MIXED MODEL OF EXTREME LEARN MACHINE TREE AND RANDOM FOREST CLASSIFIER FOR PREDICTION OF ORAL CANCER

Dr. M Natarajan

Assistant Professor, Department of Computer Science,
Thanthai Hans Roever College, Perambalur, TamilNadu, India.
E-mail ID: prof.mnrajan@gmail.com

Dr. A. Muruganandam

Principal, Department of Computer Science,
Sri Aravindar Arts and Science College, Vanur, Villupuram, TamilNadu, India.
E-mail ID: murugandbc1976@gmail.com

Abstract- Oral Cancer is one of the deadliest diseases and most of the human are infected by this crucial disease in several parts of the world. It may occur in any part of the oral cavity. The early detection and prevention of oral cancel is very critical issue but it can improve the survival chances considerably, allow for simple treatment and provided the better quality of life for survivors. In existing system, the genetic algorithm is used for feature selection and the Support Vector Machine classifier algorithm is used for classification to predict the oral cancer. The feature selection and the classification is performed separately so the time complexity of the accuracy and prediction time quite complex So to solve this issue in proposed system the firefly algorithm is used for the feature selection and for the classification, mixed model of Extreme Learn Machine (ELM) Random forest classifier technique is used to improve the classification accuracy. The proposed system is tested with normal clinical data set which is improved the classification accuracy and the prediction time compared to existing system.

Keywords: Extreme Learning Machine (ELM), Random Forest (RF) Classification, Support Vector Machine (SVM) Classification, Firefly Algorithm (FA), Genetic Algorithm (GA).

1. INTRODUCTION

The Oral Cancer is also referred as mouth cancer. The mouthy cancers are initially started as lump, bump or patch in the mouth. Sometimes that does not go away after the few weeks are automatically happened either by you, your dentist or another doctor [1]. The most mouth cancers are squamous cell carcinomas (cancer cells come from the cells lining all parts of the inside of the mouth), but salivary gland cancers and other types of cancers can arise in the mouth as well.

1.1 Pre-cancerous Oral Lesions

There are also a few common pre-malignant lesions of which you should be aware.

- Leukoplakia
- Erythroplakia

- Dysplasia
- Lichen planus

1.2 Navigating Oral Cancers

In order to known about the several types of oral cancer, bellow mentioned an overview of the basics of oral cancer.

- Buccal Cancer
- Lip Cancer
- Oral Salivary Gland Cancer
- Tonguage Cancer

In this paper, we will predict the oracle cancer by using data Mining Techniques for improving the early detection of disease. The proposed Mixed Model of Extreme Learn Machine Random Forest Classifier (MMELMRFC) to detect the oral cancer. The proposed approach is increased the classification accuracy of detecting the oral cancer

2. RELATED WORK

This section describes the previous work of various researchers in oral cancer using different Data Mining Techniques. In [2] presented new approach for detecting cancer and prevention by association rule mining. It is used to extract the association among several valuable data pertaining to clinical symptoms and history of the cancer patients. In [3] Presented to analyze the salivary metabolites and identify the metabolic profiles specific to oral, breast and pancreatic. In this analysis is taken larger number of patient samples, particularly the data from the different institutes and additional clinical variables are required for further clinical application of this recent approach. In [4] presented to assess the comprehensive awareness of final year dental undergraduates of medical universities and institutes Ukraine concerning oral cancer and precancerous lesions. In [5] presented the effect on survival of elective node dissection to improve the early detection of oral cancer. In the prospective, randomized, controlled trial is evaluated the survival of elective node dissection between therapeutic node dissection in patients with lateralized T1 and T2 oral squamous cell carcinomas. The primary and secondary analysis is used to improve overall survival and disease-free survival respectively.

3. METHODLOGY

3.1 Data- mining techniques in oral cancer prediction

Oral cancer prediction is certainly very complex and non deterministic endeavor. Estimating the probability of cancer occurrences in patients requires that many factors (both genetic and non-genetic) are evaluated and properly weighted according to their significance and/or other (contact sensitive) contribution factors [7]. Some of the approaches in this search include:

Support Vector Machine (SVM)

The SVM is the supervised machine learning algorithm which is used for both classification and regression challenges. It mostly used to solve classification problems. In this algorithm, plot each data item as a point in n-dimensional space value being the value of a particular coordinate. The SVM is used to simply co-ordinates the individual observation.

Extreme Learning Machine (ELM)

The ELM is increased the accuracy of classification, regression, clustering, sparse approximation, compression and feature learning with a single layer or multi layers of hidden nodes, where the parameters of hidden nodes (not just the weights connecting inputs to hidden nodes) need not be tuned [6]. These hidden nodes are randomly allocated and never updated or inherited from ancestors without being changed. In most cases, the weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model.

Random Forest Classifier

Random forest is an ensemble classifier which consists of many decision trees and gives class as outputs i.e., the class's output by individual trees. Random forest is given many numbers of classification trees without pruning. Each classification tree is offered a specified number of votes for each class. Among all the trees, the algorithm chooses the classification with the greatest number of votes. Random forest runs efficiently on large datasets but is comparatively slower than other algorithms. It can effectively estimate missing values and hence is suitable for handling datasets with large number of missing values.

3.2 Data classification techniques in prediction of oral cancer

The data mining classification techniques contained various methods. The different method is utilized for different purpose, each method has its own advantages and disadvantages. In the data mining classification is one of the most important tasks. It is used to maps the data in to predefine targets. It is a supervised learning as targets for predefined. The aim of the classification is to build the classifier based on some cases with attributes to present the objects or one attribute to describe the group of the objects. The classifier is used to predict the group attributes of new cases from the domain-based values of other attributes. The most used classification algorithms are exploited in the microarray analysis is to belong four categories: IFTHEN Rule, Decision tree, Bayesian classifiers and neural networks.

IF-THEN Rule:

Rule induction: is the process of extracting useful 'if then' rules from data based on statistical significance. A Rule based system constructs a set of if-them-rules. Knowledge represents has the form.

IF conditions THEN conclusion:

This type of rule is contained two phases. The rule antecedent (the IF part) is contained one or more conditions about value of predictor attributes where as the rule consequent (THEN part) is contained a prediction about the value of a goal attribute. An accurate prediction of the value of goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. Rule Induction Method has the potential to use retrieved cases for predictions.

Decision Tree

Decision tree derives from the simple divide-and conquer algorithm. In these tree structures, leaves represent classes and branches represent conjunctions of features that lead to those classes. At each node of the tree, the attribute that most effectively splits samples into different classes is chosen. To predict the class label of an input, a path to a leaf from the root is found depending on the value of the predicate at each node that is visited. The most common algorithms of the decision trees are ID3 and C4.5. An evolution of decision tree exploited for microarray data analysis is the random forest, which uses an ensemble of classification trees. Showed that the good performance of random forest for noisy and multi-class microarray data.

Bayesian classifiers and Native Bayesian

From a Bayesian viewpoint, a classification problem can be written as the problem of finding the class with maximum probability given a set of observed attribute values. Such probability is seen as the posterior probability of the class given the data, and is usually computed using the Bayes theorem, estimating this probability distribution from a training dataset is a difficult problem, because it may require a very large dataset to significantly explore all the possible combinations. Conversely, Native Bayesian is a simple probabilistic classifier based on Bayesian theorem with the (native) independence assumption. Based on that rule, using the joint probabilities of sample observation. Despite its simplicity, the Native Bayes classifier is known to be a robust method, which shows on average good performance in terms of classification accuracy, also when the independence assumption does not hold.

Artificial Neural Networks (ANN)

An artificial neural network is a mathematical model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. Neurons are organized into layers. The input layer consists simply of the original data, while the output layer nodes represent the classes. Then, there may be several hidden layers. A key feature of neural networks is an iterative learning

process in which data samples are presented to the network one at a time, and the weights are adjusted in order to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained. In a review of advantages and disadvantages of neural networks in the context of microarray analysis is presented.

3.3 Proposed Architecture

In this paper, the firefly algorithm is used for feature selection for predicting the oral cancer. The firefly algorithm is the swarm intelligence-based mete heuristic technique which is inspired by the flashing behavior of fireflies. Each firefly is represented the set of attributes of the oral cancer. Initial population of fireflies is generating the operation of the prediction. The problem of the complexity accuracy and the execution time overcome by the firefly algorithm.

Algorithm:

Step 1: Initialize the populations of fireflies (Threshold values) are initialized.
Step 2: The intensity of the fireflies is calculated.
Step 3: The attractiveness function of the firefly is determined.
Step 4: The estimation of the distance (Update) between the two fireflies is measured.
Step 5: The movement of firefly is constructed.

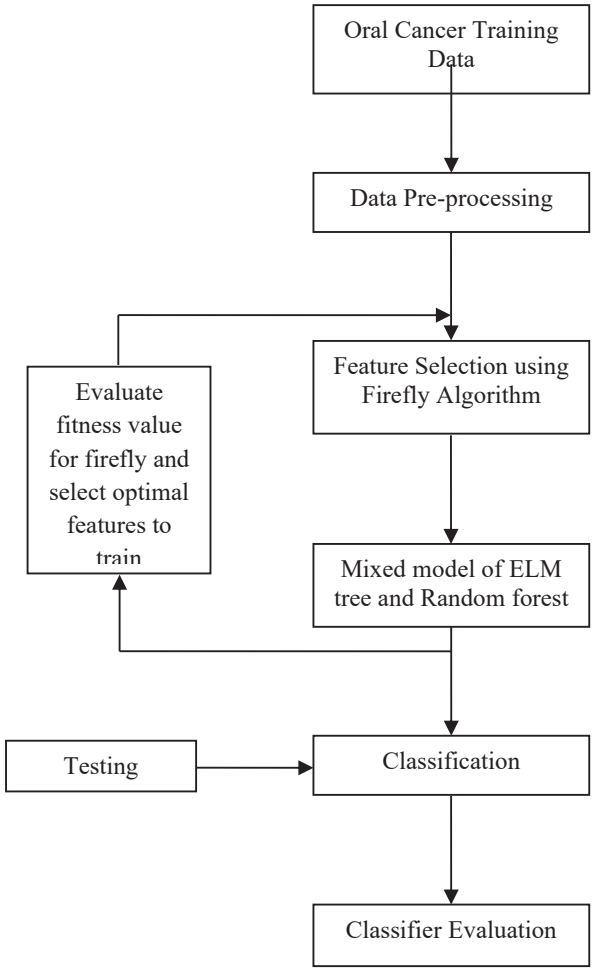

Fig 1: Architecture for prediction of oral cancer

IV. RESULT AND DISCUSSION

The performance of proposed approach is a Mixed Model of Extreme Learn Machine (ELM) tree and Random forest classifier (MMELMRFC) for prediction of oral approach is evaluated in terms of Accuracy, Precision, Recall, F-Measure and Specificity. The experimental result shows that the preposed MMELMRFC approach is achieved better result than Existing Genetic Algorithm Feature

Selection based Support Vector Machine Classifier approach (GAFSSVMC).

To evaluate the more effectiveness of the proposed method, the evaluation metrics such as Accuracy, Precision, Recall, F-Measure and Specificity are used, which is calculated using the following formulas.

Accuracy:

The accuracy is defined as the proportion of true results among the total number of cases examined. Accuracy can be calculated using this formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

Precision value is evaluated according to the feature classification at true positive prediction and false positive. It is calculated as follows:

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

Recall:

Recall value is evaluated according to the feature classification at true positive prediction and false negative.It is computed as follows:

$$\text{Recall} = \frac{\text{True positive}}{(\text{True positive} + \text{False positive})}$$

F-Measures:

F-measure is calculated from the precision and Recall. It is calculated as follows:

$$\text{F-Measure} = 2 \times \left(\frac{\text{Precision x recall}}{\text{Precision + recall}} \right)$$

Specificity:

Specificity is refer to the test ability to correctly detect patient without a condition. Specificity of a test is the proportion of healthy patients known not to have the diisease, who will test negative for it. Mathematically, this can be written as:

$$\text{Specificity} = \frac{\text{No. of True Negatives}}{\text{No. of True Negatives + No. of False Positives}}$$

Table 1: Comparisons of Performance Result

Metrics	Genetic Algorithm Feature Selection based Support Vector Machine Classifier approach (GAFSSVMC)	Mixed model of Extreme Learn Machine (ELM) tree and Random Forest Classifier (HELMRFC)
Accuracy	85	95
Precision	82	94
Recall	80	90
F-Measure	82	93
Specificity	83	95

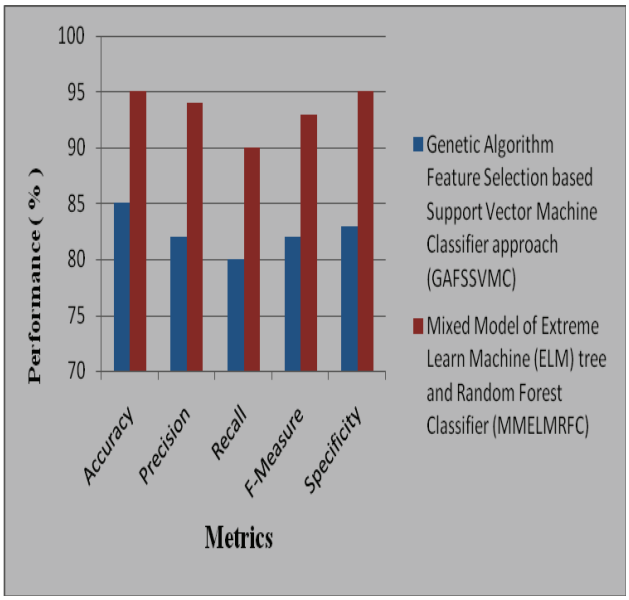

Fig 2: Comparisons of Performance Result

V. CONCULUSION

The main goal of this research is to find out the cancer based on firefly algorithm which is important task because of the disease complex in nature. The performance of any algorithm relies upon the parameters used for the process. So, each process level the accuracy can be strengthened. Here performance is improved 11 percentage compare to existing. The prediction of the oral cancer using enhanced different techniques changes from time to time because of technological advancement. Initially, the classification method is enhanced by focusing on the heterogeneous data, feature selection based on Mixed Model of Extreme Learning Machine and Random Forest Classifier. In these methods the firefly optimization algorithm is identified the more reliable features from oral cancer data and images. Ultimately these features are used in Mixed Model of Extreme Learning Machine and Random Forest Classifier to classify the oral cancer data, in order to predict the oral cancer from oral images.

VI. REFERENCES

1. Sharma, N., & Om, H. (2014). Extracting Significant patterns for oral cancer detection using apriori algorithm. Intelligent Information Management, 6(02), 30-33.

2. Sharma, N., & Om, H. (2013). Data mining models for predicting oral cancer survivability. Network Modeling Analysis in Health Informatics and Bioinformatics, 2(4), 285-295.

3. Sugimoto, m., Wong, D. T., Hirayama, A., Soga, T., & Tomita, M. (2010). Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, oral and pancreatic cancer-specific profiles. Metabolomics, 6(1), 78-95.

4. Kolenko, Y. (2016). Awareness of oral cancer and precancer among final year dental undergraduates n Ukraine. Journal of Education, Health and Sport, 6(3), 106-112.

5. D'Cruz, A. K., Vaish, R., Kapre, N., Dandekar, M., Gupta, S., Hawalder, R., ... & Kane, S. (2015). Elective versus therapeutic neck dissection in node-negative oral cancer. New England Journal of Medicine, 373(6), 521-529.

6. Chhaya Gupta, Nasib Singh Gill, (2020). Machine Learning Techniques and Extreme Learning Machine for Early Breast Cancer Prediction. International Journal of Innovative Technology and Exploring Engineering, 9(4), 163-167.

7. Ramachandran, P., Girija, N., & Bhuvaneshwari, T., (2014). Early detection and prevention of cancer using data mining techniques. International Journal of Computer Applications, 97(13), 78-85.

8. Tang, J. Y., Chuang, L. Y., Hsi, E., Lin, Y. D., Yang, C. H., & Chang, H. W. (2013). Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the apriori algorithm. BioMed research international, 2013.

9. Krishnan, M. M. R., Acharya, U. R., Chakraborty, C., & Ray, A. K. (2011). Automated diagnosis of oral cancer using higher order spectra features and local binary pattern: A comparative study. Technology in cancer research & treatment, 10(5), 443-455.

10. Exarchos, K. P., Goletsis, Y., & Fotiadis, D. I. (2012). Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Transactions on Information Technology in Biomedicine, 16(6), 1127-1134.