ISSN (Online): 3139-7689

Available online: www.ijsrmes.com

International Journal for Scientific Research in Modern Engineering and Science

Published in Conference Special Issue: Vol. 1, Special Issue 1, pp. 304-312, 2026

NON-INVASIVE DIAGNOSTICS FOR ORAL CANCER: INTEGRATING RAMAN SPECTRAL FINGERPRINTS WITH MACHINE LEARNING

Aryan Singh[1*], Srishti Kumari[2], Devesh Sharma[3], Dr Shikha Pandey[4], Dr. Mithilesh Singh[5]

[1 2 3]BTech Students, Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India
[4]Dept. of Computer Science and Engg., Bhilai Institute of Technology, Durg, India
[5]Electrical Department, Shri Rawatpura Sarkar University, Raipur, India

Abstract

Early and accurate detection of oral cancer is crucial for enhancing patient outcomes and alleviating the treatment burden. Oral squamous cell carcinoma (OSCC) develops on the mucosal epithelium of the oral cavity. According to the Global Cancer Observatory (GCO), the incidence of OSCC will rise regressively. Recent advances in Raman spectroscopy, stimulated Raman histology, AI-based spectral analysis, and multimodal data fusion have demonstrated promising capabilities for the detection of cancer without the need for extensive sample preparation. Raman Spectroscopy is a powerfull tool for identifying oral squamous cell carcinoma using machine learning algorithms. This paper represents a comprehensive framework that combines Raman spectroscopy and supervised machine learning for non- invasive oral cancer screening. A dataset consisting of 3556 Raman spectra was processed using a robust preprocessing pipeline that included spectral cropping, cosmic-ray removal, Savitzky–Golay smoothing, adaptive-smoothness-penalized least- squares baseline correction, and Min–Max normalization; feature extraction combined domain-specific peak intensities, Princi- pal Component Analysis (PCA), and entropy-based descriptors. Multiple classifiers were evaluated; XGBoost achieved the best performance with 96.0.

Keywords: *Raman spectroscopy, machine learning, oral cancer detection, XGBoost, spectral preprocessing, non-invasive diagnostics, Principal Component Analysis, baseline correction.*

* *Corresponding author*

1. Introduction

Oral cancer contributes substantially to global cancer mor- bidity and mortality. Oral and head-and-neck cancers are still some of the most widespread cancers around the world, and oral squamous cell carcinoma (OSCC) makes up the vast majority of tumors found in the oral cavity. OSCC affects more males than females, with middle-aged to elderly men being the most susceptible. OSCC results in disfiguration and functional impairments including swallowing, speech and taste, which have a substantial impact on the life quality. The real challenge is that many cases are detected late, when treatment becomes harder, and survival rates drop sharply—so identifying the disease early and quickly is incredibly important. Early- stage detection drastically improves survival rates, yet the

current gold-standard histopathology is invasive and resource- intensive. Optical techniques such as Raman spectroscopy enable the label-free and molecularly specific interrogation of tissue, providing spectral fingerprints that reflect changes in biochemical composition associated with malignancy [1]. OSCC spreads predominantly to ipsilateral lymph nodes of the neck via lymphatic outflow, but can also invade contralateral or bilateral lymph nodes. Lungs, bones, and the liver are typical locations for OSCC metastases. The integration of Raman spectroscopy with machine learning offers a approach for early detection and risk for OSCC, improving outcomes and survival rates. However, Raman data are high-dimensional and susceptible to noise and fluorescence, necessitating advanced preprocessing and machine learning (ML) for robust classi- fication. This paper presents a full pipeline from spectrum acquisition through machine learning-based classification, em- phasizing reproducibility and clinical relevance. The overall workflow is visualized in the Figure . 1

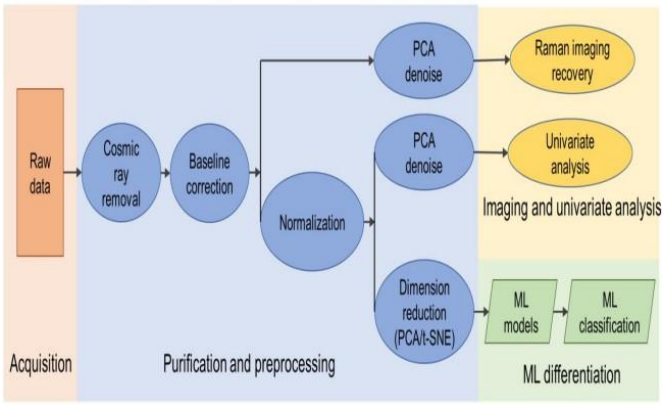


Fig. 1: Flowchart of the overall Workflow.

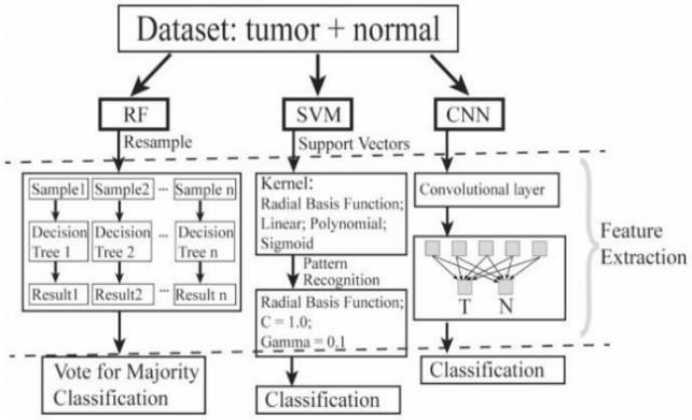


Fig. 2: Figure displays the distribution of the dataset, showing the number of samples available in each of the five classes.

2. Related Work

Recent advances in Raman spectroscopy, combined with machine learning, have significantly improved non-invasive cancer diagnostics. Han et al. demonstrated a fiber-optic Raman system integrated with deep learning models that can carry out multi-task diagnostics for oral cancer. Their study demonstrated that neural networks can concurrently distinguish tissue conditions and identify diagnostically im- portant spectral signatures,

underscoring the potential of end- to-end spectral learning for real-time clinical applications. In a separate investigation, Ghosh et al. assessed ensemble machine learning approaches using Raman cyto-spectroscopy data for early oral cancer risk prediction. Their findings showed that ensemble models—especially those incorporating multiple decision-tree-based algorithms—enhance robustness to spectral variability and noise.

In addition, machine learning has been leveraged on Raman spectra for various biomedical applications. Arau´jo et al. applied dimensionality reduction and supervised classifica- tion to pinpoint compact spectral signatures associated with melanoma, illustrating how refined feature selection enhances model performance. Similarly, Li et al. investigated multi-task deep learning for fibre-optic Raman–based diagnostics, which focuses on fast data handling and suitability for clinical work- flows. Together, these works highlight a clear movement to- ward combining sophisticated preprocessing, condensed spec- tral encodings, and deep or ensemble learning approaches to improve diagnostic accuracy and real-world applicability.

Figure 1 illustrates that Data acquisition is followed by pu- rification and preprocessing, leading to feature extraction and subsequent ML differentiation for classification and analysis. Figure 2 illustrates the general architecture of common ML models used in this domain. This work extends prior studies by implementing a rigorous preprocessing chain, interpretable feature selection, and detailed benchmarking against published methods.

3. Dataset and Problem Formulation

A. Data Description

The dataset consists of n = 3556 Raman spectra, each sampled at 1037 Raman shift values. Original labels included five classes (0: Healthy, 1: Premalignant, 2: Malignant, 3: Other conditions, 4: Noise/Ambiguous). For screening pur- poses, classes 1–4 were aggregated into a single Non-Healthy category, yielding a binary classification problem: Healthy (0) vs Non-Healthy (1).

For the study, we used the Raman spectroscopy dataset that was provided by X. Li et al., who developed a fiber-optic Raman acquisition system for multitasking diagnosis of oral cancer (OSCC). The dataset contained high-quality spectra collected from clinically validated oral tissue samples, making it a suitable benchmark for evaluating machine learning-based screening frameworks. To collect the Raman spectral data of tissues, firstly, a Raman spectra apparatus and the sample preparation of these tissues are given. Thirdly, we present cross-validation method for model estimation.

B. Objective

The objective is to develop an automated classifier that min- imizes false negatives (maximizes sensitivity) while maintain- ing acceptable specificity for screening. Models are evaluated using standard metrics derived from the confusion matrix (TP, TN, FP, FN).

4. Methodology

A. Preprocessing Pipeline

The preprocessing steps applied to each raw spectrum s(ν)
are as follows:

1) Spectral cropping to the biologically relevant fingerprint region (500–2000 cm−1).
2) Cosmic-ray spike removal (Whitaker algorithm).
3) Savitzky–Golay smoothing with window length 9 and polynomial order 3.
4) Baseline correction using penalized adaptive smoothness least squares (asPLS).
5) Min–Max normalization to scale intensities to [0, 1].

Preprocessing significantly enhances peak visibility and reduces fluorescence interference. The effect of the pipeline, specifically baseline correction, is illustrated in Figure 3.

B. Feature Extraction

The positive peaks were presented due to tumor tissue, and negative peaks appeared due to noraml tissue. Negative peaks dominated the difference for alsmost all of the dataset found in our dataset were higher under tumorous conditions.

Domain-relevant Raman features were selected at character- istic wavenumbers for healthy tissue: 608, 1437, 1526, 746, 837, 915, 1053 and 1704 cm−1. We define the feature vector for spectrum i as:

$$\mathbf{F}i = [I608, I_{1437}, I_{1526}, I_{746}, I_{837}, I_{915}, I_{1053}, I_{1704}, PC_1, PC_2, H], \quad (1)$$

Domain-relevant Raman features were selected at characteristic wavenumbers for non-healthy tissue: 1639, 543, 679, 814, 948, 1000, 1328, 1445, and 1658 cm−1. We define the feature vector for spectrum i as:

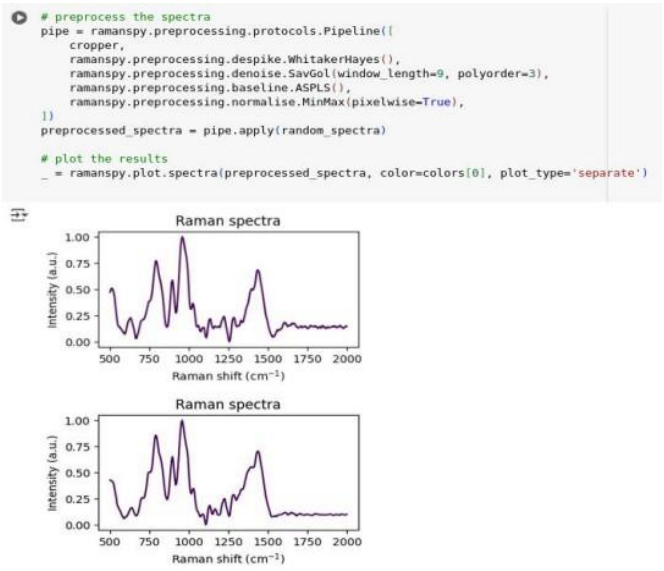


Fig. 3: The figure shows the Python implementation and the resultant Raman spectrum after baseline correction, displaying a flattened baseline and enhanced spectral peaks.

$$\mathbf{F}i = [I1639, I_{543}, I_{679}, I_{814}, I_{948}, I_{1000}, I_{1328}, I_{1445}, I_{1658}, PC_1, PC_2, H]$$

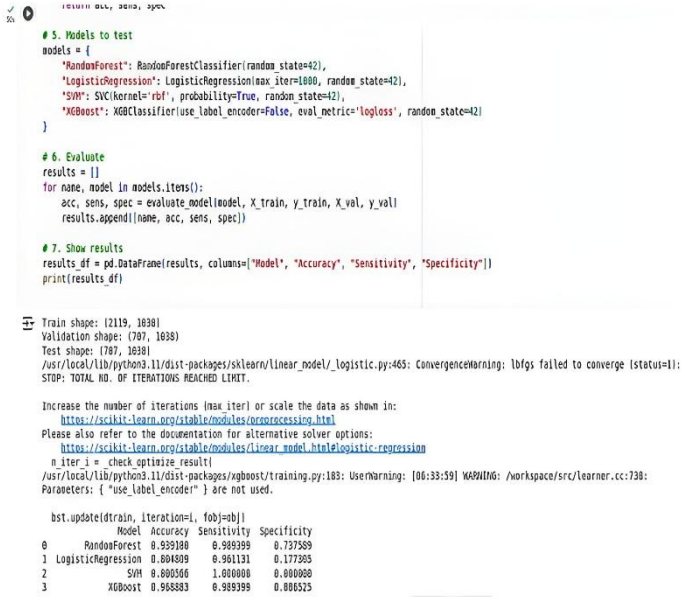
where Iλ denotes the normalized intensity at Raman shift λ, PCk are principal component scores obtained from PCA on the full fingerprint region, and H is the spectral entropy:

$$H = -\sum_j p_j \log p_j, \quad p_j = \frac{s_j}{\sum_k s_k}. \qquad (2)$$

The final feature matrix X ∈ Rn×m was used for supervised learning.

C. Classification Models and Training

We evaluated several supervised classifiers: Logistic Regres- sion, Support Vector Machine (RBF kernel), Random Forest (200 trees), XGBoost (learning rate 0.1, max depth 6), and a feedforward Neural Network with three hidden layers (ReLU activations). Receiver operator characteristics (ROC) study, the true positive rate (sensitivity) plotted against the false positive rate (1-specificity) for various values of discrimina- tion threshold to estimmate the discriminative ability of the tissue classification models. The ROC curve used to select a threshold for a classifier that maximizes true positives and minimizes false positives. It also allows us to evaluate the classifier's performance. Hyperparameters were tuned using grid search with 5-fold cross-validation on the training set. The train-validate-test split preserved class balance (70% train, 15% Validation, 15% test). The initial performance evaluation on the validation set is shown in Figure 4.



```
# 5. Models to test
models = {
    "RandomForest": RandomForestClassifier(random_state=42),
    "LogisticRegression": LogisticRegression(max_iter=1000, random_state=42),
    "SVM": SVC(kernel='rbf', probability=True, random_state=42),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)
}

# 6. Evaluate
results = []
for name, model in models.items():
    acc, sens, spec = evaluate_model(model, X_train, y_train, X_val, y_val)
    results.append([name, acc, sens, spec])

# 7. Show results
results_df = pd.DataFrame(results, columns=["Model", "Accuracy", "Sensitivity", "Specificity"])
print(results_df)
```

```
Train shape: (2119, 1038)
Validation shape: (707, 1038)
Test shape: (707, 1038)
/usr/local/lib/python3.11/dist-packages/sklearn/linear_model/_logistic.py:465: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. OF ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
/usr/local/lib/python3.11/dist-packages/xgboost/training.py:183: UserWarning: [06:33:59] WARNING: /workspace/src/learner.cc:738:
Parameters: { "use_label_encoder" } are not used.

  bst.update(dtrain, iteration=i, fobj=obj)
                Model  Accuracy  Sensitivity  Specificity
0        RandomForest  0.939180     0.989399     0.737589
1  LogisticRegression  0.804809     0.961131     0.177305
2                 SVM  0.860566     1.000000     0.000000
3             XGBoost  0.966883     0.989399     0.886525
```

Fig. 4: Initial Model Evaluation on the Validation Set. The figure compares the models and evaluates their performance using key metrics before final hyperparameter tuning.

TABLE I: Benchmarking against prior literature (test-set results)

Method	Accuracy(%)	Sensitivity(%)	Specificity(%)
XGBoost(This work)	**96.0**	**99.4**	82.1
SVM(Sahu et al.)	91.5	92.0	91.0
Random Forest(Kumar et al)	93.8	94.1	93.0

CNN(Liu et al.)	95.5	98.2	85.1
Logistic(Sharma et al.)	88.0	89.2	87.0

D. Evaluation Metrics

From the confusion matrix entries TP, TN, FP, FN, we compute:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (5)$$

5. Results

A. Model Performance

Table I summarizes the test-set performance. XGBoost outperformed other models, obtaining 96.0% accuracy, 99.4% sensitivity, and 82.1% specificity. Figure 5 presents the final performance results of the XGBoost model on the test set after hyperparameter optimization.

6. Discussion

The advancement in technology in terms of increased computaional power have allowed for improvement in model architecture of shallow machine learning. This study explored how Raman spectroscopy, supported by machine learning, can be used to characterize the biochemical alterations associated with Oral Squamous Cell Carcinoma (OSCC). The mean Raman spectra of each group were obtained for each model, to visually analyze the differences between groups. On the difference spectra between tumor tissue and normal tissue. The spectral data revealed distinct vibrational differences between healthy and cancerous tissues, indicating that molecular-level changes occurring during carcinogenesis can be detected reli- ably through Raman scattering. These differences formed the foundation for training machine learning models to recognise disease-specific patterns.

One of the important observations in this research is that the raw Raman spectra alone are not sufficient for accurate clas- sification. Preprocessing steps such as denoising, smoothing, baseline correction, and normalisation significantly enhanced the clarity of relevant features. The comparison of algorithms further showed that model behaviour varies depending on how well the classifier can adapt to variations within spectral peaks. Algorithms capable of dealing with complex, high- dimensional patterns displayed more stability and consistency.

Another key point emerging from the analysis is the sensi- tivity of Raman-based detection to sample variability. Factors such as tissue heterogeneity, patient-specific biochemical dif- ferences, and slight variations in acquisition parameters can influence the spectral profiles. This highlights the need to design robust preprocessing workflows and explore advanced model optimisation techniques such as hyperparameter tuning, dimensionality reduction, or ensemble learning.

```
# Train XGB
best_xgb = XGBClassifier(**study_xgb.best_params, eval_metric='logloss', random_state=42, use_label_encoder=False)
best_xgb.fit(X_train, y_train)
xgb_preds = best_xgb.predict(X_test)

# Metrics function
def evaluate_model(y_true, y_pred):
    cm = confusion_matrix(y_true, y_pred)
    tn, fp, fn, tp = cm.ravel()
    acc = accuracy_score(y_true, y_pred)
    sens = recall_score(y_true, y_pred)
    spec = tn / (tn + fp)
    return acc, sens, spec

# Results
rf_acc, rf_sens, rf_spec = evaluate_model(y_test, rf_preds)
xgb_acc, xgb_sens, xgb_spec = evaluate_model(y_test, xgb_preds)

# Summary table
import pandas as pd
results_df = pd.DataFrame({
    'Model': ['RandomForest', 'XGBoost'],
    'Accuracy': [rf_acc, xgb_acc],
    'Sensitivity': [rf_sens, xgb_sens],
    'Specificity': [rf_spec, xgb_spec]
})
print(results_df)
```

```
/usr/local/lib/python3.11/dist-packages/xgboost/training.py:183: UserWarning: [09:55:49] WARNING: /workspace/src/learner.cc:738:
Parameters: { "use_label_encoder" } are not used.

  bst.update(dtrain, iteration=i, fobj=obj)
          Model  Accuracy  Sensitivity  Specificity
0  RandomForest  0.947666     0.978836     0.821429
1       XGBoost  0.960396     0.994709     0.821429
```

Fig. 5: This figure presents the final, tuned results of the Ran- dom Forest and XGBoost models, highlighting the superior performance of the XGBoost model (96.0% Accuracy, 99.4% Sensitivity, 82.1% Specificity) after final tuning.

Finally, the study emphasises that Raman spectroscopy offers more than just classification potential. It provides a biochemical fingerprint of tissue composition, which gives re- searchers insights into molecular shifts associated with OSCC progression. While the diagnostic capability of this integrated approach is promising, its real value lies in enabling a deeper understanding of tumor biology. This perspective can guide fu- ture efforts to refine diagnostic tools, improve early screening protocols, and develop complementary spectroscopic methods.

7. Limitations and Future Work

Limitations include a single-centre dataset, potential class imbalance artefacts, and a lack of external clinical validation. Future directions:

(1) Expand multi-centre datasets to improve generalizability,

(2) Explore end-to-end deep spectral models and transformer architectures,

(3) Integrate hardware-in-the-loop for real-time Raman probe deployment, and

(4) Perform prospective clinical trials comparing predictions to histopathology.

8. Conclusion

In conclusion, this research demonstrates that integrating Raman spectroscopy with machine learning techniques is a reliable and effective approach for identifying Oral Squa- mous Cell Carcinoma (OSCC). The study fulfills its primary objective fulfils its primary objective by demonstrating that spectral data, when processed and analysed

with appropriate computational models, can differentiate cancerous tissues from normal tissues with meaningful accuracy. The results highlight the potential of this combined methodology as a non-invasive, rapid diagnostic tool that could support clinicians in early detection and decision-making. Although the study is limited by dataset size and requires further validation across diverse populations, it establishes a strong foundation for future advancements in spectroscopic diagnostics. Continued research in this direction may lead to improved model robustness, the development of real-time screening systems, and broader clinical applications that enhance early cancer diagnosis and patient outcomes.

Acknowledgment

We sincerely thank our research supervisor, Dr Shikha Pandey, for consistently guiding, suggesting, and supporting us throughout the course of this research. Her expertise and feedback were invaluable in shaping the direction and quality of our work.

We would also like to acknowledge the support of the Com- puter Science and Engineering Department, Bhilai Institute of Technology, Durg, and express our appreciation to our Head of Department, Dr. Sunita Soni, and the faculty members for creating a productive learning environment and providing all necessary resources required for this research.

Finally, we are grateful to our friends and family members who contributed to us through discussions, encouragement, and cooperation during various stages of the project.

REFERENCES

[1] Venkatakrishna, K., et al., “Optical pathology of oral tissue using Raman spectroscopy,” 2001.

[2] Sahu, S., et al., “In vivo subsite classification and diagnosis of oral cancer using Raman Spectroscopy,” IEEE Trans. Biomed. Eng., 2016.

[3] M. Sharma et al., “Raman Spectroscopy for oral premalignant and malignant lesion diagnosis,” J. Biomed. Opt., 2021.

[4] G. Liu et al., “Deep learning-based classification of oral squamous cell carcinoma using Raman spectroscopy,” J. Raman Spectrosc., 2024.

[5] D. Kumar et al., “Comparative study of ML algorithms for oral cancer classification,” Sensors Actuators B, 2024.

[6] J. Wang et al., “Integration of SERS and ML for ultra-sensitive oral cancer biomarker detection,” Anal. Chem., 2024.

[7] A. Khan et al., “Real-time clinical validation of an AI-powered Raman probe for oral cancer screening,” IEEE Trans. Biomed. Eng., 2025.

[8] P. M. Conforti et al., “Raman Spectroscopy and AI Applications in Cancer Grading: An Overview,” IEEE Access, vol. 12, 2024.

[9] J. Han et al., “Rapid multi-task diagnosis of oral cancer leveraging fibre- optic Raman spectroscopy and deep learning algorithms,” Anal. Chem., 2023.

[10] S. K. Ghosh et al., ”Machine learning algorithm ensembles for early oral cancer risk assessment using Raman cyto-spectroscopy,” Soft Comput., vol. 27, 2023.

[11] D. C. Arau´jo et al., “Finding reduced Raman spectroscopy fingerprint of skin samples for melanoma diagnosis through machine learning,” Artif. Intell. Med., 2021.

[12] X. Li et al., “Rapid multi-task diagnosis of oral cancer leveraging fiber- optic Raman spectroscopy and deep learning algorithms”, Frontiers in Oncology, vol. 13, 2023