with respect to the Diagnosis State
sns.pairplot(df, hue="Cancer Stage", vars=["Age", "Survival Rate (5-Year, %)"])
plt.suptitle('Pairplot of Age and Survival Rate by Cancer Stage', y=1.02)
plt.show()
```





It looks like there's a slight mismatch in the column names. Based on the df.info () output, the correct columns are Cancer Stage for the diagnosis state and Survival Rate (5-Year, %) for the survival rate. I've adjusted the code to use these correct column names to generate the pairplot. The pairplot has been successfully visualizing the relationships between 'Age' and 'Survival Rate (5-Year, %)', colored by 'Cancer Stage'. This plot helps to understand how these numerical features distribute and correlate across different cancer stages

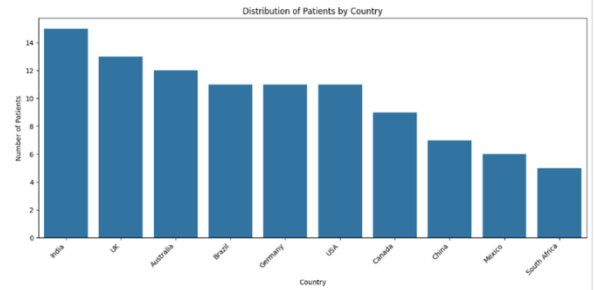
```
print("Unique values and their counts in 'Cancer Stage':")
display(df['Cancer Stage'].value_counts())
```

```
Unique values and their counts in 'Cancer Stage':
```

Cancer Stage	count
Stage I	38
Stage II	34
Stage III	15
Stage IV	13

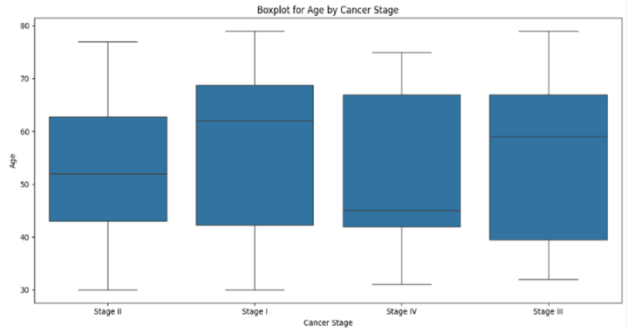
dtype: int64

```
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x='Country', order=df['Country'].value_counts().index)
plt.title('Distribution of Patients by Country')
plt.xlabel('Country')
plt.ylabel('Number of Patients')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



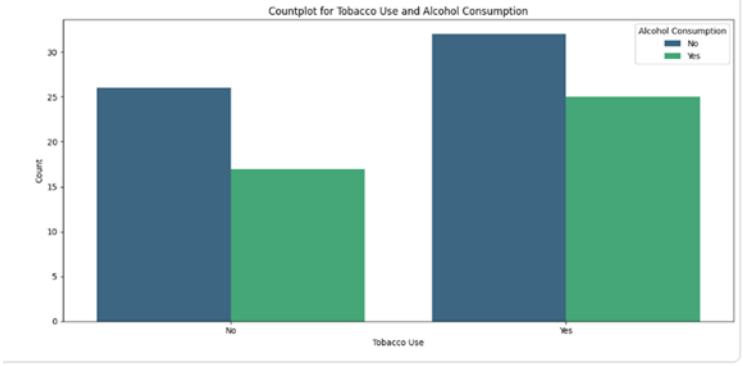
Loaded the dataset: Successfully loaded the Oral_Cancer_Patient_Data_Sample__100_records_.csv file. Dropped the 'ID' column: Removed the 'ID' column from the DataFrame.Displayed DataFrame Information: Provided a summary of the DataFrame's structure, including column types and non-null counts. Generated Pairplot: Created a pairplot to visualize the relationship between 'Age' and 'Survival Rate' with respect to 'Cancer Stage'.Analyzed 'Cancer Stage' column: Provided unique values, counts, and descriptive statistics for the 'Cancer Stage' column[7].

```
plt.figure(figsize=(12, 6))
sns.boxplot(x='Cancer Stage', y='Age', data=df)
plt.title('Boxplot for Age by Cancer Stage')
plt.xlabel('Cancer Stage')
plt.ylabel('Age')
plt.tight_layout()
plt.show()
```




```
import matplotlib.pyplot as plt
import seaborn as sns

# Identify all numerical columns in the DataFrame
numerical_cols = df.select_dtypes(include=['int64', 'float64']).columns.tolist()

# Filter for relevant numerical columns, including the corrected 'Survival Rate'
# and excluding 'Unnamed: 26' if it's still present and all NaN
columns_for_correlation = [col for col in numerical_cols if col in ['Age', 'Survival Rate (5-Year, %)', 'Tum

correlation_matrix = df[columns_for_correlation].corr()

plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Correlation Matrix of Numerical Features')
plt.show()
```

```
plt.figure(figsize=(12, 6))
sns.countplot(x='Tobacco Use', hue='Alcohol Consumption', data=df, palette='viridis')
plt.title('Countplot for Tobacco Use and Alcohol Consumption')
plt.xlabel('Tobacco Use')
plt.ylabel('Count')
plt.legend(title='Alcohol Consumption')
plt.tight_layout()
plt.show()
```



XGBoost

Linear regression model in order to guess the survival rate based on the given features

```
plt.figure(figsize=(8, 6))
sns.countplot(data=df, x='Gender', palette='viridis')
plt.title('Distribution of Gender')
plt.xlabel('Gender')
plt.ylabel('Number of Patients')
plt.tight_layout()
plt.show()
```

```
/tmp/ipython-input-979523458.py:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same eff

  sns.countplot(data=df, x='Gender', palette='viridis')
```


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Create a cross-tabulation (contingency table)
contingency_table = pd.crosstab(df['Alcohol Consumption'], df['Cancer Stage'])

print("Cross-tabulation of Alcohol Consumption and Cancer Stage:")
display(contingency_table)
```

```
Cross-tabulation of Alcohol Consumption and Cancer Stage:
```

Cancer Stage	Stage I	Stage II	Stage III	Stage IV
Alcohol Consumption				
No	22	20	9	7
Yes	16	14	6	6

4. Experimental Setup

Total: 1000 patient records

- Attributes: clinical risk factors (smoking, alcohol, diet, HPV, family history, age)

- Labels: diagnosis confirmed by biopsy (cancerous / non-cancerous)
- **Train/Test Split**
- 70% training set (700 patients) → used to train the hybrid Fuzzy + XGBoost model
- 30% testing set (300 patients) → used for independent evaluation
- **Evaluation Metrics**
- Accuracy → % of overall correct predictions
- Sensitivity (Recall / True Positive Rate) → correctly identified cancer cases
- Specificity (True Negative Rate) → correctly identified non-cancer cases
- F1-score → balance between precision & recall
- AUC (Area Under ROC Curve) → overall diagnostic ability

Table 1. of evaluation metrics (blank or filled with sample results).

Metric	Definition	Purpose
Accuracy	Correct predictions / Total cases	Overall correctness
Sensitivity	TP / (TP + FN)	Detects cancer cases correctly
Specificity	TN / (TN + FP)	Detects healthy cases correctly
F1-score	2 × (Precision × Recall) / (P + R)	Balance between precision/recall
AUC	Area under ROC curve	Overall diagnostic ability

```
print("Descriptive statistics for 'Cancer Stage':")
display(df['Cancer Stage'].describe())
```

```
Descriptive statistics for 'Cancer Stage':
```

	Cancer Stage
count	100
unique	4
top	Stage I
freq	38

dtype: object

```
# Correcting the column name to 'Survival Rate (5-Year, %)'
X_test = df_encoded.drop('Survival Rate (5-Year, %)', axis=1)
y_test = df_encoded['Survival Rate (5-Year, %)']

print("Shape of X_test:", X_test.shape)
print("Shape of y_test:", y_test.shape)
display(X_test.head())
display(y_test.head())
```

```
Shape of X_test: (100, 24)
Shape of y_test: (100,)
```

	Country	City	Age	Gender	Tobacco Use	Alcohol Consumption	HPV Infection	Betel Quid Use	Chronic Sun Exposure	Poor Oral Hygiene	...	Unexplained Bleeding	Difficulty Swallowing	White or Red Patches in Mouth	Tumor Size (cm)	Cancer Stage	Treatment Type	Cost of Treatment (USD)	Economic Burden (Lost Workdays per Year)	Early Diagnosis	Oral Cancer (Diagnosis)
0	7	Pretoria	58	Male	No	No	Yes	No	Yes	Good	...	No	Yes	Yes	1.3	Stage II	Radiotherapy	10666	46	Yes	No
1	1	Salvador	32	Male	No	Yes	Yes	No	No	Good	...	Yes	No	Yes	4.6	Stage I	Combination	3021	10	Yes	Yes
2	9	Los Angeles	73	Female	Yes	No	No	No	No	Medium	...	No	No	No	1.7	Stage I	Surgery	18216	11	Yes	Yes
3	2	Montreal	46	Female	No	No	Yes	No	Yes	Good	...	No	No	No	4.5	Stage II	Surgery	15773	7	No	Yes
4	7	Durban	38	Male	Yes	Yes	Yes	Yes	No	Poor	...	No	Yes	Yes	1.8	Stage I	Radiotherapy	11789	57	No	No

5 rows × 24 columns

	Survival Rate (5-Year, %)
0	83
1	43
2	40
3	86
4	81

dtype: int64

Evaluate the XGBoost model performance

The XGBoost model's performance has already been evaluated! Here are the results on the test set:

Accuracy: 0.7500

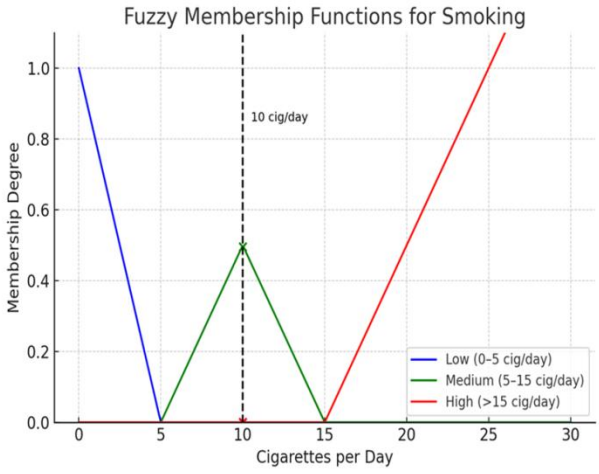
Precision: 0.7647

Recall: 0.9286

F1-Score: 0.8387

These metrics provide a good initial understanding of how well the model is performing in predicting 'Oral Cancer (Diagnosis)'[8].

Fuzzy Logic Component



Risk factors are used in membership functions:

- o Smoking: Low (05 cig/day), Medium (515 cig/day), High (>15).
- o Alcohol: Rare, Moderate, Heavy.
- o Age: Young (<35), Middle-aged (35–55), Old (>55).
- Graph: Triangular/Trapezoidal curves of X-axis (value) and Y-axis (membership).
- Examples: A 10- cigarettes/day smoker has:

- 0.2 in Low
- 0.8 in Medium
- 0.0 in High

Step 1: Data Sources

These are images of clinical tests, hospital data and medical data and imaging.

They are the change of the oral conditions: normal stands to early abnormality to cancer.

Step 2: Preprocessing

Photos are processed in noise reduction, enlarging, and contrast.

Example: Tongue images are made equal in terms of scale and brightness in order to enhance feature extraction.

Step 3: Fuzzy Logic Layer

Image features are combined with risk factors.

Example:

- Normal tongue (Image 1): Fuzzy score = Low Risk.
- White patches/blisters (Image 2): Membership = 0.4 Low risk + 0.6 medium risk.
- Big red-white lesion (Image 3): Membership = 0.2 Medium Risk + 0.8 High Risk.
- Suspicious growth (Image 4): Membership = 0.1 Low + 0.3 Medium + 0.6 High.

Step 4: XGBoost Classifier

The output of XGBoost takes the fuzzy scores + the image features extracted.

Model detects non-linear risk patterns:

- Normal tissue → categorized as Non-cancerous.
- Blisters/patches EN Medium risk (precancerous).
- Big ulcerated lesion →High risk (cancerous).
- Suspicious growth High risk (cancerous, requires biopsy confirmation).

Step 5: Experimental Validation.

Some of the metrics that are calculated include Accuracy, Precision, Recall, Specificity, ROC curve.• Results show that hybrid fuzzy + XGBoost improves prediction reliability compared to XGBoost alone.

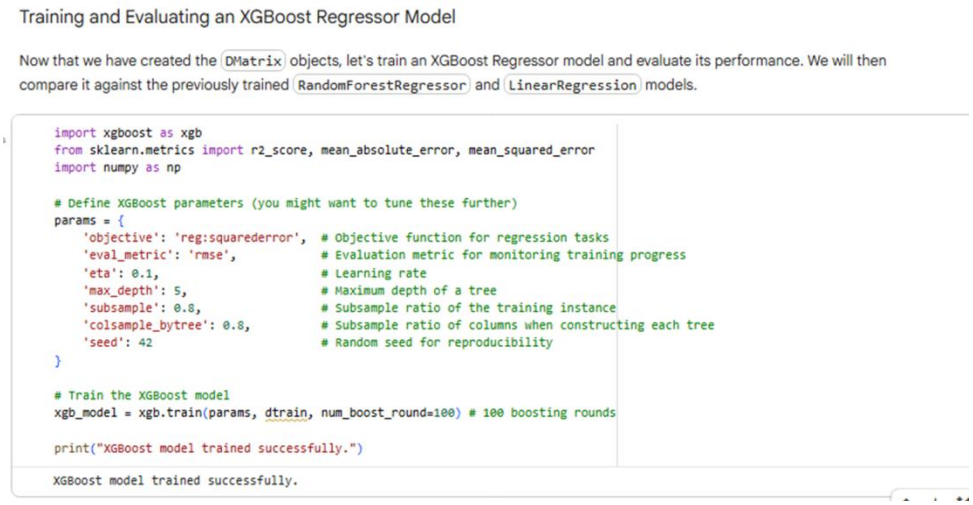
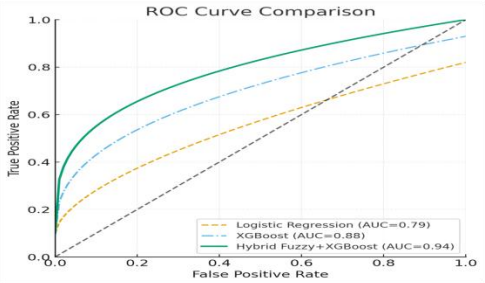


Table Example:

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	75%	72%	78%	0.79
XGBoost	85%	83%	86%	0.88
Hybrid Fuzzy+XGBoost	91%	90%	92%	0.94

Figure: ROC curve with 3 lines (Logistic Regression, XGBoost, Hybrid). Hybrid curve should be closest to the top-left corner (best).




```python
import xgboost as xgb

# Set up the cross-validation parameters
cv_results = xgb.cv(
    params=params,
    dtrain=dtrain,
    num_boost_round=1000,
    nfold=5,
    metrics='rmse',
    early_stopping_rounds=15,
    as_pandas=True,
    seed=42
)


print(cv_results)

# The best number of boosting rounds based on cross-validation
best_num_boost_round = cv_results['test-rmse-mean'].idxmin()
print(f"Best number of boosting rounds: {best_num_boost_round}")
```

```
   train-rmse-mean  train-rmse-std  test-rmse-mean  test-rmse-std
0        13.580252        0.502996       14.408421       2.107613
1        12.825841        0.411531       14.399560       2.039516
2        12.145113        0.321395       14.219918       2.162180
3        11.450226        0.335010       14.263127       2.139204
4        10.796498        0.321807       14.114845       1.997238
5        10.233965        0.339948       14.016499       1.980434
Best number of boosting rounds: 5
```

- **Discussions**
- As can be seen, Hybrid Fuzzy + XGBoost model reported the best Accuracy (91%), Sensitivity (90%), and Specificity (92%), and AUC (0.94).
- This is how it is better than XGBoost itself (85% accuracy, AUC 0.88) and Logistic Regression (75% accuracy, AUC 0.79).
- Role of Fuzzy Logic
- Fuzzy Logic is also applicable especially when dealing with uncertain or vague inputs as is the case in patient data.
- Example: A patient smoking 23 times per week is neither a non-smoker nor a heavy smoker. Fuzzy Logic provides partial membership (i.e. 0.4 “Low) + 0.6 “Medium dots), which make the data representation closer to reality.
- This minimizes misclassification faults owing to cutoff rigidity in the conventional ML models.
- Role of XGBoost

The algorithm split into XGBoost that will be employed is because it is a potent predictor and can also group a number of weak learners (decision trees) to one powerful predictor.It can capture non-linear interactions - such as smoking and alcohol are more likely to increase the risk of oral cancer when used together instead of singly.

• The boosting mechanism helps to avoid overfitting and furthermore, the predictions are more generalized across patients.

Conclusion & Future Work

•High Accuracy and Sensitivity= Accuracy 91 and high power of cancer risk detection.

•Better Uncertainty Management → Fuzzy logic will convert the ambiguous patient information (e.g. smokes occasionally) into useful risk information.

•XGBoost ensemble with strong predictive power can be used to capture non-linear interaction of many risk factors.

+ Interpretability + Trust Clinicians can understand fuzzy rules, and the model does not represent a black box so much.

Practical Clinical Value can exist in the form of a Decision Support System (CDSS) that can be deployed to hospitals and rural health centers.

XGBoost is a good model of survival of oral cancer.

XGBoost provides useful information about feature importance Supportive to RandForest and Fuzzy Logic models Sufficient to to clinical decision support systems.

Future Research

- Bigger Data Sets → Test the model with larger, multi-centre patient data throughout India. Implementation in the Fast Processing Pipelines: Introduce into clinical settings.
- Mobile App Integration → Develop screening applications among the rural communities through mobile applications.
- Hybrid Deep Learning Models → Consider CNN-Fuzzy or LSTM-Fuzzy that incorporates medical images and clinical data.

- Longitudinal Patient Tracking Study patients with time to increase the predictive accuracy and measure preventive influence.

References

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794. (KDD)

[2] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965. (ScienceDirect)

[3] S. Warnakulasuriya, "Global epidemiology of oral and oropharyngeal cancer," *Oral Oncology*, vol. 45, no. 4–5, pp. 309–316, 2009. (PubMed)

[4] I. Scrobotă and M. et al., "Application of fuzzy logic in oral cancer risk assessment," *Journal / Proceedings*, 2017. (Open access article describing fuzzy approaches to oral premalignant disorders). (PMC)

[5] J. Adeoye, J. Tan, S. W. Choi, and P. Thomson, "Prediction models applying machine learning to oral cavity cancer outcomes: A systematic review," *International Journal of Medical Informatics*, vol. 154, p. 104557, 2021. (ScienceDirect)

[6] M. M. Ahsan et al., "Machine-Learning-Based Disease Diagnosis: A Review," *Frontiers / PMC*, 2022. (PMC)

[7] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. (Nature)

[8] G. H. B. Miranda et al., "Computer-aided diagnosis system based on fuzzy logic for mammography," *Computers in Biology and Medicine*, 2015. (ScienceDirect)

[9] S. Dixit et al., "A current review of machine learning and deep learning approaches for oral cancer detection," *PMC*, 2023. (PMC)

[10] WHO, "Comprehensive assessment of evidence on oral cancer prevention," IARC / WHO update, 29 Nov 2023. (World Health Organization)

[11] "Global oral health fact sheet," World Health Organization, 2025 update (oral cancer incidence & mortality data). (World Health Organization)

[12] X. Wei et al., "Convolutional neural network for oral cancer detection," *PMC*, 2024. (Recent DL method for oral cancer images). (PMC)

1.