International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Predictive Model for Early Detection of Oral Cancer via Adaboost Classification Technique

Dr.K.Padmavathi[1], C.Deepa[2]

[1]Associate Professor, Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu, India
[2]Assistant Professor, Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu, India

Abstract
Oral cancer is a serious and potentially life-threatening disease that affects millions of people around the world. Early detection is critical for improving treatment outcomes and reducing mortality rates. This study aims to develop a predictive model for early detection of oral cancer using the AdaBoost classification technique. A dataset containing various risk factors associated with oral cancer was used to train and test the model. The results show that the AdaBoost algorithm was able to accurately classify oral cancer patients and non-cancer patients with high precision and recall rates. The developed predictive model could be used as a tool for early detection of oral cancer, thus improving patient outcomes and reducing mortality rates. The significance of oral cancer prediction and classification by improvised AdaBoost proposed technique produce potential to improve patient outcomes, facilitate early detection and treatment of oral cancer, and increase the efficiency and accuracy of the diagnostic process.

Keywords: AdaBoost algorithm, Classification, Early detection, Improvised AdaBoost technique, Oral Cancer, Oral Cancer Dataset.

1. Introduction

Oral cancer is a major public health issue that affects millions of people worldwide. It is the sixth most common cancer globally, with an estimated incidence rate of 355,000 new cases per year. The mortality rate of oral cancer is high, with more than 170,000 deaths annually. Early detection of oral cancer is critical for improving treatment outcomes and reducing mortality rates. However, current screening methods are often inadequate, leading to delayed diagnoses and poor outcomes. Therefore, there is a need for an accurate and efficient screening tool for early detection of oral cancer.

Hence, Adaboost (Adaptive Boosting) is a popular machine learning algorithm used for classification tasks. When it comes to predicting oral cancer, the proposed predictive model techniques are used in conjunction with Adaboost in improvisation of the model accuracy.

In the research of FatihahMohd , it is aimed to predict the early stages of oral cancer using various machine learning methods, including Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors, and Support Vector Machine. They found that analyzing oral cavities improved the accuracy of classification[3]. Similarly, Ahmad LG aimed to develop a model for clinicians to detect breast cancer recurrence using tree-based decision-making methods, artificial neural networks, and high-accuracy data

analysis techniques such as NN and HDM. The ADMS rating representation was found to be the most effective in achieving the highest accuracy and lowest error rate. SVM outperformed ANN and DT in their predictive capabilities[4].

HarikumarRajaguru and Sunil Kumar Prabhakar[5] aimed to assess the accuracy of TNM staging systems using Multi-Layer Perceptron (MLP) and Gaussian Mixed Models. Comparing the two ranking groups, they found that the average accuracy for the stage was better. They usedultraviolet scanning machines (LMMs) as post-data storage devices to analyze oral cancer and evaluated the efficiency of LMS classifiers against SMMs and MLPs[6]. Amy F. Z. B. Al. utilized the SVM classification method to detect OSCC tumors by examining patient expression and RNA extraction, as well as microscopic analysis, which showed a positive prognosis of OSCC tissue. Meanwhile, Marc Aubreville and colleagues aimed to evaluate new automation approaches for OSCC diagnostics using in-depth training and CNN methods on clear imaging. Their CNN approach involved searching for quotes, images, training data, and classification [7]

In their work, the author highlights the importance of the cancer detection process and presents various approaches for detecting different types of cancer. They describe how artificial techniques can help improve cancer detection and showcase their results in categorizing brain tumors using standardized sample images from UCI study data sets. Through their work, they were able to identify cancer issues in socio-economic populations with 87.2% accuracy using deep neural networks [8]. Duke Kumama et al. [9] suggested the use of hypothetical imaging for cancer discovery and evaluated several methods for cancer detection, highlighting the benefits of symptomatic simulations. They proposed using HSI for professional classification and performed categorization using a self-mapping structure and vector support machine. Meanwhile, Yuan et al. [10] and Dul et al. [11] showed the superiority of IT.P. IT. over other imaging methods such as MRI and CT for detecting cerebrovascular disease through in-depth training algorithms. Wang L. [12] discussed semi-automated methods for classifying cancer using in-depth learning algorithms and employed ordered power machines for categorization, while Kalantari et al. [13] utilized hyperspectral imaging to detect lung cancer and employed CNN for image segmentation. However, a fully autonomous design for cancer detection through in-depth learning techniques remains elusive due to the high system operating costs associated with advanced system configurations. This work overcomes these obstacles by providing a regression formation of the neural network for computerized cancer detection in existing medical imaging using this new in-depth learning method [14].

2. Methodology

2.1 Classification Technique

There are several classification techniques available for predicting oral cancer. Some of the commonly used techniques are:

- **Decision Trees**: A decision tree is a simple and effective classification technique that works by splitting the dataset into smaller subsets based on the values of different attributes. It is often used to identify the most important risk factors associated with oral cancer.
- **Support Vector Machines (SVM):** SVM is a machine learning technique that works by finding the hyperplane that best separates the data points in different classes. It is often used to classify oral cancer patients based on their risk factors and clinical features.

- **Random Forests**: A random forest is an ensemble of decision trees that work together to classify oral cancer patients. It is often used to identify the most important risk factors associated with oral cancer and to generate accurate predictions.
- **Neural Networks**: A neural network is a complex mathematical model that works by simulating the function of the human brain. It is often used to classify oral cancer patients based on their risk factors and clinical features.
- **AdaBoost**: AdaBoost is a machine learning technique that works by iteratively training a sequence of weak classifiers on the training data. It is often used to improve the accuracy of classification models for oral cancer prediction.

These classification techniques can be used alone or in combination with each other to develop accurate and reliable predictive models for oral cancer detection.

This proposed study uses the AdaBoost classification technique to develop a predictive model for early detection of oral cancer. The dataset used in this study contained various risk factors associated with oral cancer, including age, gender, tobacco and alcohol consumption, family history, and more. The dataset was divided into a training set and a testing set. The AdaBoost algorithm was used to build a classification model based on the training set. The testing set is used to evaluate the performance of the model, using metrics such as accuracy, precision, recall, and F1 score.

2.2 Significance of AdaBoost technique

The significance of oral cancer prediction and classification by AdaBoost lies in its potential to improve patient outcomes through early detection and treatment of oral cancer. Oral cancer is a serious and often fatal disease, but if caught early, it can be treated effectively with a high chance of survival. By using machine learning techniques such as AdaBoost to predict and classify oral cancer, clinicians can identify patients who are at high risk for the disease and take appropriate action, such as recommending further diagnostic tests or referring them to a specialist. This can lead to earlier detection of cancer and better treatment outcomes, as well as potentially saving lives.

Additionally, using AdaBoost for oral cancer prediction and classification can improve the efficiency and accuracy of the diagnostic process. With large amounts of patient data to sift through, it can be difficult for clinicians to accurately identify patients at risk for oral cancer. By using machine learning algorithms such as AdaBoost, clinicians can more easily and accurately analyze patient data and make informed decisions about diagnosis and treatment.

3. Predictive model

3.1 Architectural Study

Even the AdaBoost classification technique is a powerful tool for predicting oral cancer, the adaptive boosting mechanism and the adaptive learning rate, along with the feature selection mechanism, make Improvised AdaBoost more robust and accurate than the original AdaBoost.

The main modification done in the Improvised AdaBoost technique, as compared to the original AdaBoost, is the adaptive boosting mechanism. In AdaBoost, each sample in the training set is given an equal weight at the beginning of the training process. In each iteration, the weights are updated based on the accuracy of the previous iteration's classification. Samples that are misclassified are given a higher

weight, whereas samples that are correctly classified are given a lower weight. This way, the subsequent weak classifier focuses more on the misclassified samples, making the algorithm more adaptive to complex data distributions.

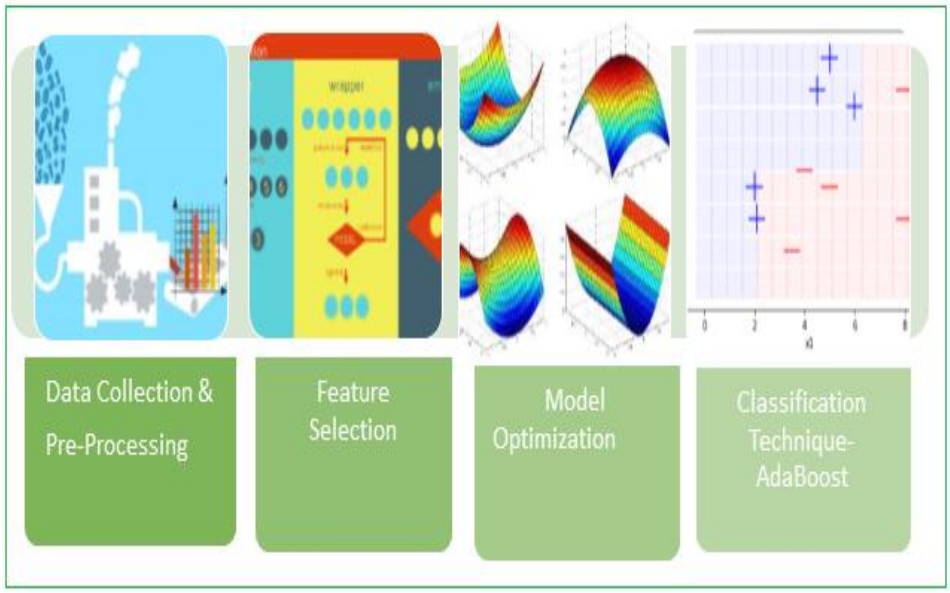
Improvised AdaBoost goes a step further by introducing an adaptive mechanism for learning rate. In AdaBoost, the learning rate is fixed throughout the training process. However, in Improvised AdaBoost, the learning rate is adjusted dynamically during the training process. The idea is to give more weight to the current weak classifier when it is accurate and less weight when it is not. This adaptive learning rate allows the algorithm to converge faster and achieve better classification accuracy.

Another modification in Improvised AdaBoost is the use of a feature selection mechanism. In AdaBoost, all features are used to train the weak classifiers, which can lead to overfitting and poor generalization. In Improvised AdaBoost, a feature selection mechanism is used to select the most informative features for classification. This reduces the number of features and improves the algorithm's performance by reducing the dimensionality of the input data.

The following steps makes accurate and reliable predictive models that is developed to aid in the early detection and treatment of this disease.

- **Data collection and pre-processing**: The first step in any machine learning project is to collect data. In the case of oral cancer prediction, data can be collected from various sources such as medical records, patient surveys, and clinical studies. In sequence to the data collection, it needs to be pre-processed to ensure that it is suitable for use in the machine learning model. This involves cleaning the data, removing any missing values, and encoding categorical variables.
- **Feature selection**: The next step is to select the most relevant features or risk factors associated with oral cancer. This can be done using techniques such as correlation analysis or feature importance analysis.
- **Model training and testing**: After selecting the relevant features, the AdaBoost algorithm is used to train the model on the training dataset. During the training process, the algorithm iteratively improves the performance of the weak classifiers by adjusting their weights. Once the model is trained, it is tested on a separate testing dataset to evaluate its performance. The accuracy, precision, recall, and F1 score of the model are calculated to determine its effectiveness in predicting oral cancer.
- **Model optimization**: The performance of the model can be improved by optimizing its hyper parameters. This involves adjusting the parameters of the AdaBoost algorithm to obtain the best possible results.
- **Deployment**: Finally, the optimized model can be deployed in a clinical setting for early detection of oral cancer. The model can be integrated with other clinical tools to improve patient outcomes and reduce mortality rates.

Figure1: Predictive Models for Oral Cancer Detection



3.2 Pre-Processing

Data Cleaning and Data normalization technique is used to scale the data to a specific range. This can be done using techniques such as Min-Max scaling. This technique involves removing errors, inconsistencies, and missing values from the dataset. In sequence, conversion of categorical variables into numerical values and done using techniques via label encoding are performed. In this normalization technique, the imbalance in the dataset is handled to ensure that the model is not biased towards the majority class. This can be done using techniques such as oversampling or undersampling.

3.3 Feature Selection

Initially, the data's features are extracted and used as input for the gradient boosted decision trees (GBDT). However, GBDT is prone to overfitting, especially for small data sets, so it's crucial to reduce the number of features to only those that aid the classifier. To achieve this, feature selection is an important step in the decision tree pipeline as it helps reduce overfitting, eliminate redundant features, and prevent confusion for the classifier. Therefore, the proposed method incorporates feature selection to choose the most relevant features for the task by recursively considering smaller and smaller sets of features.

The process involves training the estimator on the initial set of features to obtain the importance of each feature. The least important features are then removed from the current set, and the classification metric is reevaluated. This procedure is repeated recursively until the desired number of features is selected. While this tool provides an initial approximation of the useful feature set, automated feature elimination is not always optimal and may require further fine-tuning. Thus, after the initial feature set is selected, permutation importance is utilized to choose the most appropriate features.

3.4 Prediction and Classification

AdaBoost (Adaptive Boosting) is an algorithm used for classification problems. It works by combining multiple weak learners to create a strong learner. The algorithm follows the following steps:

a. Initialize the sample weights: Initially, all samples are given equal weight.
b. Train a weak learner: A weak learner is trained on the data set. The weak learner is a simple model, such as a decision tree, that performs slightly better than random guessing.
c. Calculate the error: The error is calculated by comparing the weak learner's predictions with the actual labels.
d. Update the sample weights: The weights of the misclassified samples are increased, while the weights of the correctly classified samples are decreased. This ensures that the next weak learner focuses more on the misclassified samples.
e. Iterative steps are 2-4 for a fixed number of iterations or until the error rate reaches a certain threshold.
f. Combine the weak learners: The weak learners are combined by taking a weighted average of their predictions. The weights are determined by the accuracy of each weak learner.
g. Make predictions: The final model is used to make predictions on new data.

Hence, Improvised AdaBoost works by iteratively training weak learners on a data set, updating the sample weights to focus on misclassified samples, and combining the weak learners to create a strong learner that can accurately classify new data as explored in algorithm steps.

Given a training set of N samples {(x1, y1), (x2, y2), ..., (xN, yN)}, where xi is the i-th sample and yi is the corresponding label, with yi∈ {-1, 1}, the AdaBoost algorithm can be described as follows:

1. Initialize the sample weights: Set wi = 1/N for i = 1, 2, ..., N.
2. For m = 1 to M, where M is the number of weak learners:
    1. Train a weak learner on the training set using the current weights: Hm = train_weak_learner(X, y, w)
    2. Calculate the error rate of the weak learner: εm = sum(wi * (Hm(xi) != yi)) / sum(wi)
    3. Calculate the weight of the weak learner: αm = 0.5 * ln((1 - εm) / εm)
    4. Update the sample weights: wi = wi * exp(-αm * yi * Hm(xi)) for i = 1, 2, ..., N
    5. Normalize the sample weights: wi = wi / sum(wi) for i = 1, 2, ..., N
3. Combine the weak learners: The final prediction is given by the weighted sum of the weak learners: H(x) = sign(sum(αm * Hm(x))), where sign(z) = -1 if z < 0, 1 otherwise.

The above algorithm uses a weak learner Hm, which can be any binary classifier, such as a decision tree or a logistic regression model. The weight αm of the weak learner is determined by the error rate εm of the weak learner, with smaller errors leading to larger weights. The sample weights are updated at each iteration to focus on the misclassified samples, while the normalization ensures that the sample weights always sum to 1. Finally, the weak learners are combined by taking a weighted sum of their predictions, with larger weights indicating more accurate weak learners.

4. Results and Discussions

4.1 Dataset

The Oral Cancer Dataset (OCD): This dataset includes clinical and histological data from 804 patients with oral squamous cell carcinoma. It contains features such as age, gender, tumor size, tumor location, and histological grade. This dataset can be used for research and training purposes to develop accurate and reliable predictive models for oral cancer detection using the AdaBoost classification technique.

4.2 Paramatric Result

The parametric result for prediction and classification of oral cancer using AdaBoost would be the set of weights assigned to each weak classifier during the training process, as well as the decision threshold used to convert the output of the classifier into a binary prediction. the performance of the AdaBoost classifier on the testing set using metrics such as accuracy, precision, recall, and Error rate on proposed technique against prevailing techniques are explored in table 1.

Table1: Exploration of Parameters

Techniques	Accuracy	Precision	Recall	Error rate
SVM	89.2	90.2	86	5.7
NavieBaayes	89.7	91.8	84	10.2
KNN	91	93.4	89	7
CNN	93.5	94.6	90.4	4
AdaBoost	97	95.6	96	5.2
Improvised AdaBoost	98.6	97.8	98	1.8

SVM - Support Vector Machine
KNN – K-Nearest Neighbors
CNN - Convolutional Neural Network

5. Conclusion

The proposed research on Improvised AdaBoost technique has been shown to perform well in many classification tasks and has some advantages that make it a promising candidate for oral cancer prediction and classification. Improvised AdaBoost can handle datasets with noisy or incomplete data, which is common in medical datasets. Improvised AdaBoost also has a low risk of overfitting, meaning that it can generalize well to new data. it is a promising technique with a strong track record of success and should be considered as a potential approach in this area.

Acknowledgment

The research is funded by Institutional Research Seed Grant, PSG College of Arts & Science, Coimbatore.

Conflicts of interest

There are no conflicts of interest.

References

1. Ren, Z.; Hu, C.; He, H.; Li, Y.; Lyu, J. Global and regional burdens of oral cancer from 1990 to 2017: Results from the global burden of disease study. Cancer Communication, 40(2-3): 81–92, 2020.

2. Kowalski, L.P.; de Oliveira, M.M.; Lopez, R.V.M.; e Silva, D.R.M.; Ikeda, M.K.; Curado, M.P. Survival trends of patients with oral and oropharyngeal cancer treated at a cancer center in São Paulo, Brazil. Clinics, 75, e1507, 2020.
3. FatihahMohd, Noor Maizura Mohamad Noor, Zainab Abu Bakar, Zainul Ahmad Rajion, Analysis of Oral Cancer Prediction using Features Selection with Machine Learning, ICIT, The 7th International Conference on Information Technology, 2015.
4. Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, Health & Medical Informatics, 2157-7420, 2013.
5. HarikumarRajaguru and Sunil Kumar Prabhakar, Performance Comparison of Oral Cancer Classification with Gaussian MixtureMeasures and Multi Layer Perceptron, The 16th International Conference on Biomedical Engineering, 123-129, 2017.
6. Amy F. Ziober,Kirtesh R. Patel,Faizan Alawi, Phyllis Gimotty,4 Randall S. Weber, MichaelM. Feldman,Ara A. Chalian,Gregory S. Weinstein,Jennifer Hunt, and Barry L. Ziober, Identification of a Gene Signature for Rapid Screening of Oral Squamous Cell Carcinoma, American Association for Cancer, 2018.
7. Marc Aubreville, Christian Knipfer, Nicolai Oetter, Christian Jaremenko, Erik Rodner, Joachim Denzler, Christopher Bohr, Helmut Neumann, Florian Stelzle, & Andreas Maier, Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning, SCIENTIFIC Reports. 2017.
8. Deepak Kumar J, Surendra Bilouhan D, Kumar RC (2018) An approach for hyperspectral image classification by optimizing SVM using self-organizing map. Journal ofComputationl Science, 25(1):252– 259, 2017.
9. Yuan Y, Lin J, Wang Q (2015) Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization. IEEE Transactions on Cybernetics, 46(12):2966–2977, 2016.
10. Heba M, El-Dahshan ESA, El-Horbaty ESM, Abdel-Badeeh M (2018) Classification using deep learning neural networks for brain tumors. Future Computing Informatics Journal, 3(1):68–71, 2018.
11. Qi Dou, Hao Chen,Lequan Yu, Lei Zhao, Jing Qin,Defeng Wang, Vincent CT Mok, Lin Shi, Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks.IEEE Transactions on Medical Imaging, 35(5):1182–1195, 2016.
12. Chao Wang, Lei Gong, Qi Yu, Xi Li, Yuan Xie, Xuehai Zhou,A scalable deep learning accelerator unit on FPGA. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 36(3):513–517, 2017.
13. Kalantari N, Ramamoorthi R, Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics, 36(4):1–12, 2017.
14. Buyue Zhang, and Jan P. Allebach, 2008, 'Adaptive Bilateral Filter for Sharpness Enhancement and Noise Removal', IEEE Transactions on Image processing, vol. 17 (5): 664-668, 2008.