oral

MDPI

Article

Predicting Leukoplakia and Oral Squamous Cell Carcinoma Using Interpretable Machine Learning: A Retrospective Analysis

Salem Shamsul Alam [1], Saif Ahmed [1], Taseef Hasan Farook [2,*] and James Dudley [2]

[1] Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh; salem.alam@northsouth.edu (S.S.A.); saif.ahmed02@northsouth.edu (S.A.)
[2] Adelaide Dental School, The University of Adelaide, Adelaide, SA 5000, Australia; james.dudley@adelaide.edu.au
* Correspondence: taseef.farook@adelaide.edu.au; Tel.: +61-4144-89858

Abstract: *Purpose*: The purpose of this study is to assess the effectiveness of the best performing interpretable machine learning models in the diagnoses of leukoplakia and oral squamous cell carcinoma (OSCC). *Methods*: A total of 237 patient cases were analysed that included information about patient demographics, lesion characteristics, and lifestyle factors, such as age, gender, tobacco use, and lesion size. The dataset was preprocessed and normalised, and then separated into training and testing sets. The following models were tested: K-Nearest Neighbours (KNN), Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest. The overall accuracy, Kappa score, class-specific precision, recall, and F1 score were used to assess performance. SHAP (SHapley Additive ExPlanations) was used to interpret the Random Forest model and determine the contribution of each feature to the predictions. *Results*: The Random Forest model had the best overall accuracy (93%) and Kappa score (0.90). For OSCC, it had a precision of 0.91, a recall of 1.00, and an F1 score of 0.95. The model had a precision of 1.00, recall of 0.78, and F1 score of 0.88 for leukoplakia without dysplasia. The precision for leukoplakia with dysplasia was 0.91, the recall was 1.00, and the F1 score was 0.95. The top three features influencing the prediction of leukoplakia with dysplasia are buccal mucosa localisation, ages greater than 60 years, and larger lesions. For leukoplakia without dysplasia, the key features are gingival localisation, larger lesions, and tongue localisation. In the case of OSCC, gingival localisation, floor-of-mouth localisation, and buccal mucosa localisation are the most influential features. *Conclusions*: The Random Forest model outperformed the other machine learning models in diagnosing oral cancer and potentially malignant oral lesions with higher accuracy and interpretability. The machine learning models struggled to identify dysplastic changes. Using SHAP improves the understanding of the importance of features, facilitating early diagnosis and possibly reducing mortality rates. The model notably indicated that lesions on the floor of the mouth were highly unlikely to be dysplastic, instead showing one of the highest probabilities for being OSCC.

Keywords: oral cancer; white lesion; Random Forest classifier; SHAP; predictive modelling

Citation: Alam, S.S.; Ahmed, S.; Farook, T.H.; Dudley, J. Predicting Leukoplakia and Oral Squamous Cell Carcinoma Using Interpretable Machine Learning: A Retrospective Analysis. *Oral* **2024**, *4*, 386–404. https://doi.org/10.3390/oral4030032

Academic Editor: Nejat Düzgüneş

Received: 13 August 2024
Revised: 10 September 2024
Accepted: 12 September 2024
Published: 13 September 2024

1. Introduction

Oral squamous cell carcinoma (OSCC) remains the most common type of oral cancer and accounts for nearly 90% of all oral malignancies. Approximately 300,000 new cases are diagnosed each year, with South Asia and parts of Europe having the highest prevalence due to certain lifestyle choices, social habits, and genetic predisposition [1–3]. Despite advances in treatment modalities, the five-year survival rate for OSCC remains less than 50% due to late-stage detection and the disease's aggressive nature [3]. This is likely due to a lack of information among the population, poor primary prevention campaigns, and poor symptomatology of the disease in the early stages of development [4]. The early

detection of OSCC and its precursor lesions, such as leukoplakia, is essential in improving survival rate.

Leukoplakia is a potentially malignant oral lesion that appears as white patches in the oral cavity and has a variable risk of developing into malignancy [5]. The diagnosis of leukoplakia, monitoring for dysplastic changes, and subsequent progression to OSCC has traditionally relied on histopathological examination of biopsy samples, which can also be subjective and prone to interobserver variability [5]. Clinical follow-ups confirming the malignancy potential of leukoplakia also remain complex and invasive. This variability warrants the need for adjunct diagnostic tools to assist clinicians in objective evaluation, which may lead to early detection and accurate diagnosis.

In recent years, machine learning (ML) has emerged as a transformative technology for tackling challenging dental diagnostics and analysing complex interactions within the head–neck region [6]. ML models can analyse large datasets and identify complex patterns and trends [7]. Some common forms of ML include K-Nearest Neighbours (KNN), Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest. Random Forests, for example, are becoming increasingly popular among ML models due to their robustness, ability to handle large feature sets, and ability to estimate inherent feature importance [8]. However, ML is still considered a "black box", with researchers guessing which factors influence the systems' decision making [9]. Clinical adoption of ML models necessitates high predictive performance and interpretability, ensuring that the models' decisions are transparent and understandable to healthcare professionals [10].

Several studies have investigated the application of ML to diagnosing oral cancer and its precursor lesions [11]. For example, Adeoye et al. demonstrated the utility of deep learning models in predicting malignant transformation-free survival of people with potentially malignant oral disorders, emphasising the importance of high accuracy and interpretability [12]. Similar designs of ML saw applicability in diagnosing co-dependent tumours. For example, Kutlu and Avcı used convolutional neural networks, discrete wavelet transforms, and extended short-term memory networks to classify liver and brain tumours [13]. However, in most cases, the issue with the "black box" remains unaddressed.

The present research builds on existing work by incorporating an additional layer of interpretability using SHapley Additive ExPlanations (SHAP) [14]. SHAP is based on cooperative game theory and is a method for extracting explanations from ML models about their decision-making processes [15]. It is effective in identifying the most influential features driving these models. SHAP values provide a consistent measure of feature importance, allowing us to understand how each feature contributes to the model's output [16].

This research aims to

1. Assess predictive models that analyse trends in the underlying features contributing to common white lesions, such as leukoplakia and OSCC;
2. Extract explanations for the decisions of the best performing machine learning models.

2. Material and Methods

2.1. Dataset Description

The current study adhered to the Minimum Information for Clinical Artificial Intelligence Modelling (MI-CLAIM) checklist, and all codes adhered to the PEP-8 guidelines [17]. The dataset included 237 patients, each with a diagnosis of oral cancer or leukoplakia as a potentially malignant oral lesion. Each patient was categorised into various features and variables by dental specialists at the time of diagnosis. The dataset was obtained through open access, courtesy of NDB-UFES (NDB-UFES: An oral cancer and leukoplakia dataset composed of histopathological images and patient data—Mendeley Data), and was published on 16 March 2023 [18]. The dataset was deidentified and, therefore, deemed exempt from ethical review. Twenty-four predictor variables were coded for the 237 patients (Table 1). Definitive diagnoses were ascertained through histopathological analyses and used to categorise the cases into three classes of lesion progression (Table 2).

Table 1. Descriptions of the 24 coded features.

Coded Feature Names	Description
localization_Tongue	Tongue localisation
localization_Lip	Lip localisation
localization_Floor of mouth	Floor-of-mouth localisation
localization_Buccal mucosa	Buccal mucosa localisation
localization_Palate	Palate localisation
localization_Gingiva	Gingival localisation
larger_size	Larger lesions
tobacco_use_Yes	Current habit of tobacco use
tobacco_use_Former	History of tobacco use
tobacco_use_No	No history of tobacco use
tobacco_use_Not informed	Undisclosed habit of tobacco use
alcohol_consumption_No	No alcohol consumption
alcohol_consumption_Former	History of alcohol consumption
alcohol_consumption_Yes	Current habit of alcohol consumption
alcohol_consumption_Not informed	Undisclosed habit of alcohol consumption
sun_exposure_No	No abnormal sun exposure
sun_exposure_Yes	Abnormal sun exposure
sun_exposure_Not informed	Sun exposure (not informed)
gender_M	Male sex
gender_F	Female sex
age_group_2	Age older than 60 years
age_group_1	Age between 41 and 60 years
age_group_0	Age younger than 40 years

Table 2. Class distribution in dataset.

Class Names	Sample Counts
Leukoplakia without dysplasia	57
Leukoplakia with dysplasia	89
Oral squamous cell carcinoma	91

2.2. Data Preprocessing

Initial model building was first carried out. Missing values were addressed using a 'forward-fill' method that moves the last valid observation forward, effectively filling gaps with the most recent available data. This method ensured that the dataset remained complete and usable while avoiding arbitrary imputation values. The columns without any features were removed, which removed any unnecessary information that could impair the model's performance. Following the feature selection, the remaining features were scaled with StandardScaler [19,20]. Scaling standardised the range of feature values, resulting in a mean of zero and a standard deviation of one [21]. Normalisation increases the convergence and accuracy of machine learning algorithms because it ensures that all features contribute equally to the model.

The dataset was then divided into training and testing sets to assess the models' performance. The training set consisted of 177 samples, while the testing set contained

60. A stratified split was used to keep the class distribution consistent across both sets, preserving the balance of each class during the training and testing stages.

2.3. Machine Learning Models Used

This study compared five machine learning models: Support Vector Machine (SVM) [22], Random Forest [8], K-Nearest Neighbours (KNN) [23], Naive Bayes [24], and Logistic Regression [25]. Anaconda Navigator enabled the implementation of these models in Python within Jupyter Notebook (Anaconda, Inc., Austin, TX, USA).

2.3.1. K Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a non-parametric classification algorithm that classifies samples using the majority vote of their nearest neighbours. This algorithm is simple and effective, making predictions based on the proximity of samples in the feature space. The KNN model was built with the KNeighboursClassifier from the sklearn.neighbours module and a number of neighbours (k) set to 5 [23]. The model's performance was evaluated by training and testing it on the dataset using default distance metrics.

2.3.2. Logistic Regression

Logistic Regression is a statistical model for binary classification that uses a logistic function to estimate probabilities. It calculates the probability of a binary outcome using one or more predictor variables. The Logistic Regression model was implemented with the 'LogisticRegression' class from the sklearn.linear_model module and the 'liblinear' solver [26]. The model was trained on the dataset and evaluated using the default settings to determine its predictive performance.

2.3.3. Naive Bayes

Naive Bayes is a probabilistic classifier that uses Bayes' theorem and assumes that features are independent. Despite its simplicity and strong independence assumptions, Naive Bayes is efficient and scalable, making it suitable for various classification tasks. In this study, the Naive Bayes model was implemented using the sklearn.naive_bayes module's GaussianNB class [27]. The model was trained and evaluated using the default settings to determine the classification accuracy.

2.3.4. Support Vector Machines (SVM)

The Support Vector Machine (SVM) is a robust supervised learning algorithm for classification and regression tasks. It works by determining the best hyperplane to maximise the margins among different classes in the feature space. SVM's ability to handle high-dimensional data makes it a popular choice for various classification tasks. In this study, the SVM model was implemented using the SVC class from the sklearn.svm module, with a radial basis function (RBF) kernel to handle non-linear classification tasks [22]. The model was trained on the training set and then evaluated on the testing set.

2.3.5. Random Forest

Random Forest is an ensemble learning technique that builds multiple decision trees and aggregates their predictions to improve classification accuracy while minimising overfitting. This method uses the combined power of various trees to create a robust model with improved predictive performance. In this study, the Random Forest model was implemented with the RandomForestClassifier from the sklearn.ensemble module. The model used 100 trees, with the hyperparameters set to the default values to make the process easier [8]. The trained model's performance was assessed using the testing set.

2.4. Evaluation Metrics

To evaluate the performance of the machine learning models in this study, various evaluation metrics were used to provide a complete picture of their effectiveness. These metrics included the accuracy, precision, recall, F1 score, and ROC-AUC score.

Accuracy measures the proportion of correctly classified instances out of the total number of instances. It is calculated as

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

However, accuracy alone can be misleading, particularly in class imbalance, as it may not fully reflect the model's effectiveness in distinguishing among classes with different sample sizes. Therefore, the proportion of correctly predicted positive observations to total predicted positives (precision) was used. It measures the model's ability to prevent false positives. The formula for precision is

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The ratio of correctly predicted positive observations to total observations in the class (sensitivity or recall) was then assessed. It assesses each model's ability to identify all relevant instances in a class. The formula for recall is

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 score is the harmonic mean of precision and recall, providing a metric that balances both aspects. It is beneficial when working with imbalanced datasets in which one class is underrepresented. The F1 score is calculated as

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Finally, the area under the receiver operating characteristic (ROC) curve was assessed, which compares the true positive rate (recall) to the false positive rate (1—specificity) at different threshold settings. This score measures the model's ability to distinguish among classes at various thresholds, with higher scores indicating better performance. The ROC-AUC score is computed as follows:

$$\text{ROC-AUC Score} = \int_0^1 \text{ROC Curve dFalse Positive Rate}$$

Here, dFalse Positive Rate is a differential element representing a small change in the false positive rate. In calculus, the differential dFalse Positive Rate is used to represent an infinitesimally small increment in the false positive rate. This integration process sums up the area under the ROC curve by accounting for these tiny changes in the false positive rate, ultimately providing a measure of the model's performance in distinguishing among classes across various threshold values.

3. Results

The five machine learning models—K-Nearest Neighbours (KNN), Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest—were assessed across performances for predicting leukoplakia, dysplastic changes, and OSCC. An overall performance was then derived from the three separate assessments.

3.1. Model Performances across the Board

3.1.1. Model Performance for Leukoplakia without Dysplasia

Random Forest outperformed the other models with perfect precision (1.0), high recall (0.78), an excellent F1-Score of 0.88, and the highest ROC AUC score of 0.99. (Table 3) SVM and Logistic Regression delivered balanced results with reasonable precision and recall. KNN produced moderate results but did not excel. Despite perfect recall, Naive Bayes had low precision (0.43), resulting in a high false positive rate and lower reliability. Thus, Random Forest was the best model for predicting this class, while SVM and Logistic Regression were useful but could be improved, and Naive Bayes was less reliable due to high false positives.

Table 3. Predictive performance of machine learning models for leukoplakia without dysplasia.

Models	Precision	Recall	F1 Score	ROC
KNN	0.75	0.67	0.71	0.86
Logistics Regression	0.76	0.72	0.74	0.92
Naive Bayes	0.43	1	0.6	0.86
SVM	0.86	0.67	0.75	0.76
Random Forest	1	0.78	0.88	0.99

3.1.2. Model Performance for Leukoplakia with Dysplasia

Random Forest outperformed all other models, with perfect recall (1.0), high precision (0.91), an excellent F1 score of 0.95, and a top ROC-AUC score of 0.99 (Table 4). SVM achieved high precision (1.0) but had slightly lower recall (0.81), implying that it missed some true positives. KNN and Logistic Regression performed reasonably well but were not exceptional. Despite its high precision (0.90), Naive Bayes had very low recall (0.43), which resulted in many false negatives and decreased reliability. Overall, Random Forest was the most effective model, with SVM, KNN, Logistic Regression, and Naive Bayes having different limitations.

Table 4. Predictive performance of machine learning models for leukoplakia with dysplasia.

Models	Precision	Recall	F1 Score	ROC
KNN	0.75	0.86	0.8	0.94
Logistics Regression	0.7	0.67	0.68	0.85
Naive Bayes	0.9	0.43	0.58	0.81
SVM	1	0.81	0.89	0.94
Random Forest	0.91	1	0.95	0.99

3.1.3. Model Performance for Oral Squamous Cell Carcinoma

Random Forest performed the best, with perfect recall (1.0), high precision (0.91), an F1 score of 0.95, and a ROC-AUC score of 1.0, indicating outstanding overall performance (Table 5). SVM had higher recall (0.95) but a lower precision (0.69), resulting in a lower F1 score than Random Forest. KNN and Logistic Regression demonstrated a balanced but moderate performance. Naive Bayes struggled with a low recall (0.33%) and poor overall performance.

Table 5. Predictive performance of machine learning models in oral squamous cell carcinoma.

Models	Precision	Recall	F1 Score	ROC
KNN	0.85	0.81	0.83	0.95
Logistics Regression	0.74	0.81	0.77	0.93
Naive Bayes	0.88	0.33	0.48	0.75
SVM	0.69	0.95	0.8	0.93
Random Forest	0.91	1	0.95	1

3.1.4. Overall Performance across All Three Categories

Random Forest outperformed all other models with an accuracy of 0.93 and a Kappa score of 0.90 (Table 6). SVM followed, with an accuracy of 0.82 and a Kappa score of 0.72, indicating strong performance. KNN and Logistic Regression produced moderate results; KNN had an accuracy of 0.78 and a Kappa score of 0.67, and Logistic Regression had an accuracy of 0.73 and a Kappa score of 0.60. Naive Bayes had the lowest scores, with an accuracy of 0.57 and a Kappa Score of 0.37.

Table 6. The overall predictive performance of machine learning models.

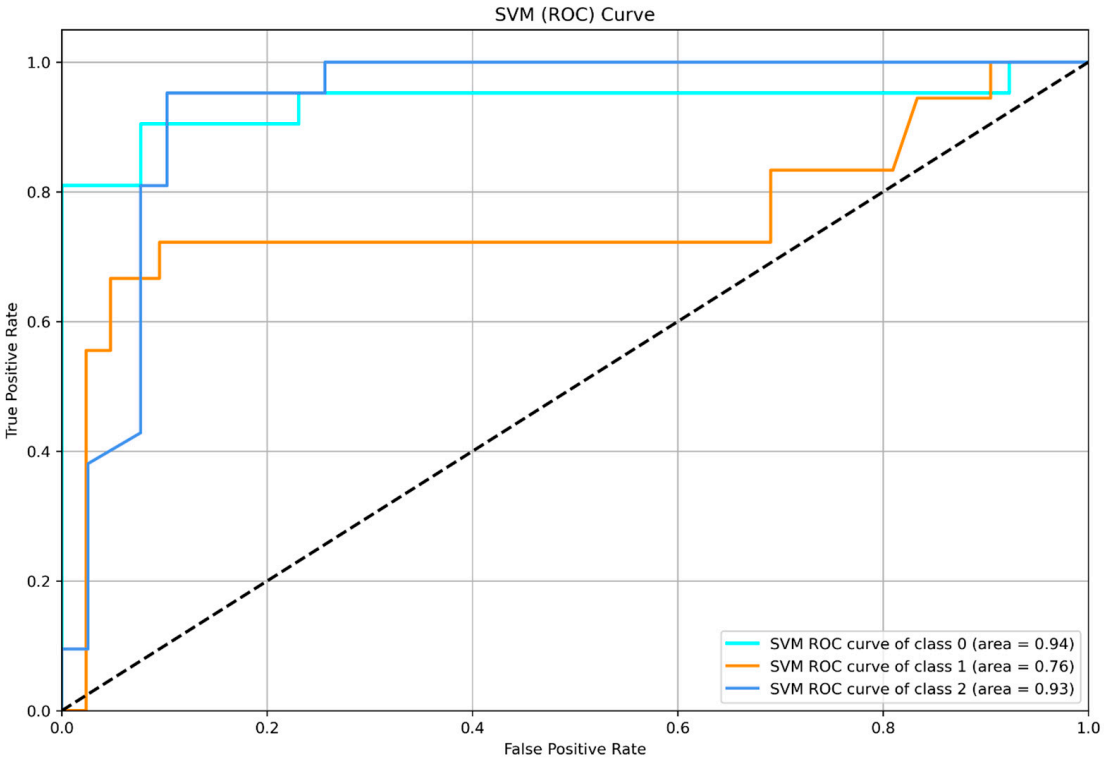
Models	Overall Accuracy	Kappa Score
KNN	0.78	0.67
Logistics Regression	0.73	0.6
Naive Bayes	0.57	0.37
SVM	0.82	0.72
Random Forest	0.93	0.9

3.2. Receiver Operating Characteristics of the Models

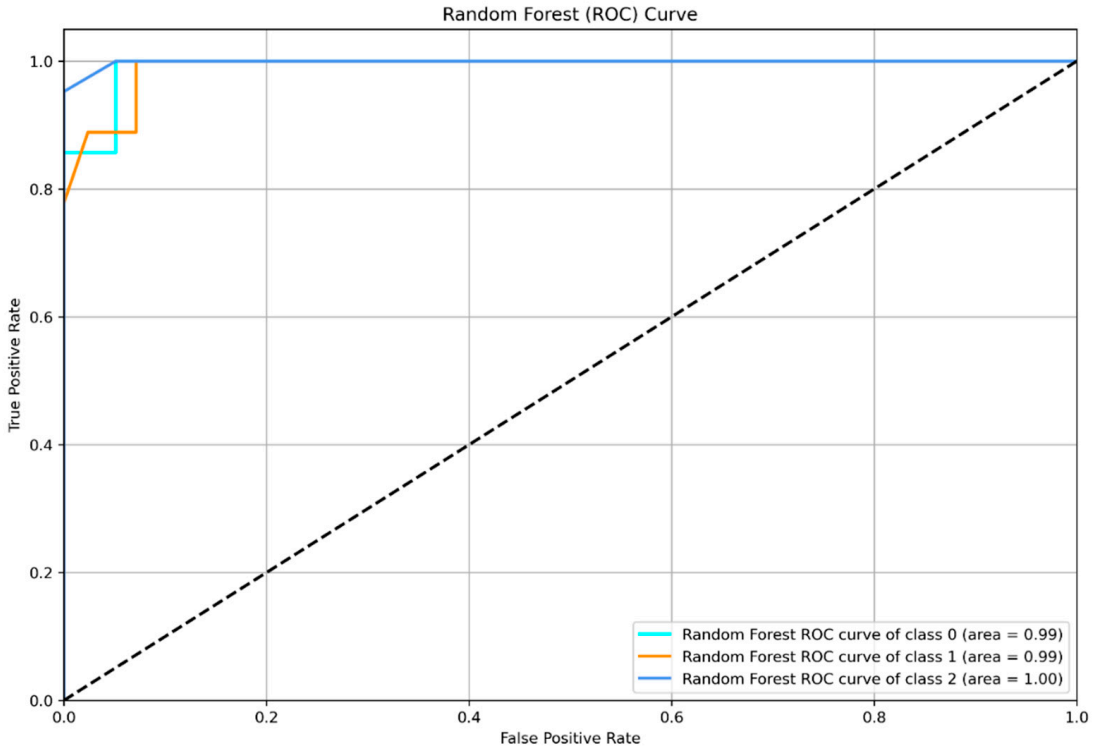
The SVM model (Figure 1A) performed exceptionally well at distinguishing between leukoplakia without dysplasia and OSCC, with AUC values of 0.94 and 0.93, respectively, but performed poorly in diagnosing dysplastic changes, with an AUC of 0.76. Random Forest (Figure 1B) outperformed all other models, with near-perfect AUC values of 0.99, 0.99, and 1.00, demonstrating its exceptional ability to classify positive instances while minimising false positives correctly. Naive Bayes (Figure 1C) performed moderately, with AUCs of 0.81, 0.86, and 0.75 indicating variability in class discrimination. Logistic Regression (Figure 1D) also performed well, with AUCs of 0.85, 0.92, and 0.93, while KNN (Figure 1E) produced balanced results (AUCs of 0.94, 0.86, and 0.95).

3.3. Confusion Matrices

The K-Nearest Neighbours (KNN) model outperformed classes 0 and 2, with higher true positive rates of 18 and 17, respectively. For leukoplakia with dysplasia, it performed more poorly, with only 12 correct predictions. Misclassifications were observed, with leukoplakia without dysplasia often being confused with dysplastic changes, as indicated by the off-diagonal values. The confusion matrix for Logistic Regression indicated balanced performance across all three classes. Classes 0, 1, and 2 demonstrated true positive counts of 14, 13, and 17, respectively. This model appeared to perform relatively uniformly, though errors occurred in classifying dysplastic changes similar to the previous model. The confusion matrix for Naive Bayes showed that it performed well for class 1 (18 true positives), but the model struggled to distinguish between basic leukoplakia and OSCC. The SVM confusion matrix showed good performance for OSCC (20 true positives), but some confusion occurred between the two classes of leukoplakia. The Random Forest model produced true positive counts of 21, 14, and 21 for classes 0, 1, and 2, respectively. The confusion matrix demonstrated minimal off-diagonal values, indicating that the model effectively minimised misclassifications. Figure 2 demonstrates the confusion matrices for all models evaluated.

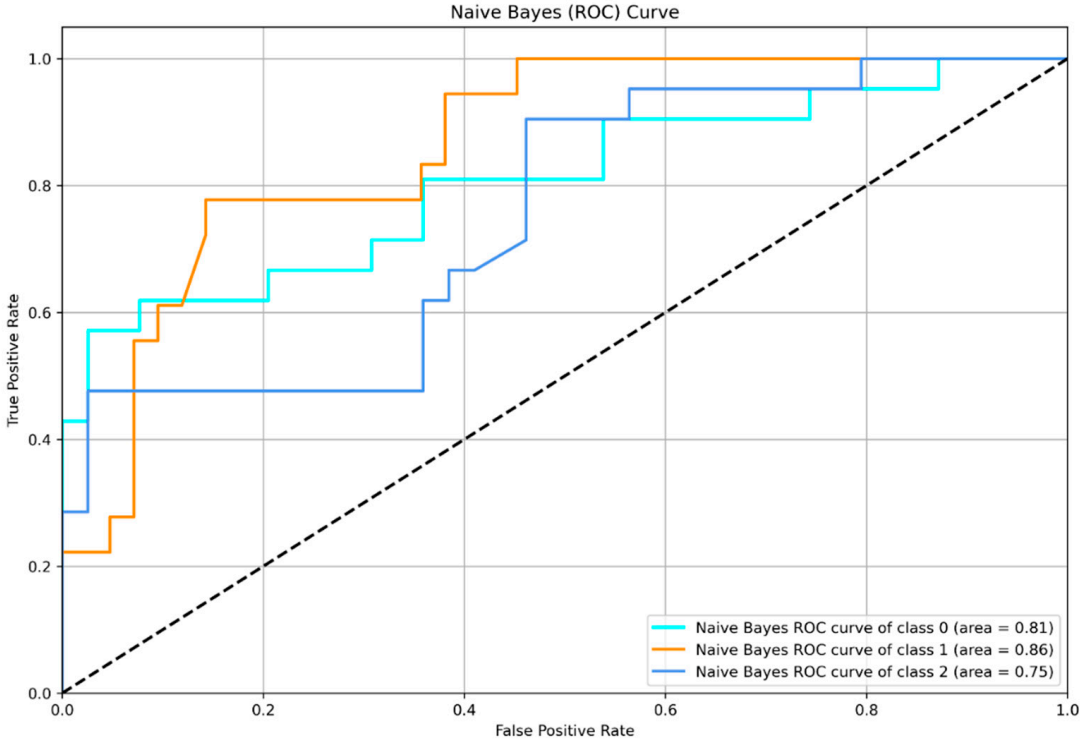


(**A**)

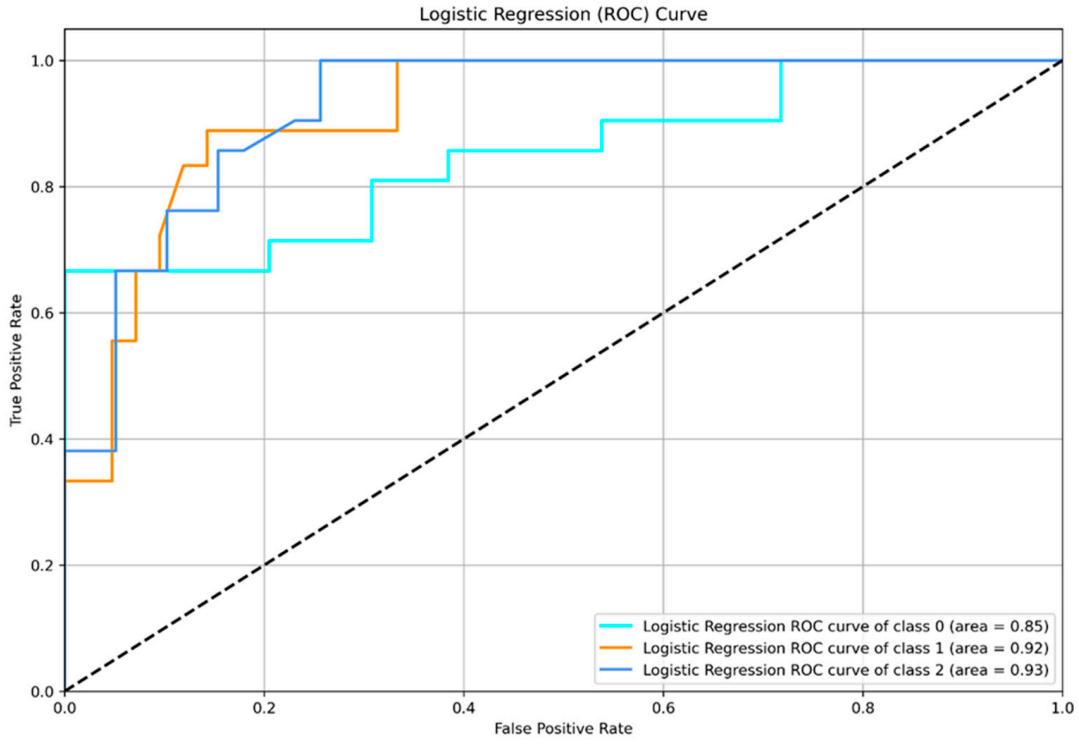


(**B**)

Figure 1. *Cont.*

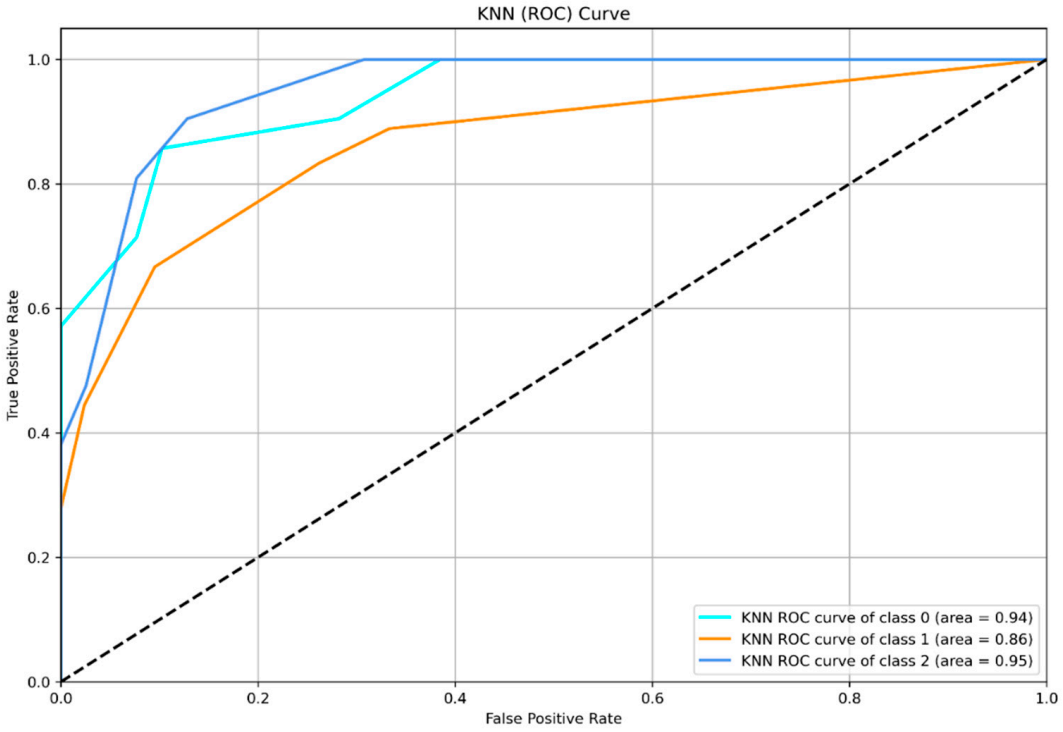


(**C**)



(**D**)

Figure 1. *Cont.*



(**E**)

Figure 1. (**A**). ROC graph for Support Vector Machine model. (**B**). ROC graph for Random Forest model. (**C**). ROC graph for Naïve Bayes model. (**D**). ROC graph for Logistic Regression. (**E**). ROC graph for K-Nearest Neighbour model.

3.4. Shapley Additive Explanations (SHAP) for Model Interpretability

The influence of each parameter in the decision-making capabilities of the ML models were assessed through SHAP values illustrated through bar charts. The length of each bar represents the feature's importance, with longer bars indicating a more substantial impact.

3.4.1. Interpretability of Leukoplakia without Dysplasia

Gingival localisation of lesion emerges as the most significant feature in this category (Figure 3A). Interestingly, larger lesions were also correlated with the class.

3.4.2. Interpretability of Leukoplakia with Dysplasia

Among the features analysed (Figure 3B), buccal mucosal lesions were the most influential factor in the models' decision-making processes, followed by the patients' ages, with individuals over 60 years showing a higher probability of positive predictions for dysplastic changes.

3.4.3. Interpretability of Oral Squamous Cell Carcinoma

Gingival localisation was the most important differentiating factor when the models distinguished between OSCC (Figure 3C) and leukoplakia. This was followed by the presence of lesions on the floor of the mouth, buccal mucosa, and tongue.

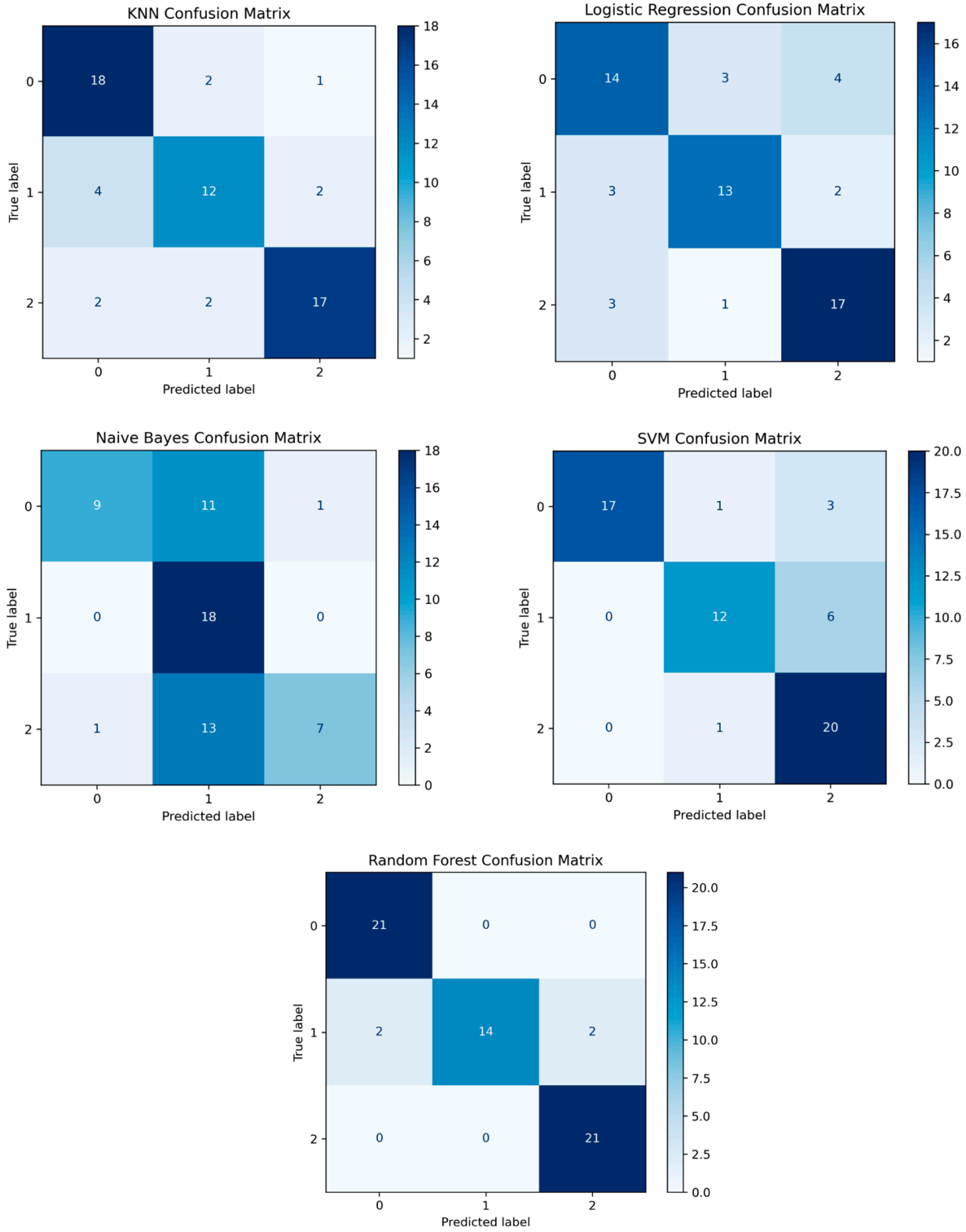
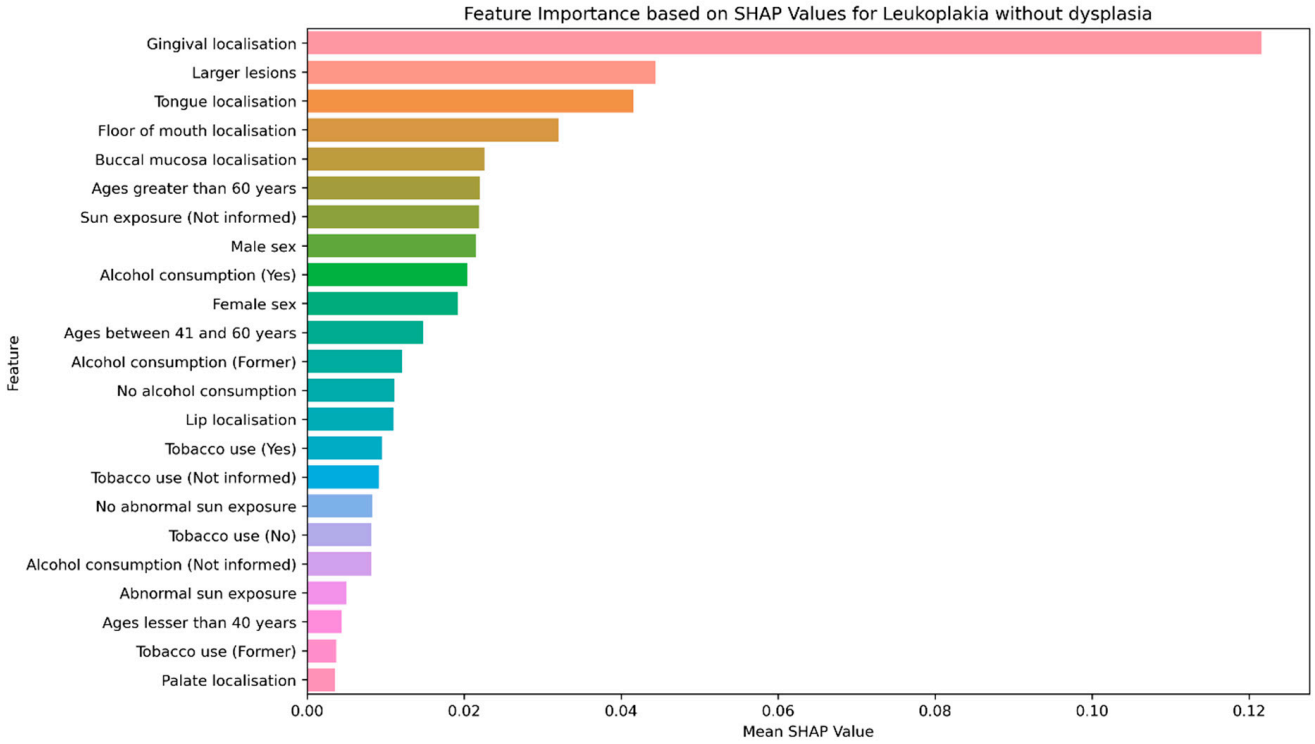
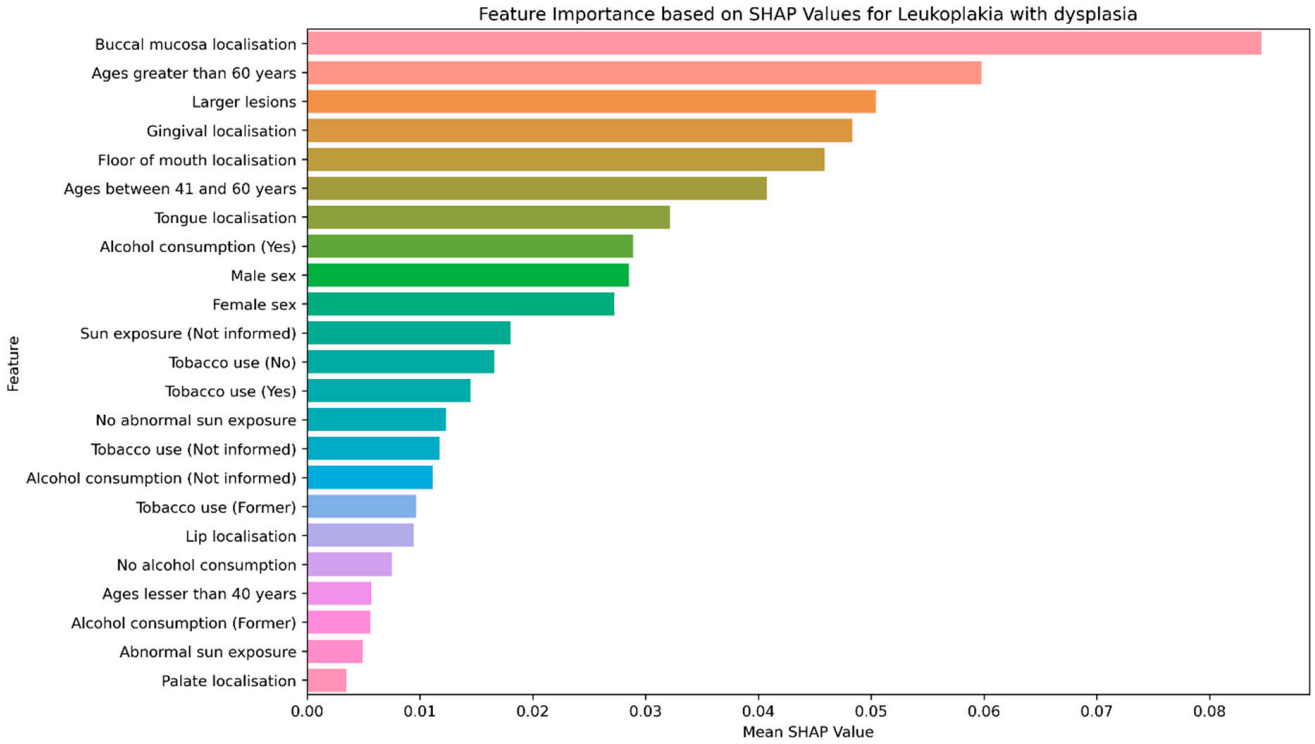


Figure 2. Confusion matrices for the machine learning models.

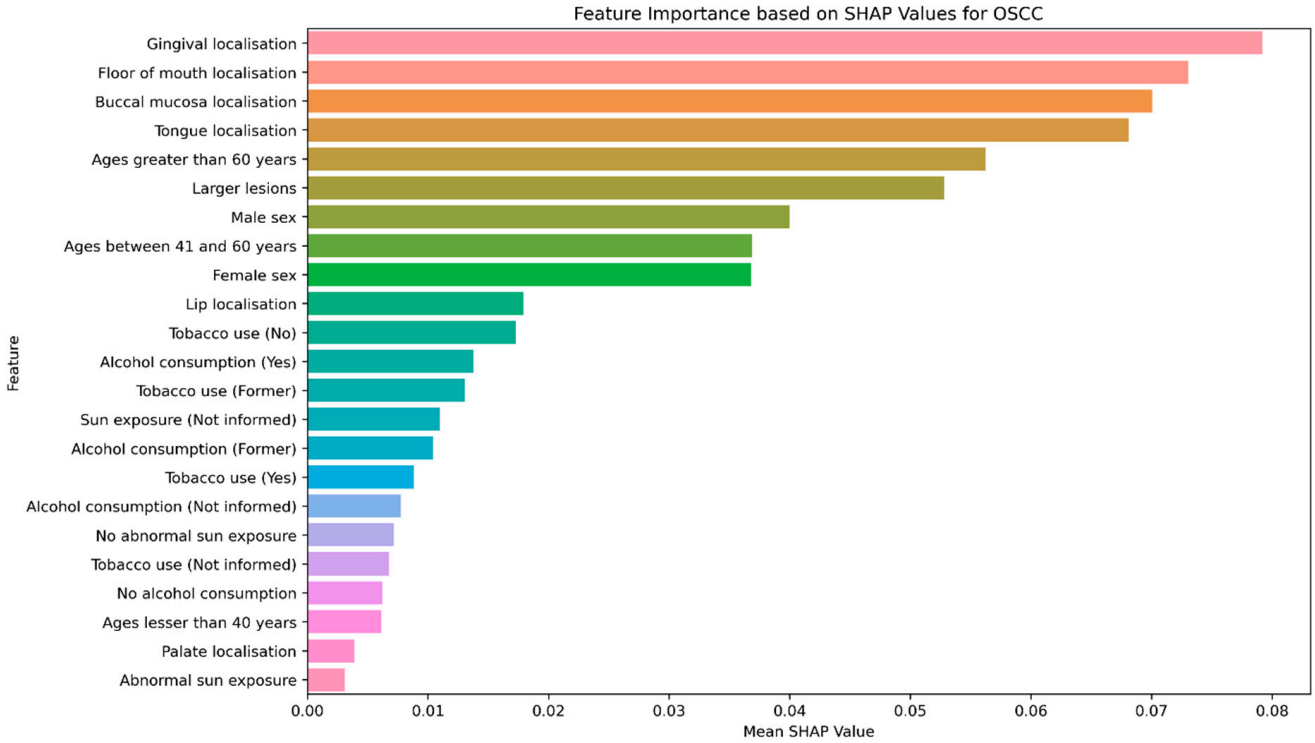


(**A**)



(**B**)

Figure 3. *Cont.*



(**C**)

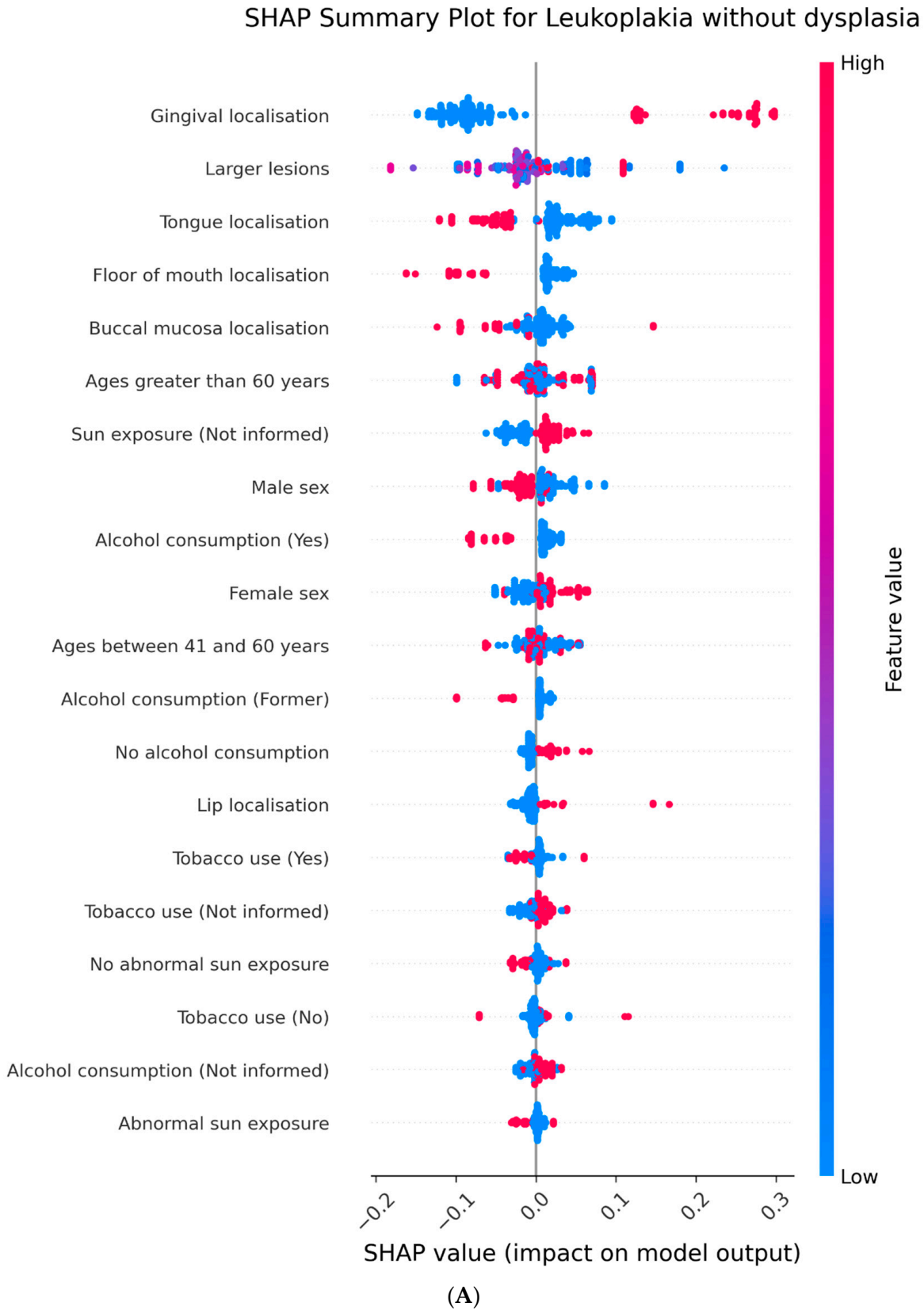
Figure 3. (**A**). SHAP value illustration for models' explainability when assessing leukoplakia without dysplasia. (**B**). SHAP value illustration for models' explainability when assessing leukoplakia with dysplasia. (**C**). SHAP value illustration for models' explainability when assessing oral squamous cell carcinoma.

4. Discussion

This research aimed to analyse known trends in the underlying features contributing to common white lesions, such as leukoplakia and OSCC, by extracting explanations from the best-performing ML models for their decisions. Previous research has emphasised the importance of model interpretability in healthcare applications, where understanding the reasoning behind predictions is essential for making informed clinical decisions and improving clinical reliance on these models [28].

To put the overall findings into a broader perspective, summary plots were constructed from the SHAP values to better reflect the explanations provided by SHAP while considering potential errors observed through the confusion matrices. The horizontal axis displays the SHAP values, which measure the contributions of each feature to the prediction of leukoplakia or OSCC. Positive SHAP values indicate that a feature's value has increased the likelihood of predicting leukoplakia or OSCC, while negative values suggest a reduced likelihood.

Leukoplakia without dysplasia was initially explored. Gingival localisation emerged as the most meaningful feature by which the model differentiated dysplastic changes, with the model ranking gingival lesions as less likely to cause malignant lesions. This aligns with the literature indicating that keratinised gingival tissue has a comparatively lower likelihood of producing neoplastic changes [29]. The model struggled with identifying whether the larger size of the lesion was a potential indicator, similar to human practitioners, who, in these cases, look for other symptoms of carcinoma in situ on a histopathological level [30]. Notably, localisation to the tongue suggested to the model that the patient might be at a higher risk of developing potential malignancies, and it relied on this metric to modify its classification for a case. A detailed assessment is illustrated in Figure 4A.

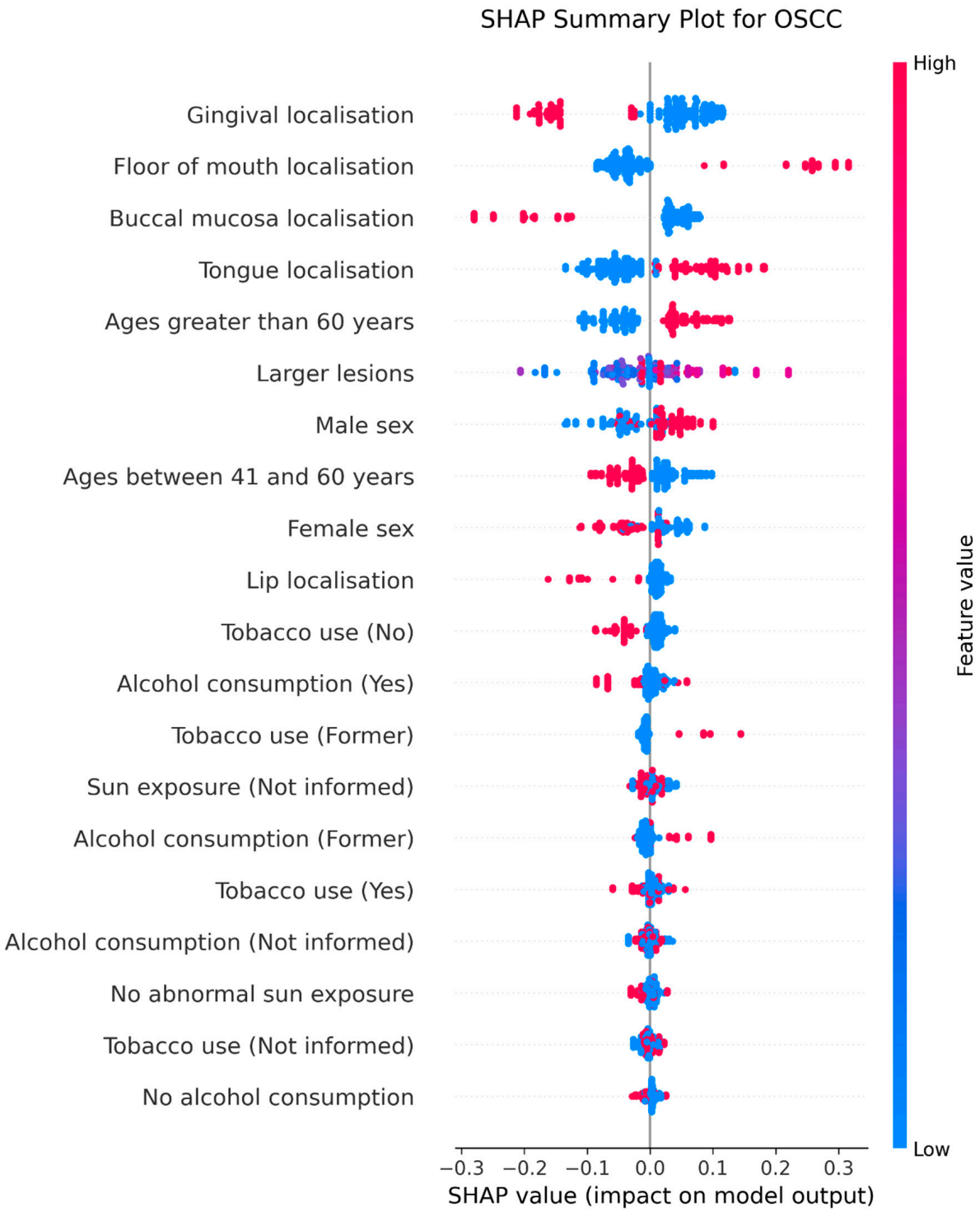


(**A**)

Figure 4. *Cont.*

(**B**)

Figure 4. *Cont.*



(**C**)

Figure 4. (**A**). Summary of interpretability for leukoplakia without dysplasia. (**B**). Summary of interpretability for leukoplakia with dysplasia. (**C**). Summary of interpretability for oral squamous cell carcinoma.

Dysplastic changes were then evaluated using summary plots (Figure 4B), with the models relying more heavily on the incidence of lesions being present on the buccal mucosa, indicating the possibility of malignant changes. This was the first time the model prioritised age over other factors, with individuals over 60 years being less likely to be categorised at this stage. Combined with the confusion the model experienced when evaluating age

in cases without dysplastic changes, the authors suggest this may be because additional health complications in individuals over 60 years could complicate diagnoses without an extensive medical history included in the potential predictor variables. This theory reemphasises the need for standardised screening and record documentation practices in general dentistry and requires further investigation [31]. This approach is similar to dental screening and triage practices, where multiple factors are considered prior to assigning individuals to a risk group with treatment needs [32]. At this stage, the model begins to consider social habits, such as smoking and alcohol intake, more heavily. Interestingly, the model determined that a lesion on the floor of the mouth was highly unlikely to be dysplastic and instead rated it as having one of the highest probabilities of being an OSCC, as discussed in the next paragraph.

Finally, the cases of OSCC (Figure 4C) were explored. As described previously, the model identified certain features as tell-tale signs that a lesion is more likely to be malignant, as indicated by the shift to positive SHAP values. These features are primarily based on the location of the lesion, such as the floor of the mouth, buccal mucosa, and the tongue. Notably, the model placed more emphasis on patient data showing a history of heavy alcohol consumption rather than an existing history of this habit. The fundamental rationale behind this decision remains unclear. It can be theorised that the model considers the possibility of accumulated damage and residual effects from long-term use of tobacco and alcohol, which may have manifested symptoms only after the cessation of these agents, allowing the body's metabolic process to undergo and possibly complete the withdrawal process [33]. Another interesting observation is that in almost all cases, the model disregarded lesions on the lip, which aligns with clinical diagnoses that favour basal cell carcinomas as occurring more frequently on the lips than OSCC [34].

The current study effectively compared machine learning models for diagnosing oral lesions, with Random Forest performing best and SHAP enhancing interpretability. However, this study was limited by the retrospective dataset of 237 samples, which may restrict the generalisability of the findings due to its size and lack of demographic and geographic diversity. While key features, such as lesion size and location, were identified, other potentially relevant factors, such as genetic predispositions or detailed histopathological data were not considered. This in turn contributed to the difficulties experienced by the models in classifying dysplastic changes. Although the analysis based on SHAP values is informative, it may oversimplify complex feature interactions and fail to address potential class imbalances that could affect model performance. This is particularly important when classifying white lesions beyond leukoplakia, extending to lesions like lichen planus, which have several variations with differing degrees of carcinogenic potential. Additionally, the study lacks external validation from data obtained from other clinics or hospitals, which might better represent real-world variations in patient distribution [9]. Future studies with larger and more diverse datasets incorporating additional features may provide more comprehensive insights. Incorporating techniques for managing class imbalances, conducting external validations, and delving deeper into feature interactions would improve the reliability and generalisability of predictions. Considering temporal and environmental factors may also enhance the model's applicability over time.

5. Conclusions

The following conclusions can be drawn from the current study:

1. The Random Forest model achieved the highest performance with an overall accuracy of 93%, showing superior class-specific precision, recall, and F1 scores for both OSCC and various types of leukoplakia.
2. SHAP (SHapley Additive exPlanations) analysis identified the top predictors influencing the model's decisions. For leukoplakia with dysplasia, these included buccal mucosa localisation, an age over 60 years, and lesion size. For leukoplakia without dysplasia, the key predictors were gingival and tongue localisation, along with lesion size. For OSCC, gingival, floor-of-mouth, and buccal mucosa localisations were the

most influential. The model notably indicated that lesions on the floor of the mouth were highly unlikely to be dysplastic, instead showing one of the highest probabilities for being OSCC.

Author Contributions: Conceptualisation: S.S.A., S.A. and T.H.F. Methodology: S.S.A. and S.A. Investigation: S.S.A. and S.A. Resources: T.H.F. and J.D. Writing—Original Draft: S.S.A. and S.A. Writing—Review and Editing: T.H.F. and J.D. Supervision: S.A. Project Administration: T.H.F. Funding acquisition: J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the University of Adelaide Kwok Paul Lee Bequest (350-75134777).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All codes and relevant information are provided online at https://github.com/Salem1901/Predicting-leukoplakia-and-oral-squamous-cell-carcinoma-using-interpretable-machine-learning.-.git (last accessed on 8 August 2024).

Acknowledgments: Generative AI (ChatGPT; OpenAI, San Francisco, USA) was used for English proofreading.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organisation. Cancer. 2020. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 9 August 2024).
2. Dhanuthai, K.; Rojanawatsirivej, S.; Thosaporn, W.; Kintarak, S.; Subarnbhesaj, A.; Darling, M.; Kryshtalskyj, E.; Chiang, C.P.; Shin, H.I.; Choi, S.Y.; et al. Oral cancer: A multicenter study. *Med. Oral Patol. Oral Cir. Buccal* **2018**, *23*, e23–e29. [CrossRef] [PubMed]
3. Di Spirito, F.; Di Palo, M.P.; Folliero, V.; Cannata, D.; Franci, G.; Martina, S.; Amato, M. Oral bacteria, virus and fungi in saliva and tissue samples from adult subjects with Oral squamous cell carcinoma: An umbrella review. *Cancers* **2023**, *15*, 5540. [CrossRef] [PubMed]
4. Warnakulasuriya, S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol.* **2009**, *45*, 309–316. [CrossRef] [PubMed]
5. Neville, B.W.; Damm, D.D.; Allen, C.M.; Bouquot, J.E. *Oral and Maxillofacial Pathology*; WB Saunders: Philadelphia, PA, USA, 2002.
6. Farook, T.H.; Haq, T.M.; Ramees, L.; Dudley, J. Predicting masticatory muscle activity and deviations in mouth opening from non-invasive temporomandibular joint complex functional analyses. *J. Oral Rehabil.* **2024**, *51*, 1770–1777. [CrossRef]
7. Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
8. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
9. Farook, T.H.; Dudley, J. Automation and deep (machine) learning in temporomandibular joint disorder radiomics. A systematic review. *J. Oral Rehabil.* **2023**, *50*, 501–521. [CrossRef]
10. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
11. Mahmood, H.; Shaban, M.; Indave, B.I.; Santos-Silva, A.R.; Rajpoot, N.; Khurram, S.A. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. *Oral Oncol.* **2020**, *110*, 104885. [CrossRef]
12. Adeoye, J.; Koohi-Moghadam, M.; Lo, A.W.I.; Tsang, R.K.-Y.; Chow, V.L.Y.; Zheng, L.-W.; Choi, S.-W.; Thomson, P.; Su, Y.-X. Deep learning predicts the malignant-transformation-free survival of oral potentially malignant disorders. *Cancers* **2021**, *13*, 6054. [CrossRef]
13. Kutlu, H.; Avcı, E. A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transform and long short-term memory networks. *Sensors* **2019**, *19*, 1992. [CrossRef] [PubMed]
14. Farook, T.H.; Haq, T.M.; Ramees, L.; Dudley, J. Predictive modelling of freeway space utilising clinical history, normalised muscle activity, dental occlusion, and mandibular movement analysis. *Sci. Rep.* **2024**, *14*, 16423. [CrossRef] [PubMed]
15. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
16. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017; Volume 30, p. 30.

17. Norgeot, B.; Quer, G.; Beaulieu-Jones, B.K.; Torkamani, A.; Dias, R.; Gianfrancesco, M.; Arnaout, R.; Kohane, I.S.; Saria, S.; Topol, E.; et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat. Med.* **2020**, *26*, 1320–1324. [CrossRef] [PubMed]
18. Ribeiro-de-Assis, M.C.F.; Soares, J.P.; de Lima, L.M.; de Barros, L.A.P.; Grão-Velloso, T.R.; Krohling, R.A.; Camisasca, D.R. NDB-UFES: An oral cancer and leukoplakia dataset composed of histopathological images and patient data. *Data Brief* **2023**, *48*, 109128. [CrossRef]
19. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112.
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
21. Farook, T.H.; Jamayet, N.B.; Abdullah, J.Y.; Alam, M.K. Machine learning and intelligent diagnostics in dental and orofacial pain management: A systematic review. *Pain Res. Manag.* **2021**, *2021*, 6659133. [CrossRef]
22. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
23. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
24. Duda, R.O.; Hart, P.E. *Pattern Classification and Scene Analysis*; Wiley: New York, NY, USA, 1973; Volume 3.
25. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
26. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1958**, *20*, 215–232. [CrossRef]
27. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; Volume 4.
28. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv* **2017**, arXiv:1712.09923.
29. Keshava, A.; Gugwad, S.; Baad, R.; Patel, R. Gingival squamous cell carcinoma mimicking as a desquamative lesion. *J. Indian Soc. Periodontol.* **2016**, *20*, 75–78. [CrossRef] [PubMed]
30. Farah, C.S.; Fox, S.A. Dysplastic oral leukoplakia is molecularly distinct from leukoplakia without dysplasia. *Oral Dis.* **2019**, *25*, 1715–1723. [CrossRef] [PubMed]
31. Razzaki, S.; Baker, A.; Perov, Y.; Middleton, K.; Baxter, J.; Mullarkey, D.; Sangar, D.; Taliercio, M.; Butt, M.; Majeed, A.; et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv* **2018**, arXiv:1806.10698.
32. Kirton, J.A.; Thompson, W.; Pearce, M.; Brown, J.M. Ability of the wider dental team to triage patients with acute conditions: A qualitative study. *Br. Dent. J.* **2020**, *228*, 103–107. [CrossRef] [PubMed]
33. Liakoni, E.; Edwards, K.C.; St Helen, G.; Nardone, N.; Dempsey, D.A.; Tyndale, R.F.; Benowitz, N.L. Effects of nicotine metabolic rate on withdrawal symptoms and response to cigarette smoking after abstinence. *Clin. Pharmacol. Ther.* **2019**, *105*, 641–651. [CrossRef]
34. Loh, T.; Rubin, A.G.; Jiang, S.I.B. Management of mucosal basal cell carcinoma of the lip: An update and comprehensive review of the literature. *Dermatol. Surg.* **2016**, *42*, 1313–1319. [CrossRef]