Comparative Analysis of Machine Learning Algorithm for Classification of different Osteosarcoma types

Sanket Mahore
Dept. of Instrumentation and Control
College of Engineering Pune
Pune, India
sanketam19.instru@coep.ac.in

Kalyani Bhole
Dept. of Instrumentation and Control
College of Engineering Pune
Pune, India
kab.instru@coep.ac.in

Shashikant Rathod
Dept. of Instrumentation and Control
College of Engineering Pune
Pune, India
rsr18.instru@coep.ac.in

2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579556

***Abstract*—Histopathological osteosarcoma data analysis gives a lucrative way to study the pathological texture of Osteosarcoma. Osteosarcoma is classifying into Viable, Necrotic, and Non-Tumor. A Viable cell indicates a cancerous cell element and a necrotic cell means the cancerous cell nuclei kill by chemotherapy or radiation therapy. Necrosis is the process where cell injury results in the death of a cell. Classification of Osteosarcoma is a very time-consuming and complicated task due to its inter-class similarities and variations. Improvement in Machine Learning(ML) algorithm and Graphical Processing Unit(GPU) gives more accurate results for the classification and prediction of Osteosarcoma. In this paper, we use four ML algorithms for the classification of OST: Decision Tree(DT), Support Vector Machine(SVM), K-Nearest Neighbors(KNN) and AdaBoost(Adaptive Boosting). The ML model performance evaluation uses performance metrics like Accuracy, Sensitivity, specificity, F1-score, Kappa, and AUC (area under the curve). All four classifiers successfully classified Osteosarcoma into three types. The overall accuracy of DT, SVM, KNN and Adboost is 81.22%, 83.80%, 86.90% and 91.70% respectively. Adaboost algorithm outperforms the other three algorithms with overall accuracy 91.70%, sensitivity 91.60%, specificity 96.00%, F1 score 90.60%, Kappa 0.87 and AUC 0.99.**

***Index Terms*—Machine Learning, Osteosarcoma, SVM, KNN, Adaboost**

I. Introduction

Osteosarcoma is a type of bone cancer which is primitively present in Mesenchymal cell which produces Osteoid [1]. The origin of Osteosarcoma is unknown, but some researcher said that it derived from the mutation of genes which is responsible for bone formation. It is mainly found in the femur (42%), tibia (19%), and humerus (10%), skull or jaw(8%), and pelvis (8%) [2]. Osteosarcoma cells categorize into three main types like Viable, Necrotic, and Normal or Non-tumor cells. Viable cells contain a cancerous element that leads to the spread of cancer all over the body. The viable cell destroys by chemotherapy and radiation therapy, where the cell's nucleus gets killed. Cells commit suicide by a process known as Necrosis, and cell is known as Necrotic cell.

The diagnosis of Osteosarcoma performs using Imaging techniques like X-Ray, CT(Computed tomography), and MRI(Magnetic resonance Imaging). Apart from this, Histopathological Imaging also uses to diagnose the OST(Osteosarcoma). In this technique, the H&E(Hematoxylin & Eosin) [3] staining agent administrate into the cell, and nuclei get blue color while tissue gets pink or red. By analyzing the texture of H&E stain images under a digital microscope, the pathologist performance diagnosis. As shown in Fig. 1, Viable cell nuclei are densely aggregates, wherein the necrotic cell nuclei density gets reduce.

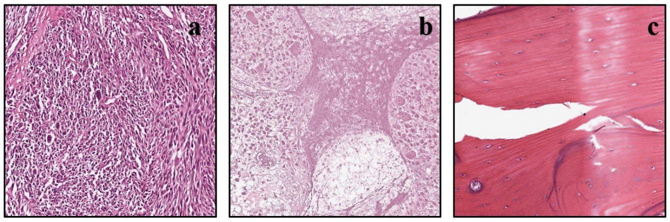

Fig. 1: Example showing the complexity of Histopathological images (a) Viable(Nuclei is densely aggregate together) (b) Necrotic(Necrosis, Fibrosis, and Osteoid) and (c)Non-Tumor (Bone, Cartilage)

Different modalities like CT, MRI, and Pathological Imaging use in the Healthcare industry to diagnose cancer. With the advancement in Deep Learning, Machine learning algorithms, and Graphics processing unit, data analysis is significantly and less time confusing with more accurate result [4]. The use of AI in the medical field increase, and it is not restricted to Medical image analysis but also use in Omics analysis and Natural language processing [4].

For segmentation and recognition using MRI modality, the researcher uses the Conditional Random Field (CRF) model, which uses features like texture. Combination of K-mean algorithm, Boosting algorithm and statistical method used with CRF which append features with grayscale values [5]. K-mean clustering uses color normalization for tumor isolation. Multi-threshold otsu segmentation technique uses to classify into viable and nobn-viable [6]. Rashika Mishra and her research team use histopathological images to train the CNN(Convolution Neural Network), which get overall accuracy of 92% [7]. Features extracted from Histopathological Images use to train the Machine learning model like Complex tree(CT), Support Vector Machine(SVM), and Ensemble learner(ENS). Overall accuracy of CT, SVM and ENS is 80.9%,89.9% and 86.8% respectively [8]. Bruno Korbar uses Histopathological images of Colorectal cancer and trains the Convolutional Neural Network model, which gives overall accuracy of 93% [9].

In this paper, we use four types of machine learning algorithms like DT, SVM, KNN, and Adaboost. In the previous study, more features are used, which take more time for training and the accuracy is more diminutive. To avoid this problem, we used fewer features and a trained model to get much better accuracy. The overall organization of the paper is as follows: Section II gives a lucrative idea about the dataset, machine learning algorithms, and performance evaluation matrix. Section III gives a comparative analysis of machine learning models with detailed discussion and Section IV shows the Conclusion.

II. Methodology

A. Block Diagram

The block diagram represents our machine learning approach's overall flow where data is used for the feature extraction process. The feature samples are pre-processed and divided into training and testing sets. The training set is used to train all four machine learning models and the testing set used to validate our ML models. Four machine learning algorithms are used like DT, SVM, KNN and Adaboost for classification and prediction. The validation of ML models is done by performance metrics which gives details statistical information of models.

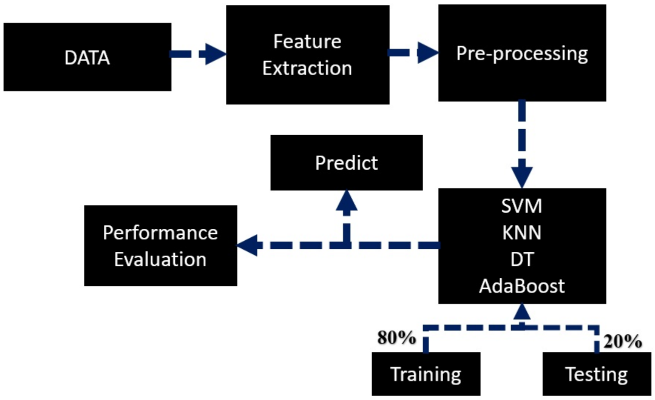

Fig. 2: Basic block diagram of Machine Learning Model

B. Dataset

We use Osteosarcoma features extracted from Histopathological images of Osteosarcoma by the researcher team of the University of Texas, Dallas. The Histopathological Image data [10] is publicly available on The Cancer Image Archive (TCIA) [11]. 1143 image samples are present where 535 Non-tumor, 340 Viable, and 268 Necrotic cells. Two types of features were extracts from images, i.e., Expert-guided feature and Cellprofilier features. The Expert-guided feature and Cellprofilier features were extracted from histopathological images using in-house software developed by UT, Dallas and Cellprofilier [12]. The Cellprofilier feature comprises second-order statistical features like entropy, variance, correlation and expert-guided feature include area, circularity, nuclei count and percentage of Red or Blue.

Feature type	Feature names
Expert-Guided	Total clusters,Red & Blue count Red & Blue percentage, Circularity and Area.
Cellprofiler	Gabor, sum Entropy, sum variance, variance sum average, correlation and angular second moment.

TABLE I: List of Expert-guided and Cellprofilier features extracted from OST images

We splited data into training and testing, where 80% data was used for training purposes and 20% for testing. As shown in TABLE II: the training set has 914 samples from which 428 is Non-Tumor, 265 Viable and 221 is Necrotic. Similarly, the testing set has 229 samples where Non-Tumor is 107; Viable is 75 and Necrotic is 47 samples. .

Image types	Training set	Testing set	Total
Non-Tumor	428	107	535
Viable	265	75	340
Necrotic	221	47	268
Total	**914**	**229**	**1143**

TABLE II: Number of data samples

C. Pre-processing

We normalize the data by data transformation techniques like MinMax normalization [13]. In this, all data values convert into a scale of 0 to 1. The following equation calculates the value:

$$X_n = \frac{X_i - min(X)}{max(X) - min(X)} max_n(X) - min_n(X) + min_n(X) \tag{1}$$

where,
X_n = new value
X_i = original value
$max(X)$ = maximum value of column
$min(X)$ = minimum value of column

D. Machine Learning Algorithm

Machine Learning is a subclass of Artificial Intelligence divided into Supervised, Unsupervised, and Reinforcement learning. We use supervised Machine learning algorithms like DT, SVM, KNN and Adaboost for classification. Labeled and structure data used for training and testing the machine learning models.

1) Decision Tree: DT is the primarily used ML algorithm for classification and regression problems [14]. It performs classification with the help of the else-if condition [15]. It consists of the root node, sub-tee, and terminal node where the root node is selected using ASM (Attribute Selective Measure). ASM is a data mining process with two main techniques, i.e., Gini Coefficient and Information Gain. Information Gain tells us about how the attribute is essential. It is calculated by

$$IG = E(parent) - AverageE(childs) \tag{2}$$

E is entropy which is defined as measure of impurity and it is calculated by

$$Entropy = \sum_{i} -P_i \log_2 P_i \tag{3}$$

where P_i is the probability of class i.

2) Support Vector Machine: The SVM algorithm was proposed by Vladimir Vapnik and Alexey Chervonenkis in 1963 [16]. SVM mainly use for classification problems, and it is a supervised ML algorithm. SVM consists of support vectors, hyperplane, and marginal distance. Support vectors are data samples which decide the hyperplane if we delete the support vectors then the position of hyper-plane change. The hyperplane is the decision boundary for data samples which helps in classification. Each data sample is separating from the hyperplane with marginal distance. The marginal line is defining by the distance between the hyperplane as the closest support vector point. The hyperplane is defined using

$$w * x + b = 0 \tag{4}$$

The hyperplane uses data to train the model, which learned and uses optimization procedure for maximizing margin. From the hyperplane, we can make a prediction of whether OST is present or not. The hypothesis of the classifier is given by

$$H(x) = \begin{cases} +1, & \text{if } w * x + b \geq 0 \\ -1, & \text{if } w * x + b < 0 \end{cases} \tag{5}$$

3) K-Nearest Neighbors: KNN is a supervised algorithm used for classification and regression, coin by Evelyn Fix and Joseph Hodges [17]. The KNN model trained using labeled and structure data where each class is divide by a margin. If we want to predict the new data sample class, then we calculate the closest distance between sample to class. We use the Minkowski distance metric with the standard Euclidean metric. Euclidean distance calculate using

$$d(x, y) = d(y, x) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

4) Adaboost: Adaboost means Adaptive Boosting [19] [18], which is an ensemble learning technique. It uses the bagging method, which has two main techniques like Bootstrapping and Aggregation. In bootstrapping, random row samples and feature samples are given to multiple decision trees. Adaboost uses a weak learner who is a decision tree with single splits called Decision stumps. It uses weak learners and combines them to form strong learners called aggregation. Aggregating multiple decision trees reduces the chances of overfitting and improves the model performance. During boosting, the wrongly classified class gets higher weight while correctly classified gets low weight. We calculate the Total Error Rate using

$$Total\, ErrorRate(TE) = \frac{p}{N} \tag{7}$$

where, p= predicted wrong value and N= total number of samples. After that, we find out the performance of stumps from

$$Performance\, of\, stumps(P) = \frac{1}{2} \log_e(\frac{1 - TE}{TE}) \tag{8}$$

From the performance of stumps, we can update the weight of samples. Incorrectly classified sample weight updated by

$$W_n = W_o * e^P \tag{9}$$

and correctly classified sample weight updated by

$$W_n = W_o * e^{-P} \tag{10}$$

This process is repeated multiple times and stopped when the error rate is minimum.

E. Performance Evaluation Matrix

We have used performance metrics to evaluate the ML models like Accuracy, Sensitivity, Specificity, ROC curve (receiver operating characteristic curve), F1 score, and Kappa coefficient.

Accuracy is the ability of the model to predicts True positive and True negative values.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

However, due to an imbalanced dataset, we do not depend on accuracy, so we take another performance metric. The ability to detect True Positive values from the given set of samples is known as Sensitivity.

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

Specificity detects True Negative values and calculates them by the given formula.

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

F1 is the harmonic mean of precision and recall where the $\beta = 1$ and calculate using

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{14}$$

Kohen Kappa is the metrics of the observed value and expected values which check the reliability among actual and predicted values [20]. It is calculated by

$$Kappa = \frac{ACC_O - ACC_E}{1 - ACC_E} \tag{15}$$

where,
ACC_O = Observe accuracy

Agreement level	Poor	Fair	Moderate	Good	Almost perfect
Min value	$\leq$	0.20	0.40	0.60	0.80
Max value	0.20	0.40	0.60	0.80	1.00

TABLE III: Level of agreement for kappa coefficient

ACC_E = Expected accuracy

ROC curve (receiver operating characteristic curve) [21] is another performance evaluation metric that represents TPR (True Positive Rate) and FPR (False Positive Rate) where

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

The AUC(Area Under Curve) values range between 0 and 1.

We are using above evaluation metrics to asses the performance of the ML models.

III. Result and Discussion

A. Proposed ML model results

We use 80% samples to train models and 20% is using for testing purposes. TABLE IV and TABLE V show the overall training and testing results of four ML models. We perform hyperparameter tuning for each model to improve the accuracy.

At the training stage, the overall training accuracy of the DT model is 96.00%, sensitivity 95.45% and specificity 97.79%. The F1 and Kappa coefficient is 95.90%, 0.94 respectively. The SVM model under-perform with 81.83% accuracy, with a sensitivity of 80%. The KNN and Adaboost train very well with 100% of accuracy, sensitivity and specificity. Even F1 score and Kappa coefficient of KNN and Adaboost in 100%. All four models train very well, but KNN and Adaboost trained 100% for given data.

Model	ACC	SEN	SPC	F1	Kappa
DT	96.00%	95.45%	97.79%	95.90%	0.94
SVM	81.83%	80.01%	90.30%	80.49%	0.72
KNN	100%	100%	100%	100%	1.0
AdaBoost	**100%**	**100%**	**100%**	**100%**	**1.0**

TABLE IV: Training performance of Ml models

For the testing set, DT gives an accuracy of 81.22% which is acceptable but is slightly low compared to KNN and SVM. DT sensitivity and specificity are 79.70% and 89.50%, which shows that the model is detecting True Negative more than True Positive due to an imbalanced dataset.The accuracy of SVM is 83.80% with sensitivity=81.40%, specificity=91.40% and F1 score is 81.75%. The kappa coefficient is 0.74, which shows the perfect agreement.

The KNN model gives overall accuracy of 86.90% and sensitivity of 86.90%. We calculate the error rate separately and choose the suitable value of ’k’ for our KNN model i.e. k=1. To improve the KNN model accuracy we have done hyperparameter optimization. The specificity of KNN is 93.10%, F1= 86.22%, and the Kappa coefficient is 0.79.

Model	ACC	SEN	SPC	F1	Kappa
DT	81.22%	79.70%	89.50%	80.68%	0.698
SVM	83.80%	81.40%	91.40%	81.75%	0.741
KNN	86.90%	86.90%	93.10%	86.22%	0.793
AdaBoost	**91.70%**	**91.60%**	**96.00%**	**90.60%**	**0.870**

TABLE V: Testing performance of Ml models

Adaboost ML model outperforms all ML models with an overall accuracy of 91.70% and sensitivity of 91.60%. We perform hyperparameter tuning to improve our model’s accuracy like we set the depth to 9 and change the learning rate to 1. The specificity of AdaBoost is 96.00%, F1= 90.60%, and Kappa coefficient is 0.87. We can say that the Adaboost model performs very well, with an almost perfect agreement between observed and expected values from the kappa coefficient.

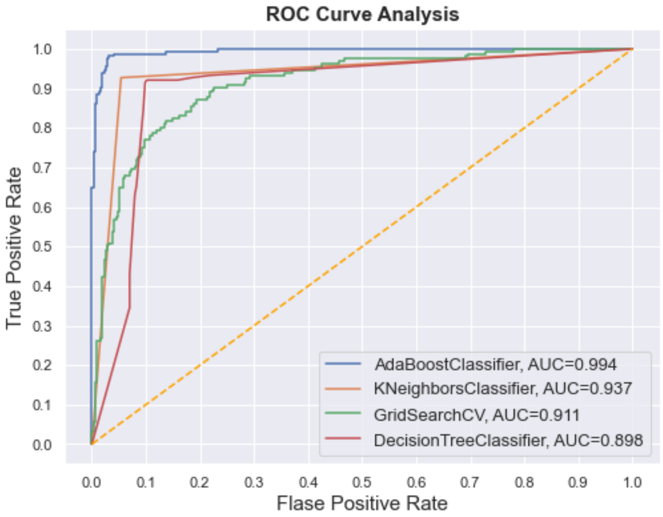

Fig. 3: Receiver Operating Characteristic (ROC) of ML models with Area Under Curve(AUC)

ROC curve gives the overall performance of the ML model at specific threshold values. DT, KNN and SVM models perform well with an AUC value of 0.90, 0.94 and 0.91. Adaboost model gets AUC score of 0.99 which shows model predict actual positive and negative value efficiently.

B. Comparative analysis with existing studies

We train our ML models with the same set of features to use by Arunachalam et al. They use three ML models like CT(Complex Tree), SVM(Support Vector Machine) and EL(Ensemble learner) [8]. Both types of features, i.e. Expert-guided and Cellprofilier, are separately and combine to check ML models’ performance. For both types of features, CT, SVM and EL overall accuracy are 81%, 90% and 87%. SVM gets higher accuracy as compared to CT and SVM.

We use both types of feature to train our models where our Adaboost model outperform SVM model of Arunachalam et

Models	Database	ACC	SEN	SPC
Complex tree [8]	[10]	81%	-	-
SVM [8]	[10]	90%	-	-
Ensemble Learners [8]	[10]	87%	-	-
Our Adaboost Model	[10]	**91.70%**	**91.60%**	**96.00%**

TABLE VI: Comparative analysis with exiting studies

al. with an overall accuracy of 91.70% as shown in TABLE V. Adaboost efficiently and accurately classify three types of Osteosarcoma cancer.

IV. Conclusion

In this paper, we compare four ML algorithms that are DT, SVM, KNN and Adaboost. We perform training and testing and get an outstanding result where Adaboost performance is higher than the other three. We have done hyperparameter tuning based on the trial and error method. After tuning the Adaboost model, we got an overall accuracy of 91.70% and the other three models, SVM, KNN and DT get overall accuracy of 83.80%, 86.90% and 81.22% respectively. For our model's performance evaluation, we use different evaluation metrics like Accuracy, Sensitivity, Specificity, F1, and Kappa coefficient. We even compare our Adaboost model with previous studies where Adaboost gets the highest result compared to SVM and CT [8]. The paper is limited to only three types of osteosarcoma tumors, i.e., Viable, Non-Viable, and Non-Tumor. The accuracy of classification can be improved by extracting relevant features from RGB histopathological images.

Thus we have concluded that the Osteosarcoma types can be successfully classified using the Adaboost machine learning technique with the desired range of accuracy 91.70%, sensitivity 91.60%, specificity 96.00%, F1 90.60% and Kappa coefficient 0.87.

Acknowledgment

The author would like to thanks the College of Engineering Pune for support and providing essential facilities. The authors also thank the UT Dallas team for their histopathological image dataset.

References

[1] Ritter, J. and Bielack, S.S., 2010. Osteosarcoma. Annals of oncology, 21, pp.vii320-vii325.

[2] Jerome, T.J., Varghese, M., Sankaran, B., Thomas, S. and Thirumagal, S.K., 2009. Tibial Chondroblastic Osteosarcoma—Case Report. Foot and ankle surgery, 15(1), pp.33-39.

[3] A. Bentaieb and G. Hamarneh, ”Adversarial Stain Transfer for Histopathology Image Analysis,” in IEEE Transactions on Medical Imaging, vol. 37, no. 3, pp. 792-802, March 2018, doi: 10.1109/TMI.2017.2781228.

[4] Hamamoto, R., 2021. Application of Artificial Intelligence for Medical Research.

[5] Huang, W.B., Wen, D., Yan, Y., Yuan, M. and Wang, K., 2016, July. Multi-target osteosarcoma MRI recognition with texture context features based on CRF. In 2016 international joint conference on neural networks (IJCNN) (pp. 3978-3983). IEEE.

[6] Arunachalam, H.B., Mishra, R., Armaselu, B., Daescu, O., Martinez, M., Leavey, P., Rakheja, D., Cederberg, K., Sengupta, A. and NI'SUILLEABHAIN, M.O.L.L.Y., 2017. Computer aided image segmentation and classification for viable and non-viable tumor identification in osteosarcoma. In Pacific Symposium on Biocomputing 2017 (pp. 195-206).

[7] Mishra, R., Daescu, O., Leavey, P., Rakheja, D. and Sengupta, A., 2018. Convolutional neural network for histopathological analysis of osteosarcoma. Journal of Computational Biology, 25(3), pp.313-325.

[8] Arunachalam, H.B., Mishra, R., Daescu, O., Cederberg, K., Rakheja, D., Sengupta, A., Leonard, D., Hallac, R. and Leavey, P., 2019. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. PloS one, 14(4), p.e0210706.

[9] Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A. and Hassanpour, S., 2017. Deep learning for classification of colorectal polyps on whole-slide images. Journal of pathology informatics, 8.

[10] Leavey, P., Sengupta, A., Rakheja, D., Daescu, O., Arunachalam, H. B., & Mishra, R. (2019). Osteosarcoma data from UT Southwestern/UT Dallas for Viable and Necrotic Tumor Assessment [Data set].

[11] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, December, 2013, pp 1045-1057.

[12] Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J. and Golland, P., 2006. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology, 7(10), pp.1-11.

[13] Zhu, Y., Ting, K.M. and Angelova, M., 2018, June. A distance scaling method to improve density-based clustering. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 389-400). Springer, Cham.

[14] Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), pp.20-28.

[15] Safavian, S.R. and Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), pp.660-674.

[16] Boser, B.E., Guyon, I.M. and Vapnik, V.N., 1992, July. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).

[17] Fix, E., 1985. Discriminatory analysis: nonparametric discrimination, consistency properties (Vol. 1). USAF school of Aviation Medicine.

[18] Freund, Y. and Schapire, R.E., 1996, July. Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

[19] Rätsch, G., Onoda, T. and Müller, K.R., 1998. An improvement of AdaBoost to avoid overfitting. In Proc. of the Int. Conf. on Neural Information Processing.

[20] Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. biometrics, pp.159-174.

[21] Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), pp.1145-1159.