ISSN 2228-9860 eISSN 1906-9642 CODEN: ITJEA8

International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies

http://TuEngr.com

Designing a Hybrid Model Using HSIC Lasso Feature Selection and AdaBoost Classifier to Classify Image Data in Biomedicine

Mukesh Madanan [1], Anita Venugopal [2]

[1] *Department of Computer Science, Dhofar University, OMAN.*
[2] *IT Foundation Unit, Dhofar University, OMAN.*
**Corresponding Author (Tel: +966(16)404-2568 Email: AAlghamdi@mu.edu.sa).*

Paper ID: 12A1G

Volume 12 Issue 1

Received 30 June 2020
Received in revised form 19 October 2020
Accepted 28 October 2020
Available online 05 November 2020

Keywords:
Artificial Intelligence, Medical Imaging, Object-Based Classification, Support Vector Machine (SVM), WEKA; AdaBoost SVM Classifier.

Abstract

In cell-based research, an effective classification approach is required for visually monitoring a large quantity of image data of cells *in vitro* treatment. It is important to classify alive and dead cells likewise in tumor cell images, detecting virus-cell images, etc. to analyze patients' situation and then provide patient-centered care. Traditionally, the classification methods employed for classifying the cell microscopy data is time-consuming and is susceptible to faults and delusion. This is a serious and crucial dilemma. Accurate classification of data set is a major task in cell-based research as it determines the treatment. This paper introduces a hybrid model that uses a nonlinear HSIC Lasso feature selection method combined with the AdaBoost SVM Classifier to classify a large quantity of data effectively and efficiently. In the proposed model, object-based classification is executed within the bounds of the Waikato Environment for Knowledge Analysis (WEKA) interface. Besides, the accuracy of the classifier is evaluated by methods like feature selection and interactive learning in WEKA. The performance comparison of the proposed model amid existing classification approaches proved that the method is better in minimizing the mean absolute error value successfully.

Disciplinary: Computer Engineering, Biomedical Technology.

Cite This Article:

Madanan, M., Venugopal, A. (2021). Designing a Hybrid Model Using HSIC Lasso Feature Selection and Adaboost Classifier to Classify Image Data in Biomedicine. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies, 12*(1), 12A1G, 1-14. http://TUENGR.COM/V12/12A1G.pdf DOI: 10.14456/ITJEMAST.2021.7

1 Introduction

Cell image analysis plays a vital role in medical imaging after the invention of optical microscopes. During analysis, it is very important to classify the Alive and Dead cells to provide a

suitable diagnosis for patients' likewise in tumor cell images and detecting virus-cell images, etc. in the samples collected. Conventionally, manual applications are used to perform the investigation in microscopic imaging via a compact count of experimental facilities. In this case, a manual investigation of thousands of microscopy images, be that as it may, is tedious and prone to error. Hence, there is a need to employ computerized devices and techniques. Nowadays, researchers are giving more attention to the computerized framework and advanced techniques to enhance the efficiency in microscopic image analysis (Xing et al., 2017).

In recent years, morphological cell analysis is a developing new methodology to perform cell image processing or pattern recognition in a computerized manner. Correspondingly, it has integrated with many frameworks in biomedical applications such as evaluation of histological tumor sections, analyzing the characteristics of morphological biomedical cells, indicating cell morphology in various cell cycle progression or grasping the drug influences and chemotactic responses (Chen et al., 2012). However, morphological cell analysis has the challenge of identifying and classifying the cell growth variations of a large number of microscopic image data in visual monitoring of the cell-based vitro method.

Generally, the analysis of microscopy images has a major task of extracting the features and classifying the data from large image data set. Most of the state of art of image analyzing systems is tending to be expensive, complex, and hard to grasp (Baatz, Arini, Schäpe, Binnig, & Linssen, 2006). Thus, Machine Learning (ML) is developed for automatic image classification to classify the shape of living cells (Li et al., 2019). Still, the performance of the classifier can be enhanced by reducing the various surplus features (Popescu & Sasu, 2014). Moreover, it can reduce the redundancy to obtain high predictive features and interpretability. Also, to achieve accurate feature selection (Fan et al., 2004) and to minimize the unbalanced classification or prediction accuracy in image processing is a challenging one. Therefore, the ensemble learning boosting technique is a sophisticated solution for minimizing the errors in ML classifier to ensure performance accuracy (Dietterich, 2000). It can effectively unite various weak classifiers into a well-built classifier, which can attain a subjectively low error rate (Sagi & Rokach, 2018). Besides, by using the boosting algorithm, the impact of prediction and computation time is enhanced (Pavlov et al., 2002).

In recent studies, many frameworks have been developed to resolve cell microscopy image classification issues. However, the results are not satisfactory in large microscopy image data. In this paper, an effective feature selection method with an AdaBoost SVM classifier to easily identify the Dead and Alive cell from large datasets based on the object-based classification method with minimum error is presented.

2 Literature Review

Various feature selection and ensemble learning algorithms developed in existing researches are reviewed in this section. Peng et al. (2010) presented the feature selection method as a Sequential Forward Floating Search (SFFS) to prevail over the drawback of filter and wrapper method that has a high cost, low computational, and classification loss. They analyzed the

performance of classification by improving the search of the feature subset through the preselection step and then evaluated the achievement of single features and feature subsets of classification via Receiver Operating Characteristics (ROC) curves and this method efficiently solved the overfitting problem. But this method did not perform well while reducing the errors in classification besides it necessitates great computational power. Theoretical analysis of the minimal-Redundancy-Maximal-Relevance (mRMR) combined with the wrapper feature selection method (Peng et al., 2005) was introduced to minimize redundancy and it showed that maximal dependency condition is equal for feature selection and they analyzed different classifiers with various datasets. The analysis results proved that the accuracy has been enhanced but it lacks in the performance of large data analysis due to higher computation time. The Fast Correlation Based on Filter [FCBF] (Yu & Liu, 2003) approach was developed to reduce the redundancy to a sufficient level with fast computation. This method does not deal with the high dimensionality of data. To reduce the noise or redundancy, Sparse Additive Models (SpAM) were introduced (Ravikumar et al., 2009). Accordingly, the back-fitting algorithm was not supported to minimize the high-dimensional feature selection issues and it obtained nonparametric regression and classification. Further, dealing with non-additive models were not explained adequately. To override this, the Spectral features selection method (Wang et al., 2016) was presented to select features based on spectral clustering and l1-Norm Graph jointly. Lack of manifold structure, Unsupervised Spectral Feature Selection with l1-Norm Graph algorithm was optimized. It reduced the redundancy or noise for high-dimensional data in an excellent manner. Nevertheless, it supported only the unsupervised method effectively. To deal with unsupervised or supervised methods along with high dimensional data, Lasso was presented (Tibshirani, 2014). Lasso penalties approaches were useful for fitting to find out the drawback in low and large dimensional feature selection with (e.g. $n < 100$ and $d > 104$), l1 regularized. In addition to this, Lasso was used to supporting linear regression, and consequently, high prediction and accuracy were obtained. Correspondingly, HSIC Lasso was implemented to take over non-linearity (Takahashi, et al., 2020).

The sequential minimal optimization is an algorithm that offered to do the training in a faster manner in Support Vector Machine (Kotsiantis, 2007) which is used for minimizing the noise in feature data and enhance computational efficiency. The research aimed to discover a boundary, to maximize the margin connecting dissimilar data points for the splitting up by Sequential Minimal Optimization (SMO) Algorithm and also cooperating with non-linear data. However, the error was not handled at the requisite level. The ensemble method was presented to construct the classifier and to gain high accurate predictions while classifying the data by weighting the vote manner (Dietterich, 2000). In this way, high accuracy classification was achieved by constructing the correlations among input attributes using an ensemble Bayesian network in microarray data (Zhang & Hwang, 2003). Nevertheless, it did not provide support for nonlinear data analysis. The authors proved that the AdaBoost classifier well performed in error-correcting when compared with traditional state and art methods. Based on this, the shape of living cells microscopy images was

analyzed by Naïve Bayes Classifier with AdaBoost (Theriault, Walker, Wong, & Betke, 2012). Thereupon, cluster mitigation was reduced and it obtained a classification in better form by minimizing false detection. This framework had high accuracy of classification but it supported only the linear model.

To enhance the classification in the machine learning method, previous researchers used various feature selection methods and classifiers but to solve the challenges in large data classification still seemed an issue. The research paper aims to introduce a significant classification methodology to provide an efficient classification of large quantities of data.

3 Classification

Based on the described existing methods, an efficient classification is needed for nonlinear high dimensional data. Also, it should be to reduce the noise, minimize time consumption, select the best feature, and avoid overfitting. To solve these constraints, the research aims at reducing the Mean Absolute Error in the proposed work. To follow the efficiency treatments in vitro, some metrics that are impacted by noise, time-consuming, misclassification is used to calculate the performance of classification in a large amount of cell microscopy images are revealed as follows.

3.1 Precision

Precision, which is a metric, is distinct as the total number of true positive divided by the sum of false positives and true positives. In biomedicine, Precision is called Positive Predictive Value. It is used to find the number of correct predictions. In classification, a low false-positive prediction means error or classification loss which is reduced the performance of classification.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

3.2 F-Measure

The harmonic (noise) mean of Precision and Recall is F – measure (Hand & Christen, 2018). It is denoted as F that is a function of Precision and Recall. It is used to measure the incorrectly classified cases in classification.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3) \quad (3)$$

3.3 Area under ROC

AUC (Area under the Curve) and the ROC (Receiver Operating Characteristics) curves combination are called as Area Under the Receiver Operating Characteristics (AU ROC). This evaluation metric is used to determine all possible classification thresholds. The AU ROC curve is based on the True Positive Rate against the False Positive Rate. AU ROC ranges value should be from 0-1. If the prediction is efficient, the value will be 1.0.

3.4 Time-Consuming

The performance of classification is affected by higher time-consuming. Thus, a fast manner calculation is important to analyze the data.

Cell-based research, as in tumor cell image and detecting the virus-cell images, is important to classify the Alive and Dead cell to detect the growth and un-growth or dead. Particularly in nonlinear high dimensional data, the efficient classification is a challenging one due to classification errors and misclassification. For attaining better classification results, we need to develop an accurate classifier. Thus, this research focuses on proposing a nonlinear HSIC Lasso feature selection method combined with AdaBoost SVM Classifier to reduce errors and misclassification efficiently.

4 The Proposed Hybrid Model

In microscopy cell image analysis, larger data classification is a major obstacle and most of the extant classifiers were unsuccessful to reduce the classification errors and are time-consuming. To solve this issue, the research introduces an object-based classification (Liu & Xia, 2010) method using nonlinear HSIC Lasso feature selection in conjunction with AdaBoost SVM Classifier to reduce the prediction error and time-consumption efficiently. This object-based classification is done using WEKA (Frank, et al., 2009) to evaluate classifier performance.

The microscopy images consist of a lot of noise and it creates a distortion of images in most cases. Besides, the noise-effected images minimize the accuracy of classification in a vulnerable way. To enhance the image quality, initially, the large cell-based image data is given as an input for pre-processing and the specific features are enhanced through correction of error and conversion of an image into an ideal format using a mathematical model. Further, in the feature selection process, the feature co-efficiency of discriminative features are found by separating the samples from different subsets. To achieve efficient feature selection, the paper proposes an HSIC LASSO feature selection algorithm to choose the best features from the training database by eliminating the redundancy features. Then, testing data and the selected features from the training data are fed into AdaBoost SVM classifier for the classification. In the classification process, the hypothesis (h (t)) is calculated by the SVM algorithm. Later with the help of AdaBoost, the training error ε_r and estimation of α_t is calculated and the weighted vectors are updated to obtain the weight of the hypothesis in the SVM classifier. AdaBoost could keep up the distribution weight of SVM iteratively and expanding its precision. Finally, the dead and live-cell data are classified with less computational time. Besides this, the Mean Absolute Error is calculated using an analysis of the metrics such as Precision, F- Measure, and Area ROC. The proposed hybrid model is depicted in Figure 1.

4.1 HSIC Lasso Algorithm

In 1996, Robert Tibshirani established the LASSO - Least Absolute Shrinkage and Selection Operator for regression or classification (Gauraha, 2018). LASSO can perform regression and feature selection in a powerful manner (Gauraha, 2018). Hence, LASSO feature selection is used to

find the admissible features in high dimensional data and facilitates to avoid redundancy and overfitting. Besides, it can achieve good prediction accuracy although it is supporting the linear data only. The research work analyzed the microscopic data in nonlinear methods. To do this, the Hilbert-Schmidt Independence Criterion Lasso (HSIC Lasso) (Yamada, et al., 2018) was employed in this research to support the non-linear high dimensionality microscopic cell image dataset.

The problem of optimization is exposed to Lasso as

$$\min_{\alpha \epsilon \mathbb{R}^d} \frac{1}{2} \|y - X^T \alpha\| \frac{2}{2} + \lambda \|\alpha\|_1 \quad (4),$$

where $\alpha = [\alpha_1 \dots \alpha_d]^T$ is a regression coefficient vector, α_k indicates the regression coefficient for the k[th] feature and $\lambda > 0$ is the regularization parameters. The feature base non-linear Lasso was proposed (Zhang et al., 2016), to get sparsely regarding features. The non-linear transformation is achieved through the feature-wise analysis. More explicitly, sample matrix X is obtained in a feature-wise aspect,

$$X = [u_1 \dots u_d]^T \in \mathbb{R}^{d \times n} \quad (5),$$

where $u_k = [x_{k,1} \dots x_{k,n}]^T \epsilon \mathbb{R}^n$ denotes the k-th feature's vectors. At that point, using the nonlinear function $\varphi(.): \mathbb{R}^n \to \mathbb{R}^p$, the feature vector of u_k and the output vector of y is transformed.

Then, the nonlinear Lasso based feature which also called HSIC Lasso[2] is

$$\min_{\alpha \epsilon \mathbb{R}^d} \frac{1}{2} \left\| \overline{L} - \sum_{k=1}^{d} \alpha_k \overline{K}^k \alpha \right\|_{Frob}^2 + \lambda \|\alpha\|_1 \quad (6).$$

Useful features are selected using non-negativity constraints as "α". Forasmuch as we utilize the output Gram matrix L to choose features in HSIC Lasso and organize the outputs via kernels. Besides, we can execute feature selection regardless of whether the training data set comprises of input x and its affinity information L, for example, connect structures amid inputs.

By using the linear combination of feature-wise input kernel matrices $\{\overline{K}^{(k)}\}_{k=1}^{d}$, regressing the output kernel matrix $\overline{L}$ is got through in Equation (6). We represent that minimum redundancy maximum relevancy (mRMR) hinged on the feature selection method for HSIC Lasso, which is a well-known feature selection procedure in ML and AI communities. Considering this, Equation (6) can be composed as

$$\frac{1}{2} \left\| \overline{L} - \sum_{k=1}^{d} \alpha_k \overline{K}^k \alpha \right\|_{Frob}^2 = \frac{1}{2} HSIC(y,y) - \sum_{k=1}^{d} \alpha_k HISC(u_{k,y}) + \frac{1}{2} \sum_{k,l=1}^{d} \alpha_k \alpha_l HISC(u_k, u_l) \quad (7),$$

where $(u_k, y) = \mathrm{tr}(\overline{K}(k)\overline{L}$ denotes empirical HSIC which is impedance matching depending upon kernel. The constant value of HSIC(y, y) is possible to be unnoticed. Additionally, if redundant features are u_k, u_l, HSIC (u_k, u_l) holds a huge value and in this manner both of α_k and α_l will in general be zero. This process implies that the redundant features wiped out by HSIC Lasso. Thus, HSIC Lasso is lead to find non-redundant features based on (mRMR) feature selection methods (Ding & Peng, 2003).

The input of the Gaussian kernel is a desirable characteristic in the feature selection method. Computing the computational characteristic is so important. This property with HSIC Lasso using

$$\frac{1}{2}\left\|vec\,(\bar{L}) - \left[vec\,(\bar{K}^{(1)}, \ldots vec\,\left(\bar{K}^{(d)}\right)\right]\alpha\right\|_2^2 \qquad (8)$$

where $vec\,(\,)$ is noted as vectorization operator. This method is expensive when features are lower than the number of the sample (n). Therefore, the table peruse method is introduced to minimize the computation time and cost.

4.2 AdaBoost SVM Classifier

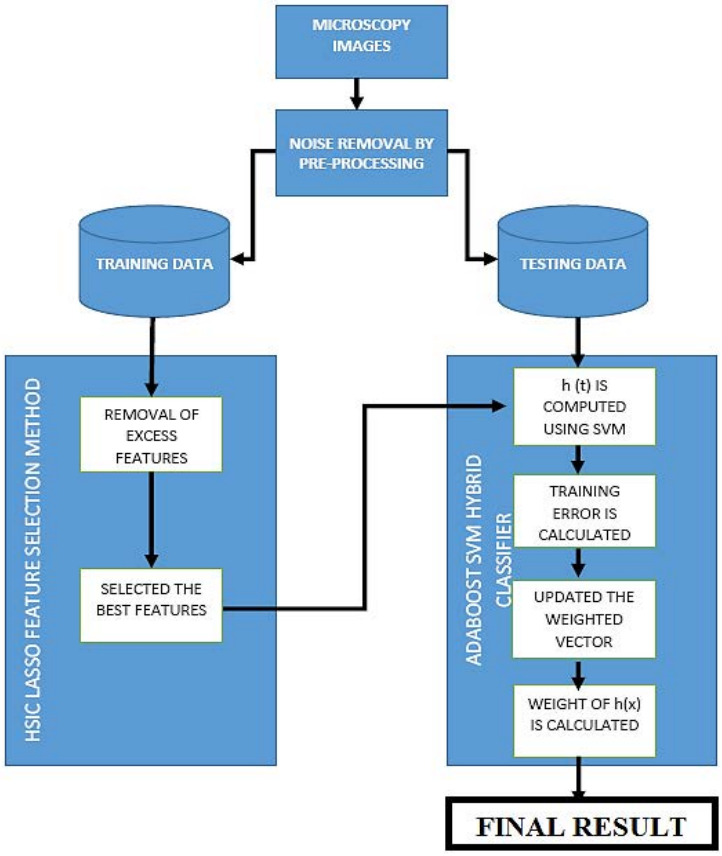

Figure 1: Proposed Model-HSIC LASSO with AdaBoost Hybrid Classifier.

The research proposes a hybrid model classifier (Ganganwar, 2012) to increase the performance of classification. Hence, the AdaBoost method with the SVM classifier is used as the base classifier. AdaBoost takes over the hypothesis weighting of the SVM method to acquire enhanced precision. The weight in misclassification error was enhanced in every cycle, the weight on the already well classified were minimized and leads to minimizing the potential weighted back in the subsequent cycle. Thus, the class (label) of hypothesis h_t was predicted.

4.3 Support Vector Machine (SVM)

Vapnik (1995) Support Vector Machine (SVM) Classifier can be used to find the decision surface which is located at a far distance from any data point. The distance amid the decision surface to the nearest data point creates the verge of the classifier. This method of development necessarily implies the decision function for an SVM and it is completely indicated by a subset of

the data which characterizes the location of the separator. These points are also known as support vectors. The Support Vector Machine is represented in Figure 2.

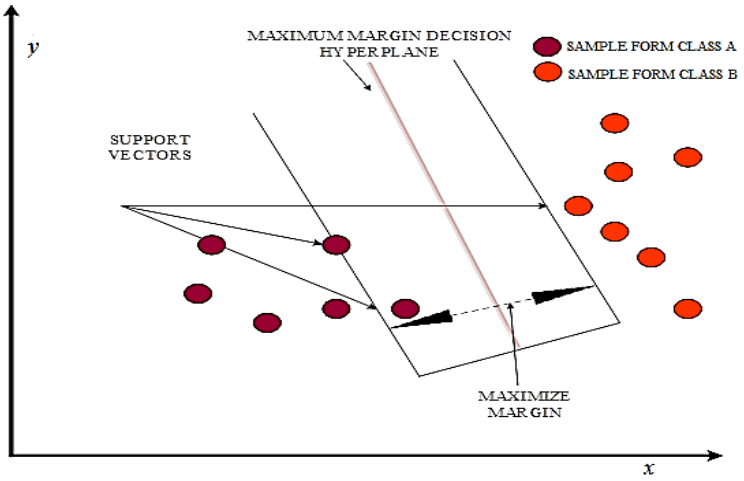

Figure 2: Support Vector Machine

It is finding the N number of features through the hyper-plane of the SVM algorithm that particularly classifies the data points. In classification, the major play of SVM is to assemble a hyperplane that can enhance the margin, the distance from the hyperplane to the nearest data. The larger margin generates a small error (Panca & Rustam, 2017). The margin was the nearest distance amid hyperplane to the closest point of each class (support vectors). Form of equation delineating the decision surface separating the classes is a hyperplane of the form as,

$$w^T x + b = 0 \tag{9}$$

where, w, x, b denotes weight vector, input vector and bias. The optimal hyperplane in SVM is shown in Figure 3.

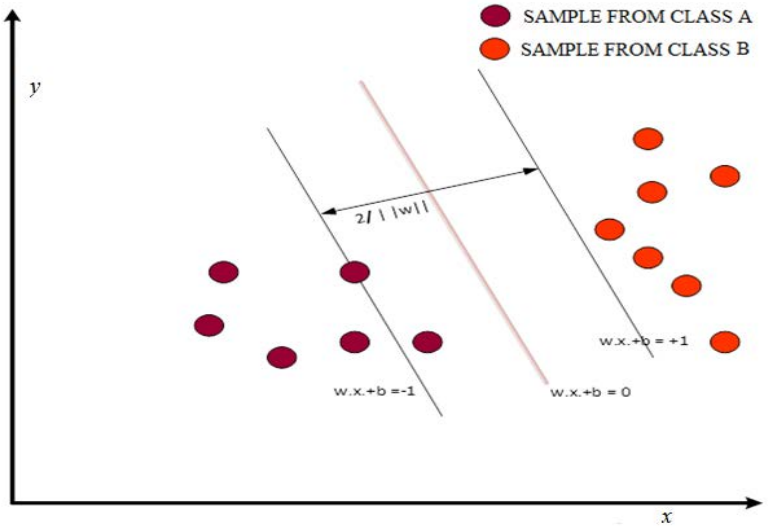

Figure 3: Optimal Hyperplane in SVM

The value of w and b are the findings through Quadratic Programming as shown in the mathematical model,

$$\min_{w,\,b} \frac{1}{2} \|w\|^2 \tag{10}$$

So that $y_i\ (w^T x_i + b\ \geq 1, i = 1, \ldots, n$

In this circumstance, the SVM finds and enhances the margins of the hyperplane to limit the classification loss. By adding C parameter and slack variable for classification error scenarios, the SVM mathematical model denotes

$$\min_{w,\,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \epsilon_i \qquad (11).$$

So that $y_i\ (w^T.x_i + b\ \geq 1 - \epsilon_i, \epsilon_i \geq 0, i = 1, \ldots n$

When enhancing margin, the algorithm attempts to keep the slack variable to zero (C >0). Nonetheless, it does not limit the number of classification loss (NP-entire issue) but affects the total distance from the margin hyperplanes. C is signified as a trade-off margin width and classification loss. The kernel function includes the key idea to obtain linearly non-separable facts. The Kernel function is

$$K\left(x_i, x_j\right) = \emptyset\ (x_i{}^T).\emptyset\left(x_j\right) \qquad (12).$$

It expresses a non-linear function and is obtained by a linear learning machine in a high-dimensional feature space while the limit of the system is constrained by a parameter that does not hang on the dimensionality of the space.

4.3.1 AdaBoost Algorithm

Yoav Freund and Robert Schapire presented AdaBoost in 1995 (Chengsheng et al., 2017). This technique has the object to keep a weight distribution **w** of the base classifier. Here, the learning algorithm takes a function from the hypothesis class, which is the set of possible classification functions. The ensemble method of AdaBoost can magnify the classification outputs by building a lot of classifiers and consolidating it. Then execute the base classifier training repeatedly for several cycles (1,2, ...T) with a given dataset. Starting weight vector $\mathbf{w}^1$ in this training was arranged equivalent to

$$\mathbf{w}_i^1 = \frac{1}{m}, \qquad i = 1,2, \ldots m \qquad (13).$$

To get the exact result, the weighted vector is updated for each iteration. Finding hypothesis ht = {-1,+1} for w_i is a major task for the base classifier in this level and by calculating the training error ε_t to measure the quality of the hypothesis,

$$\varepsilon_t = \sum_{i=1}^{m} w_i{}^t \qquad y_i \neq h_t\ (x_i) \qquad (14).$$

In this manner, training error is determined from a trained weighted vector. This process is repeated until$\varepsilon_t > 0.5$. By limiting the estimation of ε_t, the expanded estimation of α_t is attainable as follow,

$$\alpha_t = \frac{1}{2}\, In\ (1 - \varepsilon_t | \varepsilon_t) \qquad (15)$$

Thus, updating the weighted vector $w_i{}^t$ is done. The result of the hypothesis depends upon the number of weights of the hypothesis in the base classifier as

$$H\ (x) =\ sign\ (\ \sum_{t=1}^{T} \alpha_t h_t\ (x)) \qquad (16).$$

4.3.2 AdaBoost SVM Algorithm

The data set with the SVM algorithm with the number of cycles is provided as input. Then initializing the weight of the training sample is carried out and iteration is done until the last cycle. Based on the weighted training sample the hypothesis h_t is calculated using the SVM algorithm. The training error of ε_t is calculated using

$$\varepsilon_{t=}\ \sum_{i=1}^{m} w\ {}_i^t, y_i\ \neq\ \ h_t\ (x_i) \qquad (17).$$

This process is continued until If ε_t > 0,5 and stop it. Then set the weight for hypothesis h_t,

$$\alpha\ _t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \qquad (18).$$

The weights of the training samples are updated too

$$W_i{}^{t+1} = \frac{W_i{}^t \exp\{-\alpha_t y_i h_t(X_i)\}}{Z_t} = \frac{W_i{}^t}{Z_t} \times \begin{cases} \exp\ \{-\alpha_t\},\ y_i = h_t(X_i) \\ \exp\ \{\alpha_t\},\ y_i = h_t(X_i) \end{cases} \sum_{i=1}^{m} w_i{}^{t+1} = 1 \qquad (19),$$

where Z_t is normalization constant.

The result of the hypothesis in (**x**) depend on the number of weights (*T*) hypothesis of the base classifier expressed as

$$H(x) = sign\ (\sum_{t=1}^{T} \alpha_t h_t\ (x)) \qquad (20).$$

5 Results and Discussions

In this paper, the performance of the classifier is calculated by adopting the True/ False Positive Rate, F-measure, and receiver operating characteristic (ROC) area. The proposed method executes the huge number of the True Positive (TP) Rate and exactly marked the cell as ALIVE or DEAD divided by the sum of instances in the test set. Besides, many times the base classifier wrongly predicted the cell as ALIVE or DEAD, which are divided by the total number of instances in the test set and it has known as False Positive (FP). Further, the proposed model developed a ROC curve, through making the relationship of the True Positive vs False Positive of each classification threshold. In this, The Area Under ROC (AUROC) is used to calculate the accuracy rate of the classifier by changing the threshold value based on the ROC curve value. Then, the F-measure metric is used to analyze the classifier accuracy by computing the harmonic average of precision and recall. For perfect accuracy, the F-measure is on a scale of 0-1, with 1. The performance is higher if the F- measure value is on a scale of 0-1 or within 1.

WEKA's Explorer Environment is used to examine the performance of classifiers depending on the object-based model. Feature selection is a filter operation in WEKA. To evaluate the feature

selection method, ensuring the division of training data and testing data is important. The training data is designed using Sparse Attribute-Relation File Format (ARFF) file in WEKA (Bouckaert, et al., 2008) for the classifier. The suitable subsets of features are found through all possible combinations of attributes in the data. Thus, minimizing the excess features in the dataset and it is fed into the classifier. The WEKA's built-in algorithm for CV Parameter Selection was employed to object-based models. The meta-classifier of CV Parameter Selection tunes the parameter automatically as the base classifier. A specific range of values is given to perform the process frequently. The accurate value for the parameters is selected by CV within the provided range. For each object-based model, the batch size was tested over the range of 10 to 150. The tuning features are fed into WEKA's classifiers as input and the classifier performance is analyzed using ten-fold cross-validation belonging to the corresponding training set. The proposed model is compared with four existing methods and the comparison results proved that the proposed model outperformed while comparing with others in terms of False Positive Rate, area ROC, Precision, and F- Measure value.

As depicted in Table 1, the AdaBoost SVM classifier had 98% effectively classified instances. However, with Total Positive Rate, the value expanded to 98% compared with the existing classifier. The AdaBoost SVM classifier method has the build time of 0.01sec, which is equal to the Optimized SVM classifier although our proposed work has a better build in time. The AdaBoost SVM classifier reduces the False Positive Rate to a better level. The already existing classifier did not compensate AdaBoost SVM classifier in terms of precision or F-measure. Figure 4 represents the Area ROC values of Random Forest, Bayesian Network, SMO with SVM, and AdaBoostSVM algorithms.

Table 1: Result for Object Based Classifier Performance

CLASSIFIER	True Positive Rate	False Positive Rate	Area ROC	Precision	F-Measure	Time
Random Forest	0.90	0.10	0.967	0.90	0.90	0.04
Bayesian Network	0.90	0.10	0.939	0.901	0.90	1.41
Optimized SVM	0.95	0.05	0.951	0.95	0.95	0.01
AdaBoost SVM Classifier	0.98	0.03	0.985	0.98	0.98	0.01

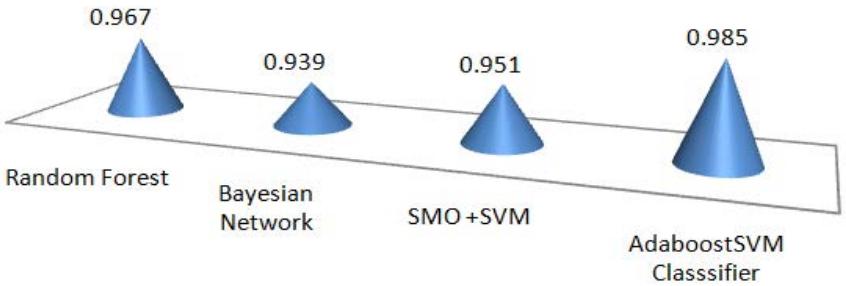

Figure 4: Area ROC Comparison

Table 2: Error Score Result for Classifiers

CLASSIFIER	Mean Absolute Error	Relative Absolute Error	Root Relative Squared Error
Random Forest	0.26	52%	61%
Bayesian Network	0.17	34.44%	80.98%
Optimized SVM	0.06	12%	48.98%
AdaBoost SVM	0.04	10%	41.25%

The high value of Area ROC indicates AdaBoost SVM classifier indicates better accuracy in classification as shown in Figure 4 Our AdaBoost SVM classifier is providing the most elevated accuracy in classification through enhancement of AUROC and effectively classified instances. The error score of the classifier is shown in Table.2. The measurement of error such as Mean Absolute Error, Relative Absolute Error, and Root Relative Squared Error for Existing classifier of Random Forest, Bayesian Network, Optimized SVM, and AdaBoost Classifier is calculated.

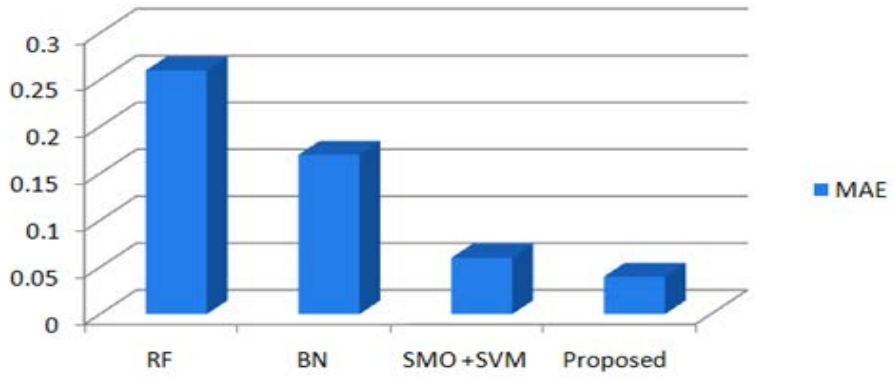

Figure 5: Mean Absolute Error Comparison of Classifiers

It is quite evident that the finest performance is accomplished by employing the proposed AdaBoost SVM classifier by reducing the Mean Absolute Error efficiently. The detection of Dead and Alive cell prediction is increased by reducing the Mean Absolute Error. Thus, the classification accuracy is improved compared with the accuracy of the existing classifier. The Mean Absolute Error comparisons are given in Figure 5.

6 Conclusion

This paper proposes a nonlinear HSIC Lasso feature selection method combined with AdaBoost SVM Classifier to obtain better classification results. In this work, the redundant features are reduced by adding the nonlinear HSIC Lasso feature selection method based on the Minimal-Redundancy-Maximal-Relevance method. Sequentially, the selected features are fed into an AdaBoost-SVM classifier. The AdaBoost algorithm boosts the performance of the classifier by adjusting the hypothesis weighting in SVM. The proposed method showed lessor error scores and the highest accuracy compared to its counterparts. Most of the existing methods failed to give better performance in huge dataset analysis although our proposed method gave the enhanced performance in terms of F-measure, Precision, and Area ROC. Also, the approach minimizes the Mean Absolute Error. This object-based classification performance is evaluated with the existing classifiers Random Forest, Bayesian Network, Sequential Minimal Optimized Support Vector Machine, and the proposed model outperformed well comparing with the existing methods. The research does not study the detailed structure of the cells while classifying. Using advanced image analysis and classification deep learning techniques such as CNN could enhance the results. For feature enhancements, very large data could be classified using neural network and pixel-based classification and can be implemented by using the feature extraction method.

7 Availability of Data and Material

Information can be made available by contacting the corresponding author.

8 References

Baatz, M., Arini, N., Schäpe, A., Binnig, G., & Linssen, B. (2006). Object-oriented image analysis for high content screening: Detailed quantification of cells and subcellular structures with the Cellenger software. *International Society for Analytical Cytology*, 652-658.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2008). *Weka manual for version 3-6-0*. Hamilton: University of Waikato.

Chen, S., Zhao, M., Wu, G., Yao, C., & Zhang, J. (2012). Recent Advances in Morphological Cell Image Analysis. *Computational and mathematical methods in medicine*.

Chengsheng, T., Huacheng, L., & Bing, X. (2017). AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*. EDP Sciences.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *International Workshop on Multiple Classifier Systems* (pp. 1-15). Berlin Heidelberg: Springer.

Ding, C., & Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference* (pp. 523-528). Stanford: IEEE.

Fan, Z.-G., Wang, K.-A., & Lu, B.-L. (2004). Feature selection for fast image classification with support vector machines. *International Conference on Neural Information Processing* (pp. 1026-1031). Berlin, Heidelberg: Springer.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I., & Trigg, L. (2009). Weka-A machine learning workbench for data mining. In O. Maimon, & L. Rokach, *The Data Mining and Knowledge Discovery Handbook* (pp. 1269-1277). Boston: Springer.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 42-47.

Gauraha, N. (2018). Introduction to the LASSO: A Convex Optimization Approach for High-dimensional Problems. *Resonance*, 439-464.

Hand, D., & Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 539-547.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval, and Pervasive* (pp. 3-24). ACM-Digital Library.

Li, C., Xue, D., Hu, Z., Chen, Y. H., Yao, U., Zhang, Y., Xu, N. (2019). A Survey for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks. *International Conference on Information Technologies in Biomedicine* (pp. 222-233). Springer.

Liu, D., & Xia, F. (2010). Assessing object-based classification: advantages. *Remote Sensing Letters*, 187-194.

Panca, V., & Rustam, Z. (2017). Application of machine learning on brain cancer multiclass classification. *International Symposium on Current Progress in Mathematics and Sciences* 2016*(ISCPMS 2016):*. Depok, Jawa Barat, Indonesia: AIP Publishing.

Pavlov, D., Mao, J., & Dom, B. (2002). Scaling-up support vector machines using a boosting algorithm. *International Conference on Pattern Recognition ICPR*. Spain: IEEE Xplore.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

1226-1238.

Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*, 15-23.

Popescu, M. C., & Sasu, L. M. (2014). Feature extraction, feature selection and machine learning for image classification: A case study. *International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)*. Bran, Romania: IEEE.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse Additive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1009-1030.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Takahashi, Y., Uek, M., Yamada, M., Tamiya, G., Motoike, I. N., Saigusa, D., . . . Tomita, H. (2020). Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. *Translational psychiatry*, 1-12.

Theriault, D. H., Walker, M. L., Wong, J. Y., & Betke, M. (2012). Cell morphology classification and clutter mitigation in phase-contrast microscopy images using machine learning. *Machine Vision and Applications*, 659–673.

Tibshirani, R. (2014). Regression shrinkage and selection via the lasso: A Retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 273-282.

Wang, X., Zhang, X., Zeng, Z., QunWu, & JianZhang. (2016). Unsupervised spectral feature selection with l1-norm graph. *Neurocomputing*, 47-54.

Xing, F., Xie, Y., Su, H., Liu, F., & Yang, L. (2017). Deep learning in microscopy image analysis: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 4550-4569.

Yamada, M., Tang, J., Lugo-Martinez, J., Hodzic, E., Shrestha, R., Saha, A., . . . Yin, D. (2018). Ultra high-dimensional nonlinear feature selection for big biological data. *IEEE Transactions on Knowledge and Data Engineering*, 1352-1365.

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the Twentieth International Conference (ICML)*, USA.

Zhang, B.-T., & Hwang, K.-B. (2003). Bayesian network classifiers for gene expression analysis. *A Practical Approach to Microarray Data Analysis*, 150-165.

Zhang, Y., Guo, W., & Ray, S. (2016). On the consistency of feature selection with lasso for non-linear targets. *The 33rd International Conference on Machine Learning*, (pp. 183-191). ML Research Press.

Mukesh Madanan is a Senior Lecturer of Computer Science at Dhofar University, Oman. He received his B.Tech in Computer Science and Engineering from M.G.University, India and got his MSc.in Software Engineering from the University of Portsmouth, UK. He is currently pursuing PhD in Information & Communication Technology at UNITEN, Malaysia. His areas of research include Machine Learning, Deep Learning, Robotics, Software Methodologies, IoT and Computer Networks.

Anita Venugopal is an IT Lecturer at Dhofar University, Oman. She is pursuing her Ph.D. in Computer Science at Motherhood University, India. The areas of research are Artificial Intelligence, Networks, and Knowledge Management.