Content lists available at LiteraPublishing
Journal of Public Health and Community Systems
NEXURA
journal homepage : https://journal.literapublishing.id/index.php/nexura

Article

Classification of Oral Cancer Using the Random Forest Algorithm

Zaskia Putri Ratnaning Pratiwi[1*], Alviani Putri Brilianti Ajie[2], Shafira Maharani Ayunindya[3], M. Ibnu Sa'ad[4], and Mudafiq Riyan Pratama[5]

[1] Departement of Health Information Management, Politeknik Negeri Jember, Jember, Indonesia; g41241272@student.polije.ac.id
[2] Departement of Health Information Management, Politeknik Negeri Jember, Jember, Indonesia; g41241115@student.polije.ac.id
[3] Departement of Health Information Management, Politeknik Negeri Jember, Jember, Indonesia; g41241223@student.polije.ac.id
[4] Departement of Health Information Management, Politeknik Negeri Jember, Jember, Indonesia; g41241249@student.polije.ac.id
[5] Departement of Health Information Management, Politeknik Negeri Jember, Jember, Indonesia; mudafiq.riyan@polije.ac.id

* Correspondence: g41241272@student.polije.ac.id

Abstract: Oral cancer ranks as the sixth most common cancer worldwide and can be prevented if detected early. The lack of universally available, reliable screening methods has led to the use of machine learning as a promising alternative approach. This study aims to develop an oral cancer classification model using a Random Forest algorithm optimized with GridSearchCV, and to implement it into a Streamlit-based web application. The data is sourced from the Oral Cancer Prediction Dataset on Kaggle, using 1,500 samples from a total of 84,922 data points. Preprocessing included removing irrelevant columns, data cleaning, encoding categorical variables, and normalizing numerical features, resulting in 17 predictor features. The dataset was stratified into training and testing sets at a 70:30 ratio. GridSearchCV optimization yielded the best hyperparameters: n_estimators = 200, max_depth = 10, min_samples_split = 10, and min_samples_leaf = 1. The model achieved an accuracy of 79.91%, a precision of 85.8%, a recall of 68.72%, and an F1-Score of 76.32%. Feature importance analysis shows that the most influential variables, in order, are Treatment Type, Age, Diet, HPV Infection, Oral Hygiene, Alcohol Consumption, Gender, Unexplained Bleeding, Difficulty Swallowing, and Betel Nut Use. The model was successfully integrated into a Streamlit application that displays real-time predictions along with probability values and follow-up recommendations. This system can support early screening for oral cancer in primary healthcare facilities. Further research is recommended using the full dataset and exploring other algorithms to improve performance, particularly the recall value to minimize false negatives.

Keywords: Classification, GridSearchCV, Oral Cancer, Random Forest, Streamlit

Citation: Z. P. R. Pratiwi, A. P. B. Ajie, S. M. Ayunindya, M. I. Sa'ad, and M. R. Pratama, "Classification of Oral Cancer Using the Random Forest Algorithm", *nexura*, vol. 1, no. 1, pp. 31–49, Jun. 2026.

Received: 27-05-2026
Accepted: 28-06-2026
Published: 30-06-2026

1. Introduction

Cancer is a cellular disease characterized by uncontrolled cell growth, followed by invasion of surrounding tissues and spread, or metastasis, to other parts of the body. Almost all cases of cancer are caused by mutations or the activation of abnormal cellular genes that play a role in controlling cell growth and division. One of the most commonly found types of cancer is oral cancer, which ranks sixth among the most common cancers and is also one of the leading causes of death worldwide.

Oral cancer is, in fact, a type of malignancy that can be prevented if detected early through the recognition of clinical signs within the oral cavity [22]. Anatomically, this cancer is characterized by the presence of a malignant tumor developing on the lips or within the oral cavity, including the tongue, cheek mucosa, gums, floor of the mouth, and

palate [5]. The majority of cases, approximately 80–90%, are diagnosed as oral squamous cell carcinoma (OSCC) [21].

This type of cancer is more commonly found in developing countries and generally affects adult men with a history of tobacco smoking, betel nut chewing, or alcohol consumption [2]. To date, the exact cause of oral cancer remains unknown, given that its occurrence is multifactorial and complex. Nevertheless, several factors are known to play a role in its development, namely local factors, external factors, and internal factors. Local factors include poor oral hygiene, chronic irritation due to inadequate restorations, and dental caries. External factors that contribute include smoking, alcohol consumption, betel nut chewing, and certain viral infections. In addition, factors originating within the body, or host factors, are equally important; these include age, gender, nutritional status, immune system condition, and genetic factors.

The complexity of these risk factors makes the early detection of oral cancer a unique challenge in clinical practice. To date, no screening method has been universally recognized as truly accurate and reliable, underscoring the need for new approaches that can identify risks more accurately and efficiently. As technology advances, machine learning has emerged as a promising approach in the medical field, particularly in the classification and detection of diseases based on clinical data [5]. Various machine learning algorithms have been tested for diagnostic purposes. Still, not all are capable of simultaneously addressing the challenges frequently encountered in medical data, such as high-dimensional data, missing values, and class imbalance.

In this study, the method used to address this problem is the random forest algorithm. This algorithm is one of the most widely used supervised learning methods because it is capable of generating accurate predictions while efficiently identifying important features in the dataset. Technically, the random forest works by building a large number of decision trees in parallel using the bootstrap aggregating (bagging) technique, then combining the prediction results from all the trees through a majority voting mechanism to produce a final decision.

In the context of oral cancer classification, this approach is highly relevant because patient clinical data is generally multivariate, encompassing various variables such as ID (identity), country of origin, age, and gender, as well as risk factors like tobacco use, alcohol consumption, HPV infection, betel nut use, and long-term sun exposure. Additionally, there is information on health conditions such as poor oral hygiene, dietary patterns particularly fruit and vegetable intake family history of cancer, and immune system status. The data also records symptoms experienced, such as oral lesions, unexplained bleeding, difficulty swallowing, and the appearance of white or red patches inside the mouth. Furthermore, medical information is included regarding tumor size, cancer stage, type of treatment received, and 5-year survival rates. Additionally, economic aspects are documented, such as treatment costs and the economic burden represented by the number of workdays lost per year. Finally, information is provided on whether the diagnosis was made at an early stage and the outcomes of oral cancer diagnosis.

The ensemble approach in random forests has proven effective in reducing the risk of overfitting, a common issue in single decision tree models, while improving the model's stability and generalization to new data. Furthermore, random forests also provide a measure of feature importance that allows for the identification of the most influential clinical variables in the oral cancer risk classification process, ensuring that the model's results are not only accurate but also medically interpretable and clinically useful for decision-making.

However, the performance of a random forest is highly dependent on its hyperparameter configuration. Therefore, to obtain optimal classification results, this study also employs GridSearchCV, a systematic search technique for identifying the best

parameter combinations through cross-validation, with the aim of optimizing model performance. This process involves tuning several key hyperparameters, such as the number of trees (n_estimators), the maximum tree depth (max_depth), the number of features considered at each split (max_features), and the splitting criterion (criterion). By combining the strengths of random forests and GridSearchCV optimization, the resulting model is expected not only to have high accuracy but also to generalize well to previously unseen patient data [5].

However, an accurate model alone is not enough if it cannot be accessed and used practically by healthcare workers in the field. To address this need, this study implemented the predictive model into a web application based on Streamlit, a Python framework that enables the creation of interactive interfaces in real time. This application is designed to be accessible to both healthcare workers and non-technical users, featuring an intuitive and user-friendly interface. Thus, this system is expected to bridge the gap between algorithmic complexity and practical field needs, particularly in the process of screening and early detection of oral cancer in primary healthcare facilities.

Based on the above description, this study aims to develop and evaluate an oral cancer classification model using a random forest algorithm optimized with GridSearchCV, while implementing it into the Streamlit platform so that oral cancer risk prediction results are more easily accessible, understandable, and usable in the context of practical and efficient early screening. Through this approach, the study is expected to make a tangible contribution to efforts to reduce the incidence and mortality rates of oral cancer in Indonesia.

2. Materials and Methods

This study employs a comparative analysis approach to identify the most optimal algorithm between Random Forest and Logistic Regression for predicting oral cancer based on key risk factors, symptoms, cancer staging, survival rates, treatment approaches, and economic burden to facilitate research and prediction modeling. The data used were obtained from a secondary source, specifically an oral cancer prediction dataset retrieved from the Kaggle website.

2.1 Data Collection

The research data was obtained from the open-source platform Kaggle, specifically from the dataset titled "Oral Cancer Prediction Dataset." The original dataset consists of 84,922 samples and 25 columns (19). Due to computational limitations, this study used only 1,500 samples randomly selected from the original dataset. Of the 25 available columns, several were removed, namely ID, Country, Tumor Size (cm), Cancer Stage, Survival Rate (5-Year, %), Cost of Treatment (USD), and Economic Burden (Lost Workdays per Year). These columns were removed because the information is only known after the diagnosis is confirmed, so it cannot be used as a predictive feature (5). The complete list of all attributes in the dataset along with their status is shown in Table 1 below.

Table 1. Dataset Attributes

Fitur Name	Description	Status
ID	Unique identification number for each patient	Deleted
COUNTRY	Patient's country of origin	Deleted
AGE	Patient's age in years	Feature
GENDER	Patient's gender (Male or Female)	Feature
TOBACCO_USE	Status of tobacco use habit	Feature
ALCOHOL_CONSUMPTION	Alcohol consumption habit	Feature
HPV_INFECTION	Human Papillomavirus (HPV) infection status	Feature
BETEL_QUID_USE	Habit of chewing betel/areca nut	Feature
CHRONIC_SUN_EXPOSURE	Chronic sun exposure	Feature
POOR_ORAL_HYGIENE	Poor oral hygiene condition	Feature
DIET_FRUITS_VEGETABLES	Level of fruit and vegetable consumption in the patient's diet	Feature
FAMILY_HISTORY_OF_CANCER	Family history of cancer disease	Feature
COMPROMISED_IMMUNE_SYSTEM	Weakened or impaired immune system condition	Feature
ORAL_LESIONS	Presence of lesions or sores in the oral cavity	Feature
UNEXPLAINED_BLEEDING	Bleeding with no explainable cause	Feature
DIFFICULTY_SWALLOWING	Difficulty swallowing food or beverages	Feature
WHITE_OR_RED_PATCHES	Presence of white or red patches inside the mouth	Feature
TUMOR_SIZE_CM	Tumor size in centimeters (numeric)	Deleted
CANCER_STAGE	Stage of cancer the patient has	Deleted
TREATMENT_TYPE	Type of treatment received by the patient	Feature
SURVIVAL_RATE_5YEAR	5-year survival rate percentage (%)	Deleted
COST_OF_TREATMENT_USD	Treatment cost in United States Dollars (USD)	Deleted

Fitur Name	Description	Status
ECONOMIC_BURDEN_LOST	Economic burden due to lost working days per year	Deleted
EARLY_DIAGNOSIS	Status of early oral cancer diagnosis in the patient	Feature
ORAL_CANCER_DIAGNOSIS	Target label, indicating whether the patient has oral cancer or not	Target

2.2 Data Pre-Processing

The pre-processing stage was conducted to ensure the quality of the dataset before it was used in the oral cancer classification modeling process using the Random Forest algorithm. The dataset used consisted of patient clinical data covering various risk factors and health conditions related to oral cancer. The first step was to remove columns that were not predictive. A total of 7 columns were removed from the dataset: the ID and Country columns, as the data in these columns did not contain clinically informative information; and the Tumor Size, Cancer Stage, Survival Rate, Cost of Treatment, and Economic Burden columns, as the data in these columns were the result of the diagnostic process, not predictor variables. The Treatment Type column was retained in the dataset because it reflects the treatment status currently being or previously undergone by the patient; not all patients receiving a specific treatment are diagnosed with oral cancer, so this column has the potential to provide relevant predictive information for the model. Including diagnostic result columns in the model training process has the potential to cause data leakage, a condition in which information that should not be known at the time of prediction actually influences the model, resulting in invalid performance evaluations [3]. The second step is data cleaning to identify and address invalid values, null values, missing values, and inconsistencies in the remaining attributes. The third step is encoding categorical variables using LabelEncoder, which was applied to the 16 categorical features in the dataset to convert categorical data into numerical values that can be processed by machine learning algorithms. This is done because machine learning algorithms, particularly those based on statistics and mathematics, are designed to process and analyze numerical data; therefore, categorical data must be converted into numerical form as part of the feature engineering stage [11]. The fourth step is the normalization of numerical data using a MinMaxScaler applied specifically to the Age column, so that the values in that column are scaled within the range of 0 to 1 and do not dominate the model training process. Normalization is applied only to the Age column because it is the only numerical feature with a value range that differs significantly from other features, potentially dominating the model if not scaled first.

2.3 Splitting of training and test data

The dataset used in this study was sourced from a publicly available oral cancer dataset containing a total of 84,922 data points. However, only 1,500 data points from the entire dataset were used in this study. This partial data selection was made based on several considerations, including maintaining computational efficiency during model training, reducing the processing load during the GridSearchCV phase, which involves a comprehensive search for hyperparameter combinations and ensuring that the experimental process can proceed in a controlled and reproducible manner. Although only 1,500 data points were used, the target class distribution in this data subset was maintained to remain representative of the class distribution in the original dataset.

After undergoing data preprocessing, the dataset was then divided into two parts: a 70% training set and a 30% testing set. This division was performed stratified by target

class to ensure that the distribution of the Yes and No classes remained proportional in both data subsets. This approach aims to train the model to independently recognize characteristic patterns, while also evaluating its performance in classifying oral cancer cases in previously unseen data. This approach is also crucial for minimizing the risk of overfitting and ensuring the model has good generalization capabilities when applied in real-world classification scenarios.

2.4 Classification of Oral Cancer Using the Random Forest Algorithm with GridSearchCV

The classification stage was conducted using the random forest algorithm combined with hyperparameter optimization techniques using GridSearchCV. This method enables a systematic search for the optimal parameter combination through a cross-validation scheme [5]. The dataset used in this stage consists of 1,500 samples randomly selected from a total of 84,922 data points in the Oral Cancer Prediction dataset. From the 24 available predictor variables, data cleaning was performed to identify the 17 most relevant predictor variables for use in the classification process, namely: Age, gender, tobacco use, alcohol consumption, HPV infection, betel nut use, chronic sun exposure, poor oral hygiene, dietary patterns (fruit and vegetable intake), family history of cancer, weakened immune system, oral lesions, unexplained bleeding, difficulty swallowing, white or red patches in the mouth, early diagnosis, and type of treatment. The adjusted parameters include:

1. n_estimators: 50, 100, 200 - the number of decision trees used in the ensemble. The value 50 serves as a conservative baseline, 100 is a standard commonly used in medical classification research, and 200 provides additional stability. This value is set to 200 because a dataset of 1,500 samples does not require a larger ensemble to achieve optimal convergence.
2. max_depth: None, 5, 10 - to set the maximum depth of each decision tree. The value None is used as a baseline, allowing the tree to grow fully; the value 5 is chosen to maintain a balance between bias and variance, while the value 10 represents a medium depth.
3. min_samples_split: 2, 5, 10 - the minimum number of samples required to split an internal node in a decision tree. These values are standard, commonly used options and are appropriate for a dataset of 1,500 samples, as they provide flexibility ranging from aggressive to conservative splitting.
4. min_samples_leaf: 1, 2, 4 - the minimum number of samples in each leaf node of the tree, which serves to improve the model's generalization ability. This combination of values is appropriate and does not need to be changed, as it aligns with the characteristics of a medium-sized dataset such as the one used in this study.

The use of GridSearchCV aims to identify the optimal model configuration that strikes a balance between accuracy and complexity. The best parameter combination identified is then used to train the final Random Forest model on the training data, resulting in more stable and reliable performance in predicting oral cancer.

2.5 Model Evaluation

Model evaluation was conducted to assess the performance of the Random Forest model in classifying oral cancer. In this study, model evaluation utilized several performance metrics commonly used in the field of machine learning, namely the confusion matrix, accuracy, precision, recall, and F1-score.
The definitions and formulas for each metric are as follows:

The confusion matrix is used to describe the distribution of correct and incorrect predictions, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The general structure of a confusion matrix is presented in below [6]:

Table 2. Confussion matrix

Actual	Predicted Positive	Predicted Negatif
Positif	TP (True Positive)	FN (False Negative)
Negatif	FP (False Positive)	TN (True Negative

These values form the basis for calculating the following metrics:

Accuracy is an evaluation metric used to measure the percentage of correct predictions out of the total number of predictions made by the model. This metric provides an overall picture of how often the model makes correct predictions, for both positive and negative classes.

$$(accuracy) = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is used to measure the accuracy of the model's positive predictions by calculating the ratio of True Positives (TP) to the total number of positive predictions (TP + FP). Precision can also be applied to oral cancer predictions to ensure that positive predictions are truly accurate, reduce false positive errors, and enhance the model's reliability [10].

$$(precesion) = \frac{TP}{TP + FP} \times 100\%$$

Recall is used to measure the model's ability to detect all positive cases by calculating the ratio of True Positives (TP) to the total number of positive cases (TP + FN). Recall plays a crucial role in minimizing false negative predictions to ensure that all positive cases are detected [10].

$$(recall) = \frac{TP}{TP + FN} \times 100\%$$

The F1 score is an evaluation metric used to calculate the harmonic mean of Precision and Recall, providing a balance between the accuracy of positive predictions and the ability to detect all positive cases. The F1 score is ideal for datasets with class imbalance [10].

$$(f1 - score) = 2x\frac{recall\ x\ precision}{recall - precision} \times 100\%$$

2.6 Implementing the Model in a Web Application Using Streamlit

The best model obtained from the training process and subsequent evaluation was implemented as a web application using the Streamlit framework. Streamlit is a framework specifically designed for Python, intended for building web applications in the fields of machine learning and data science. This framework was chosen for its ability to create interactive, real-time user interfaces.

Thus, users can easily access and use the prediction system without requiring technical knowledge of programming [5].

3. Results and Discussion

3.1. Preprocessing Results

Data preprocessing was performed prior to training the oral cancer classification model using the Random Forest algorithm. Data preprocessing is necessary to produce reliable analytical results and includes removing duplicates, identifying and handling invalid values, normalizing measurements, and transforming categorical information. In this study, preprocessing was conducted systematically through five main steps, as described below.

```
# ── Pre-processing ──────────────────────────────────
status_text.text("⚙️ Langkah 1/6 – Cleaning data...")
progress_bar.progress(10)

df = df_raw.drop(columns=KOLOM_HAPUS, errors='ignore').copy()
df = df.drop_duplicates()
for col in df.columns:
    if df[col].isnull().sum() > 0:
        if df[col].dtype == 'object':
            df[col].fillna(df[col].mode()[0], inplace=True)
        else:
            df[col].fillna(df[col].median(), inplace=True)

# ── Target Encoding ─────────────────────────────────
status_text.text("⚙️ Langkah 2/6 – Target and feature encoding..")
progress_bar.progress(25)

le_target = LabelEncoder()
y = le_target.fit_transform(df['Oral Cancer (Diagnosis)'].astype(str))
X = df.drop(columns=['Oral Cancer (Diagnosis)']).copy()
```

Figure 1. Data Preprocessing Steps Implemented in Python

Figure 1 shows the data preprocessing steps implemented in Python. The process begins with data cleaning (Step 1/6), which consists of three sequential operations. The first operation removes unnecessary columns that have been predefined in a list, ensuring that only relevant features are retained for the modeling process. The second operation eliminates duplicate rows using drop_duplicates(), preventing redundant data from affecting the model's learning process. The third operation handles missing values conditionally, if a column is categorical (dtype == 'object'), the missing values are imputed using the mode via fillna(df[col].mode()[0]), whereas if a column is numerical, the missing values are imputed using the median via fillna(df[col].median()). This conditional approach ensures that each column is treated appropriately based on its data type, preserving the statistical characteristics of the original data.

The process then continues with the target encoding stage (Step 2/6), where a LabelEncoder is applied to convert the target column Oral Cancer (Diagnose) into a numerical format using fit_transform(). This conversion is necessary because machine learning algorithms, including Random Forest, require numerical input rather than categorical string values. Following the encoding, the feature matrix X is separated from the target variable y by dropping the target column from the dataframe. A progress bar is implemented to visually monitor each step as the execution proceeds in the Streamlit application.

3.1.1. Removing Irrelevant Columns

The first step is to remove columns that do not contribute to the model's predictive power. A total of 7 columns were eliminated from the dataset: the ID and Country columns, which do not contain meaningful clinical information; and the Tumor Size (cm), Cancer Stage, Survival Rate (5-Year, %), Cost of Treatment (USD), and Economic Burden (Lost Workdays per Year) columns, which are diagnostic outcome variables. After removal, the dataset was reduced to 18 columns. Including these outcome variables in the

training process has the potential to cause data leakage, a condition where data leakage in machine learning classification often leads to overfitting, overly optimistic performance estimates, and poor reproducibility, thereby compromising the reliability of the resulting model [3].

3.1.2. Data Cleaning

The data cleaning phase was conducted to identify and address data duplicates and missing values. The review identified 7 duplicate rows, which were subsequently removed from the dataset. Meanwhile, the review of missing values revealed no missing values across all remaining attributes, so no further imputation was required. Data cleaning is crucial because differences in scale between attributes can cause one variable to dominate others, resulting in an unbalanced model.

3.1.3. Encoding the Target Variable

The target variable "Oral Cancer (Diagnose)," which was originally categorical, was converted to a numerical format using a LabelEncoder. The encoding results in a mapping of "No" → 0 and "Yes" → 1. This is done because machine learning algorithms, particularly those based on statistics and mathematics, are designed to process and analyze numerical data; therefore, text or categorical data must be converted into numerical form as part of the feature engineering stage.

3.1.4. Encoding Categorical Features

A total of 16 categorical features in the predictor variables were converted to numerical form using LabelEncoder, including Gender, Tobacco Use, Alcohol Consumption, HPV Infection, Betel Quid Use, Chronic Sun Exposure, Poor Oral Hygiene, Diet (Fruit & Vegetable Intake), Family History of Cancer, Compromised Immune System, Oral Lesions, Unexplained Bleeding, Difficulty Swallowing, White or Red Patches in the Mouth, Treatment Type, and Early Diagnosis. A Label Encoder is a method that converts categorical data in a dataset into numerical values, making it easier for machine learning algorithms to process the data; during the encoding process, each category is represented as a unique number such as 0 or 1 [11]. After the encoding process, all features in the dataset are in numerical form and ready for use in model training.

3.1.5. Normalization of Numeric Data

Normalization was performed specifically on the Age column using MinMaxScaler with a feature_range of [0, 1]. Data normalization is not applied to all datasets, but only to those with scale differences among features, where such differences can cause features with larger values to potentially dominate others, necessitating normalization to balance each feature's scale to a smaller ratio [4]. With MinMaxScaler normalization, the values in the Age column are scaled within the range of 0 to 1 so that they do not dominate the model training process. After all preprocessing steps are complete, the final dataset has 17 predictor features.

3.1.6. Training and Test Data Split

The dataset used was sourced from a public oral cancer dataset containing a total of 894,922 records, from which 1,500 samples were randomly selected (random sampling). After data cleaning, which involved removing 7 duplicates, the total number of samples used in the modeling process was 1,493. The dataset was then split in a 70:30 ratio using stratified splitting, resulting in 1,045 training samples and 448 test samples. The use of a

70/30 stratified split is a common approach in machine learning-based research, where this division aims to maintain the class proportions in the training and test data so that model performance evaluation can be conducted in a representative and unbiased manner.

The class distribution in the training data shows 553 'No' class samples and 492 'Yes' class samples, while the test data contains 237 'No' class samples and 211 'Yes' class samples. This relatively balanced distribution between the training and test data indicates that the stratified splitting approach successfully maintains proportional class representation, thereby minimizing the risk of evaluation bias due to class distribution imbalance.

3.2. Training the Random Forest Model with GridSearchCV

The model was developed using the Random Forest algorithm, which is an ensemble learning method that works by building multiple decision trees simultaneously, then combining the predictions from all of these trees through a majority voting mechanism to produce a final prediction. The more trees that are built, the more stable and accurate the predictions become, because errors from one tree can be corrected by the others.

```
# ── GridSearchCV ──────────────────────────────────
total_iter = (
    len(PARAM_GRID['n_estimators']) *
    len(PARAM_GRID['max_depth']) *
    len(PARAM_GRID['min_samples_split']) *
    len(PARAM_GRID['min_samples_leaf'])
) * 5
status_text.text(f"⚙️ Step 4/6 — GridSearchCV ({total_iter} iterasi)... Please wait")
progress_bar.progress(55)

rf_base = RandomForestClassifier(class_weight='balanced', random_state=42)
grid_search = GridSearchCV(
    estimator=rf_base, param_grid=PARAM_GRID,
    cv=5, scoring='accuracy', n_jobs=-1, verbose=0
)
grid_search.fit(X_train, y_train)
best_model = grid_search.best_estimator_
```

Figure 2. Steps Implemented in Python

Figure 2 illustrates the implementation of GridSearchCV used for hyperparameter tuning in the Random Forest Classifier model. The model is initialized with two important parameters. The first parameter is `class_weight='balanced'`, which is used because the oral cancer prediction dataset may have an imbalance in the number of data points between the Positive and Negative classes. With this parameter, the model automatically assigns a higher weight to the minority class so that the model does not tend to predict only the majority class. The second parameter is `random_state=42`, which locks the random seed used during the decision tree building process, ensuring that running the code with the same data will produce an identical model and that the research results are reproducible.

Next, the initialized Random Forest model is passed into GridSearchCV as the primary estimator. GridSearchCV is responsible for comprehensively searching for the best hyperparameter combination within the specified parameter space by automatically evaluating each combination through 5-fold cross-validation. Thus, the training process does not rely on a single parameter configuration but instead tests all possible combinations to find the most optimal settings.

Once the GridSearchCV configuration is complete, the actual training process begins via the command `grid_search.fit(x_train, y_train)`. At this stage, the Random Forest builds a number of decision trees in parallel, where each tree is trained using a subset of data selected randomly via bootstrap sampling. Additionally, each tree only considers a random subset of features at each branching point, resulting in trees that are diverse from

one another. This diversity is the primary strength of Random Forest, as it makes the model more resistant to overfitting compared to a single decision tree [5].

GridSearchCV automatically evaluates every hyperparameter combination using 5-fold cross-validation, in which the training data is divided into five parts in turn and used as validation data to ensure good model generalization. From all the tested combinations, the best parameters were obtained as follows:

Best Parameters from GridSearchCV

	max_depth	min_samples_leaf	min_samples_split	n_estimators
Best Value	10	1	10	200

Figure 3. Best Parameter Results from GridSearchCV

Figure 3 shows the best hyperparameter configuration obtained through GridSearchCV optimization. max_depth = 10, min_samples_leaf = 1, min_samples_split = 10, and n_estimators = 200. This best model is then used to objectively evaluate performance on the test data, and is saved as the file rf_oral_cancer_model.pkl using pickle, so that the model can be reloaded at any time to make predictions without having to repeat the entire training process from the beginning.

3.3 Model Performance Evaluation

The model was evaluated using several performance metrics commonly used in the field of machine learning, namely the confusion matrix, accuracy, precision, recall, and F1-score. The following are the results of the tests that were conducted:

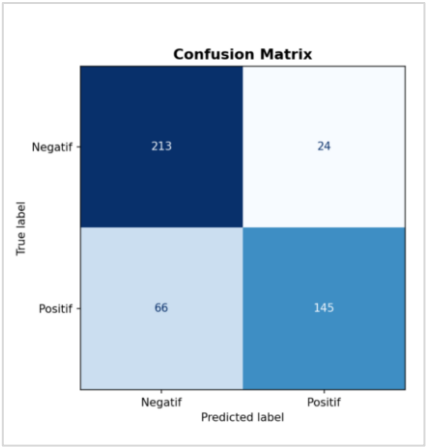

Figure 4. Confusion Matrix Results for the Random Forest Prediction Model

Figure 4 shows the confusion matrix of the model's prediction results using the random forest algorithm. The model successfully classified 213 negative patient data points (TN) and 145 positive patient data points (TP) accurately, with 2 negative patient data points incorrectly predicted as positive (FP) and 66 positive patient data points incorrectly predicted as negative (FN). This indicates that the model has a high level of specificity in identifying healthy (negative) patient data, but still shows a tendency toward underdiagnosis, as evidenced by the high False Negative (FN) rate of 66 patients.

Classification Report

	precision	recall	f1-score	support
Negatif (No)	0.76	0.90	0.83	237
Positif (Yes)	0.86	0.69	0.76	211
accuracy			0.80	448
macro avg	0.81	0.79	0.79	448
weighted avg	0.81	0.80	0.80	448

Figure 5. Classification Report: Precision, Recall, F1-Score

Figure 5 presents the classification report with a model accuracy of 80%. For the negative label, the model shows a precision of 76%, supported by an optimal recall of 90%. For the positive label, the model shows a precision of 86% with a lower recall of 69%. This difference in sensitivity results in an F1-Score of 83% for the negative label and 76% for the positive label, indicating that the model is more effective at confirming healthy patients than at detecting patients at risk for oral cancer.

3.4 Feature Importance Results

Feature importance is necessary to identify the independent variables that contribute most significantly to predicting oral cancer [13]. Based on the results obtained from the model training process in Google Colab, the following results were obtained:

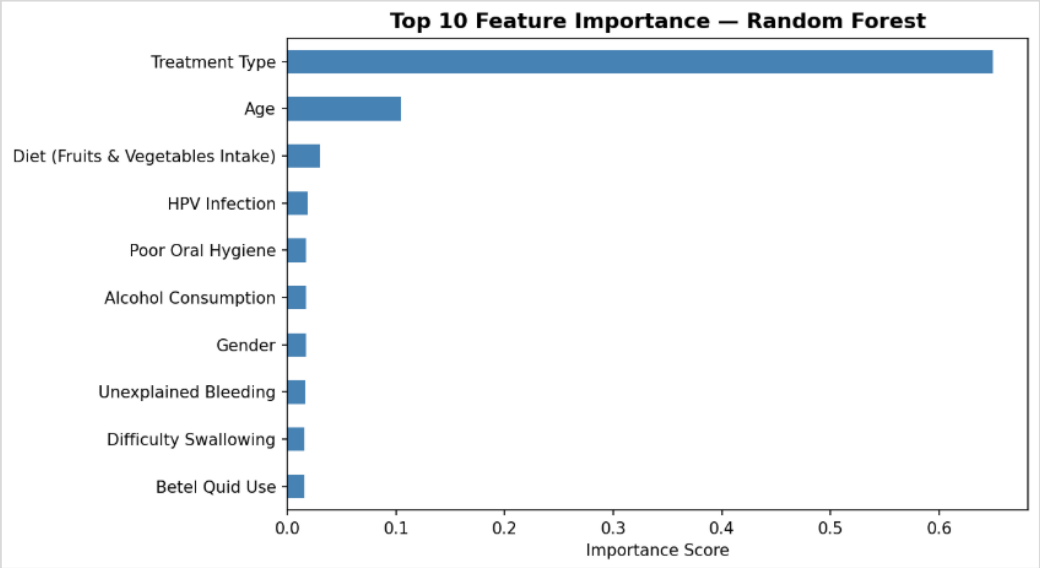

Figure 6. Top 10 Feature Importance

Figure 6 show the results of the Feature Importance analysis show that TREATMENT TYPE is the most influential variable in predicting oral cancer, with a contribution value of 0.649; this indicates that the type of treatment a patient receives is the primary factor influencing the model's prediction of oral cancer. The AGE variable, with a contribution value of 0.104, indicates that a person's risk of developing oral cancer increases with age[5].

The DIET variable (FRUITS & VEGETABLES INTAKE), with a value of 0.03, also contributes to the prediction of oral cancer, reflecting dietary patterns and lifestyle.

Meanwhile, risk factors such as HPV INFECTION, POOR ORAL HYGIENE, ALCOHOL CONSUMPTION, GENDER, UNEXPLAINED BLEEDING, DIFFICULTY SWALLOWING, and BETEL QUID USE are also important variables in supporting the prediction of oral cancer.

3.5 Implementation of the Streamlit Application

The oral cancer classification model that had been trained and optimized was implemented into an interactive web application using the Streamlit framework. Streamlit is an open-source Python framework specifically designed for data scientists and machine learning engineers to build interactive web applications without requiring extensive knowledge of HTML, CSS, or JavaScript [20]. The application interface is divided into four main tabs: Model Training, Prediction, Model Performance, and About the System

3.5.1 Training Model Tab

The Training Model tab serves as the entry point for users who need to train the classification model from scratch. As shown in Figure 7, this tab displays the application header 'Oral cancer Prediction' with a blue gradient background, followed by the four navigation tabs at the top. The tab contains an instruction box with a yellow background explaining that the dataset must be uploaded in .xlsx format and that training only needs to be done once, as the trained model will be saved automatically as .pkl files for future predictions without retraining [20]

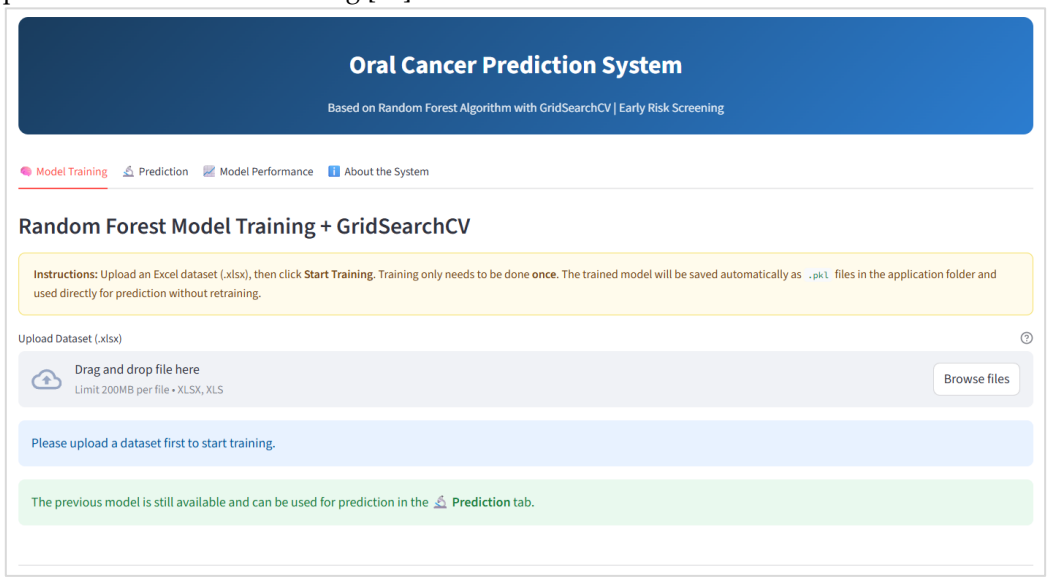

Figure 7. Training Model Tab Interface of the Oral Cancer Detection Application

Below the instruction box, a file uploader widget is displayed with a drag-and-drop area labeled 'Upload Dataset (.xlsx)' and a 'Browse files' button. Two notification banners are shown below the uploader: a blue banner informing the user to upload the dataset first, and a green banner confirming that the previous model is still available and can be used in the Prediction Tap.

3.5.2 Prediction Tap

The Prediction Tap is the main interface for end users to perform oral cancer risk screening. As shown in Figure 6, this tab displays a patient data questionnaire form with the heading 'Patient Data Questionnaire' along with a brief instruction at the top. The form is divided into two columns to accommodate all 17 input features in a structured and readable layout. [20]

The left column is organized into two groups. The first group, labeled Demographic Data s, contains an age slider ranging from 18 to 89 years with a default value of 45, and a gender radio button with options Male and Female. The second group, labeled Habits & Lifestyle, contains five Yes/No radio buttons for Tobacco Use, Alcohol Consumption, Betel Quid Use, Chronic Sun Exposure, and Poor Oral Hygiene, as well as a dropdown menu for diet level with options Low, Moderate, and High.

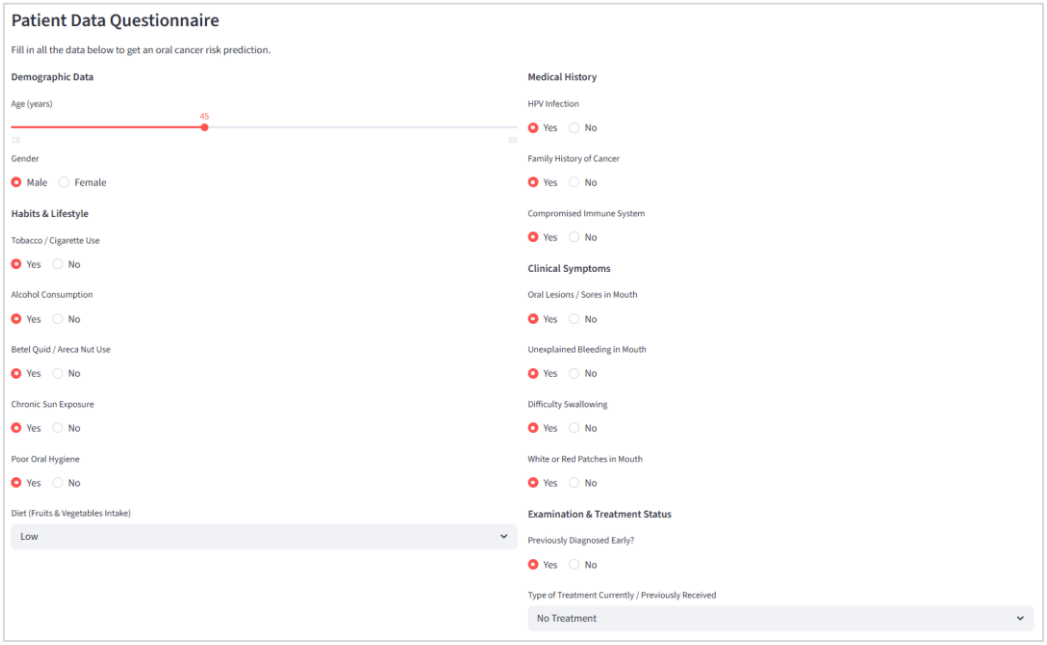

Figure 8. Patient Data Questionnaire Form of the Oral Cancer Detection Application

Figure 8 show the right column is organized into three groups. The first group, labeled Medical History, contains three Yes/No radio buttons for HPV Infection, Family History of Cancer, and Compromised Immune System. The second group, labeled Clinical Symptoms, contains four Yes/No radio buttons for Oral Lesions, Unexplained Bleeding, Difficulty Swallowing, and White or Red Patches in Mouth. The third group, labeled Examination & Treatment Status, contains a Yes/No radio button for Early Diagnosis status and a dropdown menu for Treatment Type with five options: No Treatment, Surgery, Radiation, Chemotherapy, and Targeted Therapy. After all fields are completed, the user clicks the Predict Now' button to obtain the prediction result.

3.5.3 Prediction Output Display

The prediction output is displayed through three components rendered sequentially below the form after the user clicks the prediction button: metric cards, a colored result box, and a probability distribution chart. The output display varies depending on whether the model predicts a positive or negative diagnosis.

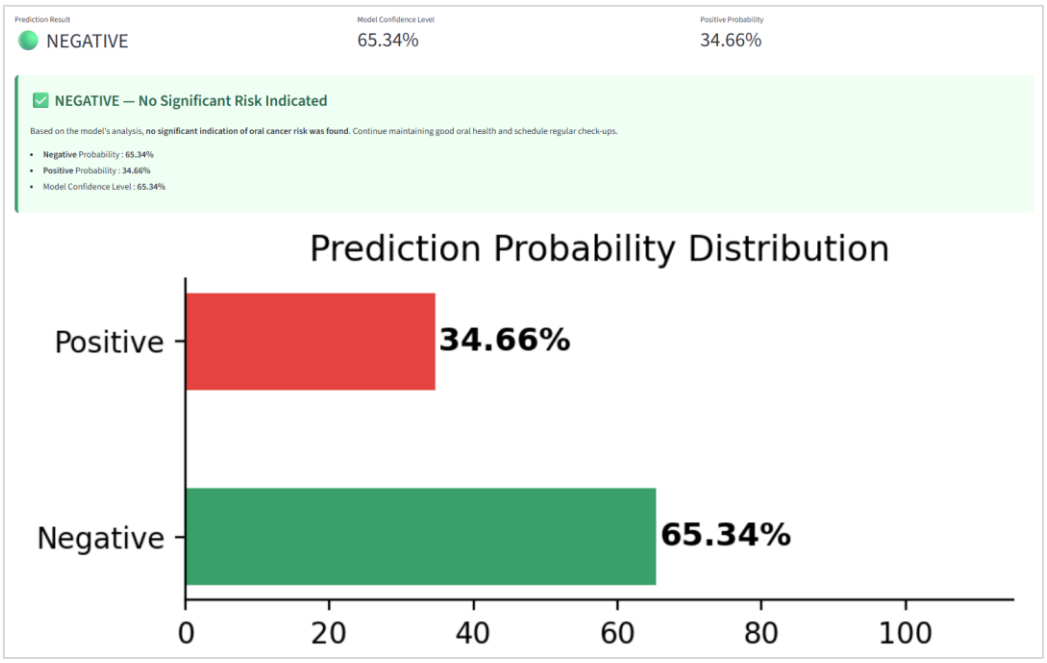

Figure 9. Positive Prediction Result with Probability Distribution (67.03% Positive)

Figure 9 shows the prediction output when the model detects a positive indication of oral cancer risk. The three metric cards at the top display the prediction result label POSITIVE, the model confidence level of 67.03%, and the positive class probability of 67.03%. Below the metric cards, a red-bordered box with a light red background appears containing the heading '⚠ POSITIVE — Oral Cancer Risk Indicated ' along with a description stating that the patient potentially has symptoms or risk factors of oral cancer, a breakdown of the probability values, and a recommendation to immediately consult a doctor or specialist. A Matplotlib horizontal bar chart then visualizes the probability distribution, showing the positive bar in red at 67.03% and the negative bar in green at 32.97%, giving users an intuitive understanding of the model's confidence level beyond just the binary label [16,20].

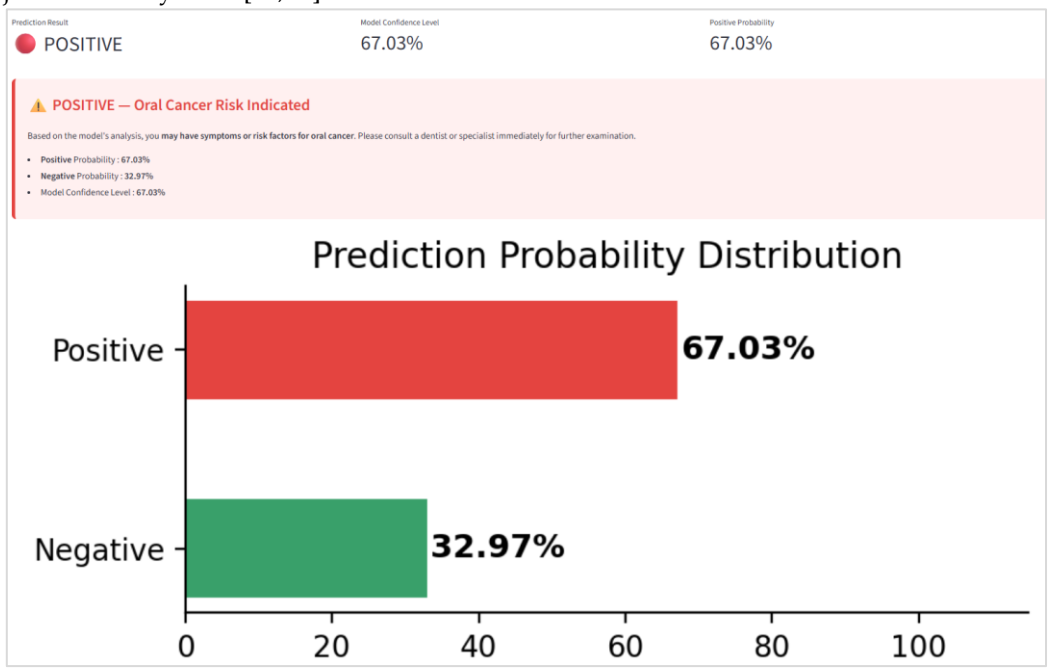

Figure 10. Negative Prediction Result with Probability Distribution (65.34% Negative)

Figure 10 shows the prediction output when the model does not detect a significant indication of oral cancer risk. The three metric cards display the prediction result label NEGATIVE, the model confidence level of 65.34%, and the positive class probability of 34.66%. A green-bordered box with a light green background appears containing the heading '☑ NEGATIVE — No Significant Risk Indicated' along with a statement that no significant indication of oral cancer risk was found, a breakdown of the probability values, and preventive health recommendations such as maintaining oral hygiene and scheduling routine dental checkups. The probability distribution chart shows the negative bar in green at 65.34% and the positive bar in red at 34.66%, confirming that the model is more confident in the negative classification [16,20]

3.5.4 Model Performance Tab

The Performa Model tab displays the evaluation results of the trained Random Forest model.

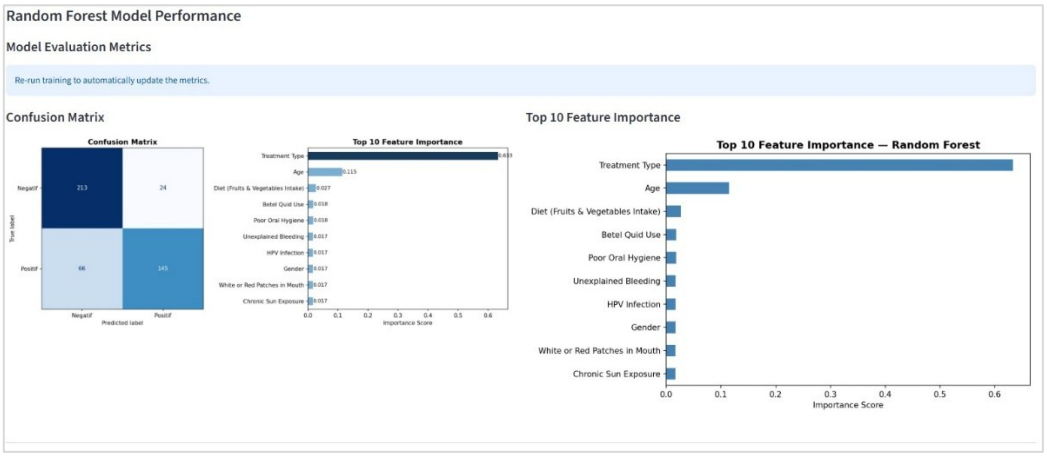

Figure 11. Model Performance Tab Showing Confusion Matrix and Top 10 Feature Importance

As shown in Figure 11, this tab contains the heading ' Random Forest Model Performance' followed by a blue information banner reminding the user to rerun the training to update the metrics automatically. The tab presents two side-by-side visualization panels. The left panel displays the Confusion Matrix, which shows the distribution of correct and incorrect predictions across the two classes (N Negative and Positive) in a blue-toned matrix [16]. The right panel displays of the Top 10 Feature Importance chart, which shows the ten most influential features in the Random Forest model ranked by their importance score. Based on the chart. Both images are generated automatically during the training process and loaded dynamically when this tab is opened [23,16].

3.5.5 About the System Tab

The About the System tab provides non-technical users with a comprehensive overview of the application system.

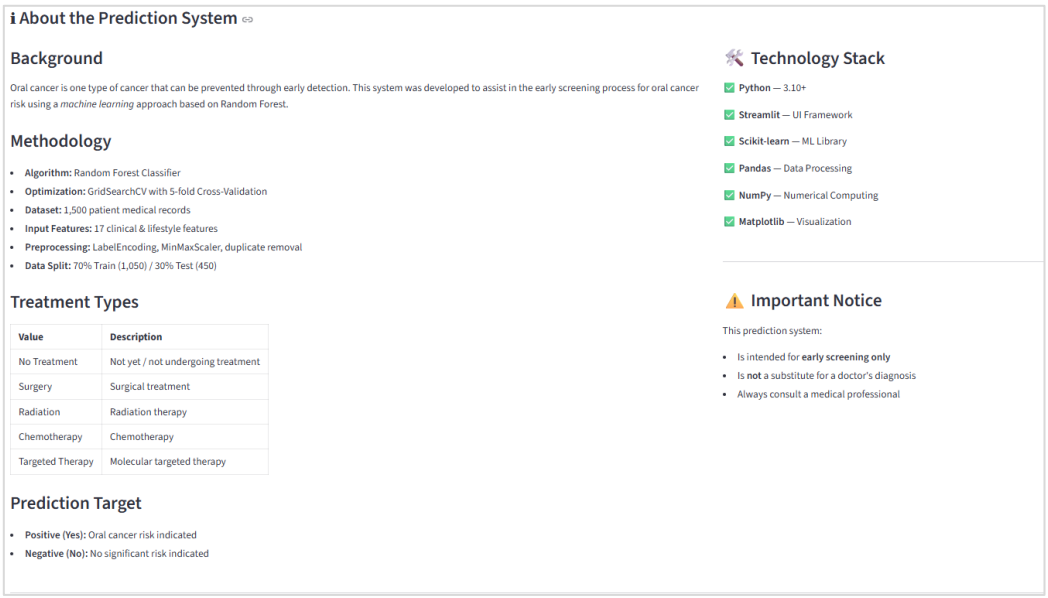

Figure 12. About the System Tab of the Oral Cancer Detection Application

As shown in Figure 12, this tab is divided into two columns. The left column presents four sections: Background, which explains that the system was developed to assist in early oral cancer risk screening using a Random Forest-based machine learning approach; Methodology, which lists the algorithm, optimization technique, dataset size, feature count, preprocessing steps, and data split ratio; Treatment Types, which presents a table describing the five treatment type values and their meanings; and Prediction Target, which defines the two output classes where Positive (Yes) indicates oral cancer risk and Negative (No) indicates no significant risk.

The right column presents two sections. The Technology i section lists six technologies used in the application development: Python 3.10+, Streamlit as the UI framework, Scikit-learn as the ML library, Pandas for data processing, NumPy for numerical computation, and Matplotlib for visualization. The Important Notice section displays a disclaimer stating that the system is intended only for early screening purposes, is not a substitute for a doctor's diagnosis, and users should always consult with medical professionals [20].

Overall, the Streamlit application successfully integrates the entire machine learning pipeline — from model training to real-time prediction — into a single accessible web interface with four well-organized tabs. Each tab serves a distinct purpose: the Training Model tab for model building, the Prediction tab for individual patient screening, the Model Performance tab for evaluation visualization, and the About the System tab for system documentation. This design supports both research reproducibility and practical usability in clinical screening contexts [16,20].

4. Conclusions

This study successfully developed an oral cancer classification model using the Random Forest algorithm optimized with GridSearchCV and implemented it into a Streamlit-based web application for early screening purposes. Based on the results obtained, the following conclusions can be drawn.

First, the data preprocessing stage — which included removing 7 irrelevant columns, eliminating 7 duplicate records, encoding 16 categorical variables using LabelEncoder, and normalizing the Age column using MinMaxScaler — successfully produced a clean dataset comprising 17 predictor features from 1,493 valid samples, ready for use in the model training process.

Second, the hyperparameter optimization process using GridSearchCV with a 5-fold cross-validation scheme successfully identified the best parameter combination: n_estimators = 200, max_depth = 10, min_samples_split = 10, and min_samples_leaf = 1, yielding a Random Forest model with an optimal balance between bias and variance.

Third, model evaluation on the test set demonstrated an accuracy of 79.91%, a precision of 85.8% for the positive label, a recall of 68.72%, and an F1-Score of 76.32%. While the model showed strong specificity in confirming negative (healthy) patients with a recall of 90%, there remains a tendency toward under-diagnosis, as indicated by 66 false-negative cases. This finding warrants careful consideration in the context of health screening, where minimizing missed positive cases is clinically critical.

Fourth, feature importance analysis revealed that Treatment Type, Age, Diet (Fruits & Vegetables Intake), HPV Infection, and Poor Oral Hygiene are the most influential variables in predicting oral cancer risk. These findings can serve as a clinical reference for healthcare professionals in identifying priority risk factors that require monitoring and early intervention.

Fifth, the implementation of the trained model into a Streamlit-based web application successfully produced an interactive, accessible, and user-friendly prediction system. The application enables both healthcare professionals and the general public to perform early oral cancer screening in real time. Nevertheless, the system's predictions are intended solely as an initial screening tool and cannot replace an official medical diagnosis from a licensed healthcare professional.

For future research, it is recommended to use the full dataset of 84,922 samples to improve model generalization, and to explore alternative algorithms or ensemble methods to further improve performance — particularly the recall value — in order to minimize false-negative predictions in oral cancer screening.

5. Patents

Author Contributions: Conceptualization, A.P.B.A. and S.M.A.; methodology, Z.P.R.P., A.P.B.A., S.M.A. and M.I.S.; software, Z.P.R.P. and M.I.S.; validation, Z.P.R.P. and S.M.A.; formal analysis, Z.P.R.P. and A.P.B.A.; investigation, M.I.S.; resources, M.I.S.; data curation, M.I.S. and S.M.A.; writing—original draft preparation, A.P.B.A. (abstract, introduction, classification framework, and model training), Z.P.R.P. (model evaluation, implementation, performance evaluation, and feature importance), S.M.A. (introduction, preprocessing, data splitting, preprocessing results, and recommendations), and M.I.S. (data collection, Streamlit application, and conclusion); writing—review and editing, Z.P.R.P., A.P.B.A., S.M.A. and M.I.S.; visualization, Z.P.R.P. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The authors would like to express their sincere gratitude to:

1. Niyalatul Muna, S.Kom., M.T.
2. Muhammad Yunus, M.Kom.
3. Mudafiq Riyan Pratama, S.Kom, M.Kom.
4. Ika Widiastuti, S.Kom.,M.Kom.,P HD

As the lecturer of the Artificial Intelligence course, for the continuous guidance, constructive feedback, and support throughout the development of this research. We also thank all individuals who have provided moral and technical encouragement during the writing process.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.

[2] R. Amtha *et al.*, "Pelatihan Deteksi Dini Kanker Mulut dengan SAMURI pada Komunitas Penyintas Kanker Love and Healthy Tangerang," *ABDI MOESTOPO: Jurnal Pengabdian Pada Masyarakat*, vol. 5, no. 1, pp. 10–21, Jan. 2022, doi: 10.32509/abdimoestopo.v5i1.1749.

[3] A. Ichwani, R. Indra Kesuma, A. Setiawan, E. Wicaksono, and R. Hanifah, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning," *Jurnal Teknik Informatika (JUTIF)*, vol. 7, no. 1, 2026, doi: 10.52436/1.jutif.2026.7.2.5490.

[4] P. Palinggik Allorerung, A. Erna, M. Bagussahrir, and S. Alam, "Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," 2024

[5] D. Aprilianto and E. Rizal, "Klasifikasi Penyakit Kanker Paru Menggunakan Algoritma Random Forest Berbasis Streamlit," vol. 9, p. 2025, doi: 10.47002/metik.v9i2.1076.

[6] M. Fadli Kurniawan and D. Ayu Megawaty, "Comparison of Logistic Regression, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithms in Diabetes Prediction," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[7] S. Fitriani, E. Budiman, M. Fadli, M. Surono, and H. Sulistiani, "Optimalisasi Metode Random Forest menggunakan Particle Swarm Optimization dalam Prediksi Prestasi Mahasiswa," *Prosiding Seminar Nasional Teknologi Komputer dan Sains*, vol. 3, no. 1, pp. 406–415, 2025, [Online]. Available: https://prosiding.seminars.id/sainteks

[8] I. Muhamad and M. Matin, "Hyperparameter Tuning menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware."

[9] R. Mishra, "Biomarkers of oral premalignant epithelial lesions for clinical application," Oral Oncol., vol. 48, no. 7, pp. 578–584, Jul. 2012. doi: 10.1016/j.oraloncology.2012.01.017

[10] A. D. Dini, P. Stroke, M. Stramlit, T. Krisna Amarya, R. Firliana, and A. Ristyawan, "Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) 2025 453," Online.

[11] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technologia : Jurnal Ilmiah*, vol. 15, no. 1, p. 93, Jan. 2024, doi: 10.31602/tji.v15i1.13457.

[12] W. Apriliah *et al.*, "SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[13] F. S. Salam Nagalay, D. Rahma Aryanti, M. Liandani, F. Sains dan Teknologi, U. Wira Buana, and F. Ilmu Kesehatan, "PENGEMBANGAN APLIKASI PREDIKSI RISIKO PENYAKIT JANTUNG BERBASIS MODEL KLASIFIKASI RANDOM FOREST MENGGUNAKAN FRAMEWORK STREAMLIT," *Jurnal Kesehatan Wira Buana*, vol. 9, pp. 2541–5387, 2025.

[14] A. A. Saputra and F. Ardilah, "Program Studi Teknik Informatika," 2026.

[15] "randomforest2001".

[16] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[17] N. F. Ayuningtyas, "Analisis pengetahuan tentang kelainan mulut berpotensi ganas pada komunitas sepeda motor pria di Indonesia," Universitas Airlangga, 2023. [Online]. https://unair.ac.id/analisis-pengetahuan-tentang-kelainan-mulut-berpotensi-ganas-pada-komunitas-sepeda-motor-pria-di-indonesia/

[18] Supriatno, F. Adityawan, F. Dentawan, and A. S. Pritama, *Kanker mulut*. UGM Press, 2023. [Online]. Available: https://books.google.co.id/

[19] A. Panday, "Oral Cancer Prediction Dataset," Kaggle, 2024.

[20] Streamlit Inc., "Streamlit: The fastest way to build and share data apps," 2019. [Online]. Available: https://streamlit.io.

[21] N. W. Johnson, P. Jayasekara, and A. A. H. K. Amarasinghe, "Squamous cell carcinoma and precursor lesions of the oral cavity: epidemiology and aetiology," *Periodontology 2000*, vol. 57, no. 1, pp. 19-37, 2011, doi: 10.1111/j.1600-0757.2011.00401.x.

[22] S. Warnakulasuriya, N. W. Johnson, and I. van der Waal, "Nomenclature and classification of potentially malignant disorders of the oral mucosa," Journal of Oral Pathology & Medicine, vol. 36, no. 10, pp. 575–580, Nov. 2007. doi: 10.1111/j.1600-0714.2007.00582.x

[23] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001