International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Aggregate Linear Discriminate Analyzed Feature Extraction and Ensemble of Bootstrap with Knn Classifier for Malicious Tumour Detection

S.SubashChandraBose, T.Christopher

Abstract: ***Tumour detection medical applications utilize classification techniques to categorize malicious and non-malicious tumour features to provide an efficient medical diagnosis of the human individual under investigation. One way to enable efficient classification, Feature extraction methods are used to eliminate the redundant features and obtain the most relevant features. However, the challenges concerning the dimension and quantum of tumour dataset persist. Toward this goal, this paper aims to maximize the malicious tumour classification accuracy using two reliable ensemble classifiers namely Bootstrap Aggregation and k-nearest neighbour. Tumour features extracted by Aggregate Linear Discriminate Analysis (LDA) and the feature distance is calculated with iterative scattering matrix algorithm. The extracted features are further refined by aggregation to select most effective feature values. After this, an ensemble classifier technique is employed to construct malicious and non-malicious tumour classes. The tumour classification based on an ensemble of bagging and k-nearest neighbour. Simulation is carried out on Tumour Repository data set to show that proposed ensemble classifiers have considerably better tumour detection accuracy than existing conventional techniques. Numerical performance evaluations show that 8% improvement by proposed method in tumour classification accuracy for malicious tumour detection in human individuals.***

Keywords: ***Feature extraction, ensemble-based classifiers, Bootstrap Aggregation, Aggregate Linear Discriminate Analysis, k-nearest neighbour, Gene expression.***

I. INTRODUCTION

Early detection of tumour detection is necessary for successful treatment for tumour classification. Three ensemble classification structures that fuse information from multiple sensors to identify abnormalities in the breast was investigated in [1] using cost-sensitive support vector machines.They were feature fusion approach, classifier fusion approach and ensemble selection approach respectively.

By applying a cost-sensitive ensemble classification model, the algorithms were able to select threshold values in such a way to reduce the false positive rate below a specified maximum value. However, the performance of classification minimized when the algorithms apply to the clinical dataset.

This motivated the improvement of measurement system and procedure, and the further development of classification algorithms to provide measures for dimensionality related issues. With this objective, an Aggregated Linear Discriminate Analysis-based Feature Extraction (ALDA-FE) investigated in this work. The ALDA-FE offers an improvement of the measurement system and procedure that considers the sum of within-class and between-class scattering matrices, ensuring dimensionality reduction.

Classification of cancer by molecular level has gained the interest of researchers because it provides the diagnosis of disease in a more systematic, accurate and objective manner for several types of cancer. An ensemble system, a set of individually trained classifiers presented in [2] whose decisions integrates with majority voting, weighted voting and Naïve Bayes combination.

The ensemble method in [2] also provided solutions enhancing the accuracy of the result, applied ensemble technique to more cancer types, and avoiding the problems related to over fitting. However, to improve tumour classification accuracy, different classifiers have to be used as base members. This is addressed by investigating ensemble classifiers, namely, Bootstrap Aggregation and K Nearest Neighbour in the proposed work and thus improving the tumour classification accuracy.

Our work focuses on the development of two strong ensemble classifiers, Bootstrap Aggregation and K Nearest Neighbour screening system. The ensemble classifiers and the associated algorithms can process measurements and decide as to whether a malicious tumour is present or not.

II. RELATED WORKS

Most of the previous work on tumour classification has concentrated on cost-sensitive classification methods. An efficient Cost-Sensitive Classifier with Gentle Boost Ensemble (Can-CSC-GBE) presented in [3] for the classification of breast cancer based on the protein amino acid features. The application of ensemble methods resulted in the minimization of misclassification cost. However, accuracy was not said to identify the concentrated work.

Manuscript published on 30 September 2019
* Correspondence Author
S.SubashChandraBose*, Research Scholar, PG and Research Department of Computer Science Government Arts College, Udumalpet, bose.milestone@gmail.com
Dr.T.Christopher, Assistant Professor, PG and Research Department of Computer Science, Government Arts College, Coimbatore, chris.hodcs@gmail

An ensemble model, Hierarchical Multi-level classifiers bagging with Multi-objective optimized voting (HM-BagMOOV) in [4] achieved the highest accuracy. Though, accuracy rate ensures the prediction of recurrence.

Hybrid Computer-aided diagnosis system for Prediction of Breast Cancer Recurrence (HPBCR) designed in [5] that provided a promising means for prediction of breast cancer recurrence. Machine learning methods were applied in [6] to reduce the human workload of matching records, minimizing the computation cost.

Recently, specific research has explored the application of proximity relations, in particular, classifiers, for feature selection. Many weak classifiers were combined into a strong optimized classifier [7], resulting in the best features identifies the selected result based on the performance. Though features being selected were found to be the best, the diagnosis rate was not concentrated.

Weighted Vote-based Ensemble model was investigated in [8] using different classifiers resulting in effective breast cancer diagnosis. Another hybrid intelligent system was developed in [9] by applying feature extraction and prediction module. This hybrid intelligent system resulted in the effective diagnosis of cancer at an early stage. RNA-sequencing data was used in [10] to differentiate Histologic Grade1 (HG1) and Histologic Grade3 (HG3) with high accuracy.

There has also been some recent work towards the disease diagnosis at an early stage for malicious tumour detection. In [11], a web-based clinical expert system designed that overcome the limitations of the individual as well as other ensemble classifiers. To enhance the prediction performance, SVM and SVM ensembles were used in [12] resulting in the early prediction of breast cancer.

Another multimodality model using wavelets and machine learning investigates the feature [13]. A multistage classification model proposed [14] for breast cancer diagnosis applying a histogram equalization and nonlocal means filtering. This multistage classification model resulted in the improvement of classification accuracy for breast cancer diagnosis.

In the area of machine learning, the model's training heavily depends on the performance of data. But, the distribution of two classes behaves in an imbalanced manner, and the scenario is found to be ubiquitous in real life. In [15], a novel ensemble method called, Bagging of Extrapolation Borderline-Synthetic Minority Oversampling Technique SMOTE SVM [16] (BEBS) designed for dealing with imbalanced data.

A selective ensemble method is combining, KNN, SVM and Naïve Bayes present in [17] for breast cancer diagnosis. Performance analysis of data mining algorithms for breast cancer diagnosis provided [18]. A comparative study of breast cancer classification models using ensemble classifiers presented for tumour classification [19]. Another Neural Network ensemble method was designed in [20] to enhance the rate of classification accuracy.

This paper presents a novel application of ensemble classification methods using tumour repository dataset collected from UCI to detect the presence of a tumour. Our main contributions when compared to state-of-the-art work are the following:

a) We employ an Aggregate Linear Discriminate Analysis technique to extract the best features to improve the sensitivity (true positive rate) while minimizing the specificity (true negative rate);
b) We design two ensemble classification architectures to construct malicious and non-malicious tumour classes and thus reducing the tumour classification time.
c) We demonstrate the performance of our ensemble classification methods using data collected in a clinical trial with freshly excised breast tissues made at different frequencies.

This paper extends the previous work by designing an ensemble selection-based classification method which significantly outperforms existing methods. It also provides a more detailed description of our algorithms, Iterative Scattering Matrix algorithm and Ensemble-based Classification algorithms that factor in the tumour detection accuracy, and a complete performance evaluation involving UCI repository datasets, Breast tissue and Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

The remainder of the paper organized as follows. Section 2 reviews the related works. Chapter 3 introduces our system, algorithm, and the ensemble classifier. Experimental results reported in Section 4 with the detailed discussions provided in Section 5. A summary of concluding remarks included in Section 6.

III. PROPOSED SYSTEM

In this paper, we propose a selective ensemble classification method called, Aggregate Linear Discriminate Analysis-based Ensemble Classifier (ALDA-EC) for malicious tumour classification and detection. The proposed ALDA-EC method is designed with the objective of maximizing the classification performance of malicious tumour detection with minimal time as compared to state-of-the-art works. This aim of proposed ALDA-EC method is achieved by the application of Ensemble-based Classification (i.e. integration of Bootstrap and k-nearest neighbour classifiers). As Ensemble-based Classification applied Weighted Voting Scheme to classify difficult instances, where each instance is classified by measuring the total vote based on the K Nearest Neighbours. Then, the highest vote of the base classifier is assigned to each sample with a lower amount of time utilization. Thus, ALDA-EC method increases the classification accuracy for effective malicious tumour detection as compared to existing works.

On the contrary to existing classification works, Aggregate Linear Discriminate Analysis technique is used in a proposed method to extract the best features for malicious tumour classification.

In addition to existing classification works, a proposed method designed an Ensemble-based Classification technique by integrating a Bootstrap aggregation with k-nearest neighbour classifiers using the weighted voting rule to reduce the classification errors of malicious tumour detection as compared to state-of-the-art works.

Extracting the feature using UCI machine learning repository and cancer gene expression profiles namely Breast tissues (BT), WDBC (Wisconsin Diagnostic Breast Cancer), respectively, we sampled with 10-fold cross-validation to train the base classifiers. Cross-validation is used in proposed ALDA-EC method to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. With the objective of improving the tumour classification accuracy or the malicious tumour detection rate, the proposed work integrates Bootstrap and k-nearest neighbour classifiers and ranks them based on the trained tumour indicator value. The better-ranked class is then selected by majority vote to detect malicious tumour classes. The flow chart of the proposed work ALDA-EC is shown in figure 1.

As shown in the figure, we consider Aggregate LDA-based Ensemble Classifiers in the tumour classification systems based on cancer data sets to achieve a high degree of accuracy. The proposed system consists of two main modules: the feature extraction and the ensemble classification.

To start with the features in the training dataset are extracted using Aggregated LDA. With the extracted features, each base classifier is trained based on the assigned weight for each sample.

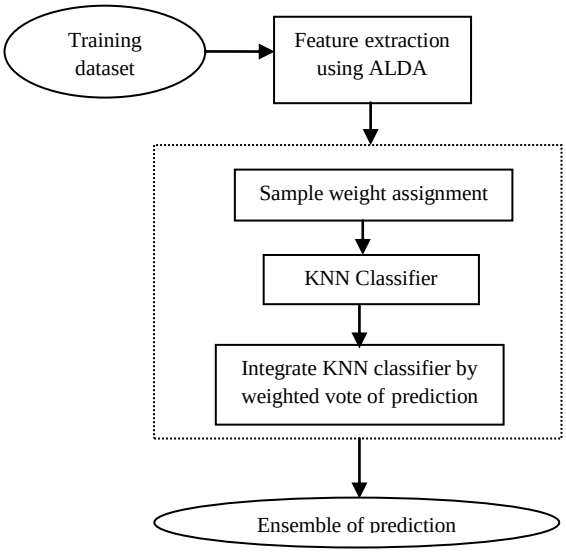

Figure 1 Block diagram of Aggregate LDA-based Ensemble Classifier

These base classifiers are combined using the weighted voting rule resulting in an ensemble-based classifier. Two well known Bootstrap Aggregation (based on a weighted vote of prediction) and KNN base classifiers are used to arrive at to classify ensemble classifier into malicious or non-malicious tumour classes, achieving ensemble of prediction.

A. Aggregate LDA-based Feature Extraction

The extraction of data follows a standard procedure depicted in figure 2, where the extraction process summarized in three main steps. In ALDA, feature extraction is performed by constructing a matrix within the class and between the classes. Feature distance is evaluated using the Iterative Scattering Matrix algorithm for reducing dimensionality in feature extraction.

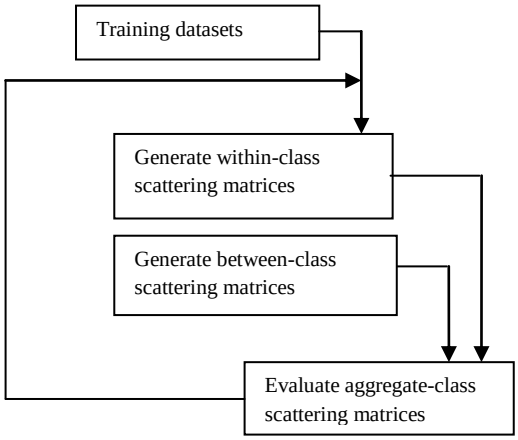

Figure 2 Main steps in the feature extraction process

The Iterative Scattering Matrix algorithm infers to the evaluation of within-class and between-class scattering for extracted data. Finally, the extracted features are further refined using aggregate-class scattering matrices. The aggregate scattering matrices reduce the distance between samples in similar classes and improving distance between different classes. In this manner, the most predominant features selected.

Let us consider a training dataset 'td', with a total of 'N' features, comprising of$m$' rows and 'n' columns. To extract relevant features, the ALDA-EC method obtains an initial set of measured data with which builds desired values using aggregate class. The aggregate class in the ALDA-EC method consists of the summation of within-class scattering matrices and between-class scattering matrices with which the features extracted. In addition to arriving at aggregate class, with the objective of minimizing the classification errors, the ALDA-EC method uses a local mean and global mean values for neighbourhood processing. The global mean value is obtained as given below,

$$\mu = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} \qquad (1)$$

From (1), the global mean value 'μ' for neighbourhood processing for measured data 'x_{ij}' is obtained. Next, the local mean value 'm_i' is obtained through '$4-neighbor\ pixel$' and is as given below.

$$(m_i) = [(m-1,n),(m+1,n),(m,n-1),(m,n+1)] \quad (2)$$

From (2), the '$4-neighbor\ pixel$' is symbolized as($m-1,\ n,\ m+1,\ n,\ m,\ n-1,\ m,\ n+1$' respectively. With the obtained local and global mean values, within-class matrices and between-class matrices formed efficiently, which forms the basis for feature extraction. The within-class matrix obtained is as given below.

$$S_w = \frac{1}{N}\left(x_{ij} - m_i\right)\left(x_{ij} - m_i\right)^{MIT} \quad (3)$$

From (3), the within-class matrices 'S_w' is the product of the differences between the measured data 'x_{ij}' and local mean 'm_i' to that of the transposed measured data and local mean '$\left(x_{ij} - m_i\right)^T$' respectively. The between-class matrix obtained is as given below.

$$S_b = \frac{N_i}{N}(m_i - \mu)(m_i - \mu)^T \quad (4)$$

From (4), the between-class matrices 'S_b' are arrived from the product of the differences between the local mean 'm_i' and the global mean 'μ' to the transposed local and global mean '$(m_i - \mu)^T$' respectively. Finally, to select the most predominant features, the aggregated class scattering matrices is obtained as given below.

$$S_a = S_w + S_b \quad (5)$$

$$FE = S_a \quad (6)$$

From (5), the aggregated-class forms the summation of the within-class and between-class matrices. From that, the distance between test samples in similar classes reduced. This helps to extract the features and reduces the dimensionality. The objective behind the design of aggregated-class is to reduce the within-class scattering while increasing the between-class scattering. As a result, the ALDA-EC method minimizes the misclassification probability by selecting features that reduce the distance between the samples in similar classes and improve the distance between different classes. The pseudo code for Iterative Scattering Matrix manipulationgiven in algorithm

Input: Total features '??', rows '??', columns '??', measured data '??.'

Output: Features extracted

1: **Initialize** measured data '??.'
2: **Begin**
3: **Repeat**
4: Obtain global mean '??' using (1)
5: Obtain local mean '?? ??' using (2)
6: Measure within the class '??' using (3)
7: Measure of class '??' using (4)
8: Measure aggregate class '??' using (5)
9: **Until** (all features processed)
10: **End**

Algorithm 1 Aggregate LDA

As given in the algorithm 1, the Iterative Scattering Matrix algorithm obtains the target point in the features based on class separation. These target point in the features then forms the basis for ensemble classifiers explained in the following sections. Creating an aggregation class using the within and between class not only extracts more relevant features for classification but also reduces the redundant features present in the class. In this way, the ALDA-EC method reduces the misclassification probability or improves the true positive rate of the features extracted and therefore proves the way for tumour prediction.

B. Ensemble classifiers using Bootstrap Aggregation and K Nearest Neighbour

Ensemble classifiers are learning algorithms that construct a set of classifiers. It further classifies new data points through a weighted vote of their predictions. The objective behind the use of ensemble methods is to use a combination of models to increase tumour classification accuracy. In recent years, several methods for constructing ensembles have developed.

In this work, we consider bagging methods that create individuals for its ensembles by training each classifier on a random redistribution of the training set, called as Bootstrap replicate. The technique therefore referred to as Bootstrap Aggregation. Followed by this, the distance between the test sample and the nearest neighbour predicted and measured via KNN. This in turn, therefore, minimizes the complexity involved using the Euclidean distance measure.

On the contrary to existing works, the weighted voting rule is used in proposed ALDA-EC method in order to minimize the classification errors of malicious tumour detection. With help of Weighted Voting Rule scheme, an ensemble-based classification is performed in ALDA-EC method according to the class that acquires the most frequent votes. Therefore, ensemble-based classifiers reduce the classification error rate of malicious tumour detection than that of the base classifiers. Thus, the proposed ALDA-EC method provides good classification performance for malicious tumour detection.

The Ensemble-based Classifier model used the majority vote of the samples for the decision of n^{th} class to effectively classify the samples into a malicious or non-malicious tumour classes. From that, the majority vote of positive samples of base classifiers is classified as non-malicious tumour classes. In the same way, the majority vote of negative samples of the base classifiers is classified as a malicious tumour class. Thus, the ensemble method enhances the classification performance by combing all base classifiers results into a strong classifier. Figure 3 shows the Ensemble-based Classifier model.

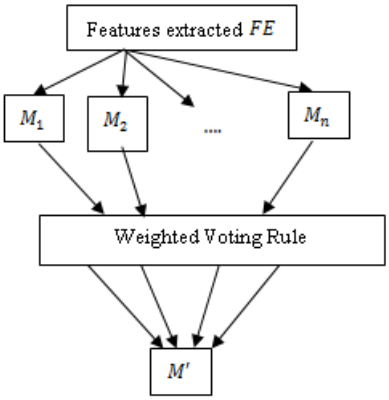

Figure 3 Block diagram of the Ensemble-based Classifier model

Let us consider a training set of size 'k', with 'k' possible random instances; then, the ensemble methods combine a series of 'k' learned models or random instances '$M_1, M_2 \ldots M_k$' with the aim of creating an improved modelM''. Given a training set 'T' (i.e. extracted features 'FE' using ALDA-EC) using 't' tuples, each iteration, 'i', a training set 'T_i' or 'FE_i' of 't' tuples is sampled with replacement from 'T'. A classifier model 'M_i' is learned from each training set 'T_i'.

To classify an unknown sample 's', each classifier 'M_i' returns its class prediction by Weighted Voting Rule scheme as given below. In this Weighted Voting Rule scheme, a classification performed according to the class that acquires the most frequent votes.

$$Class\ (P) = MAX\ (\sum g\ (q_k(s), c_i)) \quad (7)$$

From (7), '$q_k(s)$' represents the classification of the 'kth' classifier with 's' denoting the sample to be classified and '$g(q, c)$' is a scaling function and is mathematically written as given below.

$$g\ (q, c) = \begin{cases} 1, & q = c \\ 0, & q <> c \end{cases} \quad (8)$$

With the Weighted Voting Rule scheme, using KNN classifier, the classes are classified by a majority vote of its neighbour features with the features being assigned to the class most common among its K Nearest Neighbours. To minimize the computational complexity of base classifiers and selecting the best value of k, the ALDA-EC, select '$k = 1$'. To find the majority vote mathematical formula given as below.

$$Class_n^{1NN}\ (s) = q_{(1)} \quad (9$$

From (9), the ALDA-EC method uses the nearest neighbour type classifier as the one nearest neighbour classifier '$1NN$' that assigns a sample 's' to the class '$Class$' of its closest neighbour. This is because in the ensemble based classifiers having an error rate lesser than that of the base classifiers '$Base_c$' is enough to arrive at good performance. With this, an ensemble-based classifier is generated by combining the base classifiers using the Weighted Voting Rule scheme to classify given test samples'.

Then, the total vote for each class evaluated mathematical formulas as given below to obtain the excellent performance.

$$TV_c = \log \frac{1}{\frac{Base_c}{1 - Base_c}} \quad (10)$$

From (10), the total vote for each class 'TV_c' obtained by logarithmic values of the base classifier classes '$Base_c$'. With this, the sample 's' is assigned to class 'c' based on highest vote 'TV_c'. Given below is the pseudo code representation of the Ensemble-based Classification algorithm.

Input: Features Extracted 'FE', samples'
Output: Tumour classification
1: **Begin** 2: **For** each extracted features 'FE' and '$k = 1$.' 3: **For** each class 'c' and samples' 4: Evaluate Weighted Voting Scheme using (7) 5: Measure KNN distance classification using (9) 6: Measure total vote using (10) 7: Assign the highest vote 8: **End of** 9: **End of** 10: **End**

Algorithm 2 Ensemble-based Classification

As given in the pseudo code, the ensemble-based classification starts with the extracted features obtained using ALDA-EC. Followed by this for each class and samples, each iteration includes two steps. The first step involves the Weighted Voting Scheme that is applied to arrive at classifying difficult instances, where each instance is classified by measuring the total vote, based on the K Nearest Neighbours. Next, the highest vote assigned to each sample where the final classification represents the simple majority vote. In this way, malicious tumour classification accuracy is improved based on the strong ensemble classifiers, Bootstrap Aggregation and k-nearest neighbour.

IV. PERFORMANCE EVALUATION

In this section, we provide an experimental evaluation of the proposed algorithm. We implement our method using JAVA with various experiments conducted on Breast tissues (BT) and WDBC (Wisconsin Diagnostic Breast Cancer) data sets.

To prove the efficiency of the proposed algorithm, and made the comparison with the results of several existing ensemble classifier methods, including ensemble selection-based algorithm [1] and the ensemble system for cancer classification [2]. The datasets used and summarized in Table 1.

Table 1 UCI Machine Learning Datasets

Dataset	Source	k	n	a
Breast Tissues	UCI	6	106	10
WDBC	UCI	2	569	32

The two datasets Breast tissues and WDBC include 106 and 569 number of instances, each comprising of 10 and 32 attributes and 6, two classifications have made. The breast tissues dataset included the impedance measurements of freshly excised breast tissue were made at the frequencies: 15.625, 31.25, 62.5, 125, 250, 500, 1000 KHz. Regarding the WDBC dataset, features collected and computed from the digitized image of a fine needle aspirate (FNA) of a breast mass. The scanned image described the characteristics of cell nuclei present in the image. When conducting experimental work using Breast Tissue dataset, ALDA-EC method used the following attributes for classification of malicious tumour

Table 2 Attribute information's about Breast Tissue dataset

Attribute Name	Description
I0	Impedivity (ohm) at zero frequency
PA500	phase angle at 500 KHz
HFS	the high-frequency slope of the phase angle
DA	impedance distance between spectral ends
REA	area under spectrum
A/DA	area normalized by DA
MAX IP	maximum of the spectrum
DR	the distance between I0 and real part of the maximum frequency point
P	length of the spectral curve

When performing experimental work using Breast Tissue dataset, ALDA-EC method used below attributes for classification of malicious tumours.

1. Sample code number
2. Clump Thickness
3. Uniformity of Cell Size
4. Uniformity of Cell Shape
5. Marginal Adhesion
6. Single Epithelial Cell Size
7. Bare Nuclei
8. Bland Chromatin
9. Normal Nucleoli
10. Mitoses

The metrics used to measure the performance evaluation of ALDA-EC are tumour classification accuracy, tumour classification time, sensitivity and specificity. The tumour classification accuracy 'CA_s' of an individual sample 's' depends on the number of samples correctly classified (true positives plus true negatives) and is evaluated by the formula as given below.

$$CA_s = \frac{CC}{n} * 100 \quad (11)$$

From (11), 'CC' is the number of sample cases correctly classified and 'n' is the total number of sample cases. The tumour classification time 'CT_s' of an individual sample 's' depends on the number of samples correctly classified (true positives plus true negatives) and the time taken to classify the correct classified samples. The tumour classification time is evaluated by the formula as given below.

$$CT_s = Time\left(\frac{CC}{n} * 100\right) \quad (12)$$

From (12), 'CC' is the number of sample cases correctly classified and 'n' is the total number of sample cases.

Sensitivity also called the true positive rate measures the proportion of positives that correctly identified as such. In other words, sensitivity or true positive rate refers to the percentage of malicious classes which correctly identified as having the condition. Higher the rate of sensitivity, more efficient the method is said to be.

$$Sensitivity = \frac{No.of\ true\ positives}{Total\ no.of\ malicious\ tumour\ classes\ in\ a\ sample} \quad (13)$$

Specificity also called the true negative rate measures the proportion of negatives that correctly identified as such.

In other words, specificity or true negative rate refers to the percentage of non-malicious tumour classes which correctly identified as not having the condition

$$Specificity = \frac{No.of\ true\ negatives}{Total\ no.of\ non-malicious\ tumour\ classes\ in\ a\ sample} \quad (14)$$

V. DISCUSSION

Firstly, the proposed ALDA-EC method compared for malicious tumour detection with two other tumour detection methods, ensemble selection-based algorithm [1] and ensemble system for cancer classification [2]. Four performance measures were employed to evaluate the tested tumour detection method. The first measure was the tumour classification accuracy, which measures the relative monotonicity between the sample cases correctly classified and the total number of sample cases. The second measure, tumour classification time which measures the time taken to classify the tumour cells, if detected. Finally, the third and fourth measure, sensitivity and specificity are evaluated.

C. Performance evaluation using WDBC dataset

Table 2 records the measure of four metrics, namely, tumour classification accuracy, tumour classification time, sensitivity and specificity obtained from (11), (12), (13) and (14). It proves from Table 2 that our malicious tumour detection method, ALDA-EC provides better performance measure among all the other methods.

Table 2 Performance evaluation of ensemble selection-based algorithm and ensemble system for cancer classification using WDBC dataset

Metrics	ALDA-EC	Ensemble selection-based algorithm	Ensemble system for cancer classification
Tumour Classification Accuracy (%)	68.15	65.23	61.45
Tumour Classification Time (ms)	75.13	89.13	88.28
Sensitivity (%)	65.13	61.23	59.15
Specificity (%)	62.13	59.19	56.28

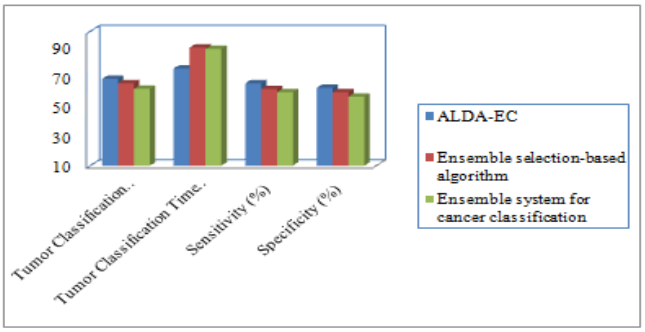

Figure.4.Performance comparisons with ensemble selection-based algorithm [1] and ensemble system for cancer classification [2] using WDBC dataset

Figure 4 shows that the tumour classification accuracy, tumour classification time, sensitivity and specificity have better performance improvement as the features extracted increases from 10 to 90. The reason is that an ensemble of the classifier used in the ALDA-EC method.

The ensemble is performed using bagging and the K Nearest Neighbour. With these two classifiers, ALDA-EC method performs tumour classification efficiently.

Furthermore, by applying the ensemble-based classification algorithm in the ALDA-EC method, Weighted Voting Rule scheme, using KNN classifier measures difficult instances. Here, each instance is classified by measuring the total vote, based on the K Nearest Neighbours, by assigning the highest vote to each sample.

This helps in achieving the performance improvement of tumour detection using ALDA-EC method when compared to the existing methods [1] and [2] respectively.

D.Performance evaluation using Breast Tissue dataset

Table 3 shows the performance of the methods tested across all views when Breast Tissue dataset have used. It is evident that the proposed ALDA-EC performs consistently better than the compared methods across all aspects. Note that all methods achieve the lowest classification time in the frontal view. However, the ALDA-EC significantly improves the performance attained by the other methods in this view. We attribute this to the fact that ALDA-EC performs the classification using bagging that builds several instances of a black-box estimator on random subsets of the original training set. It then aggregates individual predictions to form a final prediction, where the classification of the features from the frontal view facilitated.

Table 3 Performance evaluation of Breast Tissue dataset

Metrics	ALDA-EC	Ensemble selection-based algorithm	Ensemble system for cancer classification
Tumour Classification Accuracy (%)	71.17	68.25	64.48
Tumour Classification Time (ms)	68.08	72.09	81.23
Sensitivity (%)	70.23	66.33	64.25
Specificity (%)	67.33	54.33	51.48

Having the full features extracted from each training dataset, we can calculate how much correlation there is between an individual evaluation an the overall mean.

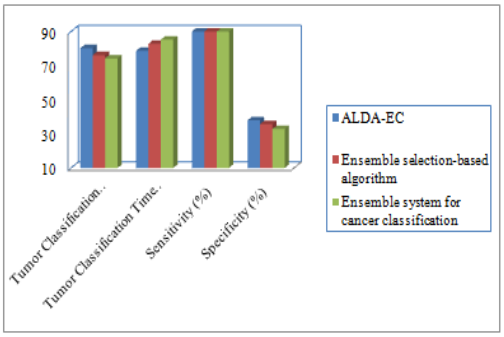

Figure 5 Performance comparisons with ensemble selection-based algorithm [1] and ensemble system for cancer classification [2] using Breast Tissue dataset

Figure 5 compares the proposed ALDA-EC with two other methods, ensemble selection-based algorithm [1] and the ensemble system for cancer classification [2] four best-performing algorithms even further. It is worth noting that the tumour classification accuracy is higher than the state-of-the-art methods [1] and [2]. Since aggregated predictions are used to form a final prediction, ALDA-EC reduces the variance of the base estimator by introducing randomization into its construction stage and then creating an ensemble out of it. As the proposed framework using ensemble methods, named, bootstrap aggregation and K Nearest Neighbour, it provides a way to reduce overfitting and therefore reducing the tumour classification time also using Breast Tissue dataset. The improvement was found to be 4% and 10% compared to [1] and [2], reducing the classification time by 6% and 16% respectively.

VI. CONCLUSION

In this paper, we proposed a method for selective ensemble classification method for malicious tumour detection. By observing that dimension and quantum of a tumour usually compromised, they are utilized to obtain relevant features, the ALDA-EC method, initial set of measured data are obtained, which then builds desired values using an aggregate class with the aid of Iterative Scattering Matrix algorithm. A significant feature of our method is that the dimensionality of features, especially the most relevant features obtained at the same time, the Iterative Scattering Matrix algorithm reduces the dimensionality. That is, our methods allow the doctors or the medical practitioners to reduce the tumour misclassification error with the aid of local and global mean for neighbourhood processing. This goal is achieved using the concept of aggregation of linear discriminate analysis. On the other hand, our method allows the correct classification of a tumour using the ensemble-based classifiers. This goal is achieved using the techniques of Bootstrap Aggregation and KNN. Through experimental study, we show that ensemble aggregation used in ALDA-EC can help to reduce the total tumour classification time by 15% and 11% compared to the ensemble selection-based algorithm using two different datasets, WDBC and Breast Tissue dataset respectively. The future enhancement of ALDA-EC method can proceed with different boosting techniques to reduce the false positive rate of malicious tumour classification.

REFERENCES

1. Yunpeng Li, Emily Porter, Adam Santorelli, Milica Popovic, Mark Coates, “Microwave breast cancer detection via cost-sensitive ensemble classifiers: Phantom and patient investigation”, Elsevier, Biomedical Signal Processing and Control, Volume 31, January 2017, Pages 366-376.
2. Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman,“Gene expression based cancer classification”, Elsevier, Egyptian Informatics Journal, Available online 20 December 2016, Pages 1-9.
3. Safdar Ali, AbdulMajid n, SyedGibranJaved, MohsinSattar, “Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentle boost Ensemble for breast cancer classification using protein amino acids and imbalanced data”, Elsevier, Computers in Biology and Medicine, Volume 73, 1 June 2016, Pages 38-46.
4. Saba Bashir, Usman Qamar, Farhan Hassan Khan, “IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework”, Elsevier, Journal of Biomedical Informatics, Volume 59, February 2016, Pages 185-200.
5. Mohammad R. Mohebian, Hamid R. Marateb, Marjan Mansourian, Miguel Angel Mañanas, Fariborz Markarian, “A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning”, Elsevier, Computational and Structural Biotechnology Journal, Volume 15, 2017, Pages 75–85.
6. Yannick Siegert, Xiaoyi Jiang, Senior Member, Volker Krieg, Sebastian Bartholomaus, “Classification-Based Record Linkage with Pseudonymized Data for Epidemiological Cancer Registries", IEEE Transactions on Multimedia, Volume 18, Issue 10, October 2016, Pages 1929 – 1941.
7. Taiping Zhang, Pengfei Ren, Yao Ge, Yali Zheng, Yuan Yan Tang, and C.L. Philip Chen, “Learning Proximity Relations for Feature Selection”, IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 5, May 2016, Pages 1231-1244.
8. Saba Bashir, Usman Qamar and Farhan Hassan Khan, “Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble”, Springer, Quality & Quantity, Volume 49, Issue 5, September 2015, Pages 2061–2076.
9. A. H. El-Baz, “Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis”, Springer, Neural Computing and Applications, Volume 26, Issue 2, February 2015, Pages 1437–446.
10. Mei Wang, Daniel Klevebring, Johan Lindberg, Kamila Czene, Henrik Grönberg and Mattias Rantalainen, “Determining breast cancer histological grade from RNA-sequencing data”, Springer, Breast Cancer Research, December 2016, Pages 1-13.
11. Saba Bashir & Usman Qamar & Farhan Hassan Khan, “WebMAC: A web-based clinical expert system", Springer, Information Systems Frontiers, Pages 1–17.
12. Min-Wei Huang, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, Chih-Fong Tsai, “SVM and SVM Ensembles in Breast Cancer Prediction”, plus one, Volume 12, Issue 1, January 2017, Pages 1-14.
13. Khalid Usman, Kashif Rajpoot, “Brain tumour classification from multi-modality MRI using wavelets and machine learning”, Springer, Pattern Analysis and Applications, Volume 20, Issue 3, August 2017, Pages 871–881.
14. Idil Isikli Esener, Semih Ergin, and Tolga Yuksel, “A New Feature Ensemble with a Multistage Classification Scheme for Breast Cancer Diagnosis”, Hindawi, Journal of Healthcare Engineering, Volume 2017, June 2017, Pages 1-16.
15. Qi Wang, ZhiHao Luo, JinCai Huang, YangHe Feng, and Zhong Liu, “A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM”, Hindawi, Computational Intelligence and Neuroscience, Volume 2017, January 2017, Pages 1-12.
16. Gregory P. Way, Robert J. Allaway, Stephanie J. Bouley, Camilo E. Fadul, Yolanda Sanchez and Casey S. Greene, “A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma”, Springer, BMC Genomics 2017, February 2017, Pages 1-11.

17. Jinyu Cong, Benzheng Wei, Yunlong He, Yilong Yin, and Yuanjie Zheng, "A Selective Ensemble Classification Method Combining Mammography Images with Ultrasound Images for Breast Cancer Diagnosis", Hindawi, Computational and Mathematical Methods in Medicine, Volume 2017, June 2017, Pages 1-8.
18. Subrata Kumar Mandal, "Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree", International Journal Of Engineering And Computer Science, Volume 6, Issue 2 February 2017, Pages 20388-20391.
19. Milan Joshi, Anurag Joshi, "On Comparative Study of Breast Cancer Classification Using Ensembles in Statistical Modelling", International Journal of Computer Science and Technology, Volume 8, Issue 1, January - March 2017, Pages 18-21.
20. Fengying Xie, Haidi Fan, Yang Li, Zhiguo Jiang, Rusong Meng, and Alan C. Bovik, "Melanoma Classification on Dermoscopy Images using a Neural Network Ensemble Model", IEEE Transactions on Medical Imaging, Volume 36, Issue 3, March 2017, Pages 849 – 858.