International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I6P202
Volume 69 Issue 6, 8-13, June 2021

Comparative Analysis of Various Tree Classifier Algorithms for Disease Datasets

Sajithra. N [#1], Dr.D.Ramyachitra [*2]
[#]*Research Scholar, Department of Computer Science, Bharathiar University, Tamilnadu, India*
Assistant Professor,[1]Department of Computer Science, Bharathiar University, Tamilnadu, India
[1] sajithramidhun@gmail.com, [2] jaichitra1@yahoo.co.in

Abstract *— Tree-Based Classification technique is one of the commonly used techniques called White box classification. It targets foreseeing to the having a place of cases or articles in the classes of a particular variable from their estimations on at least one prescient factor. This research work analyzes the concert of five tree-based classification algorithms, namely Decision Stump, J48, Logistic Model Tress (LMT), Random Forest, and REPTree. Various disease datasets such as breast cancer, Pima diabetes, and hypothyroid are utilized for calculating the performance of the classification algorithms by applying the 10-fold cross-validation parameter based on the given class label. Finally, the comparative analysis is held out, using the classification accuracy, kappa value, performance factors, and the error rate measures on all of the algorithms. From the experimental outcomes, it is derived that the LMT provides better results for all the disease datasets than the existing algorithms such as Decision Stump, J48, Random Forest, and REPTree.*

Keywords *— Decision Stump, J48, LMT, Random Forest, REPTree.*

I. INTRODUCTION

The reason for the arrangement trees is to anticipate or clarify reactions on a characterized subordinate variable. The techniques that can be obtained have a lot in common with the techniques used in the modern traditional methods of Cluster Analysis, Nonlinear Estimation, Discriminant Analysis, and Nonparametric Statistics Classification trees enthusiastically lend themselves to the graphically displayed creatures, making them easier to understand than they would be if only a strict numerical analysis were possible. Within the weka tool, there are several classification tree algorithms exist, such as Decision stump, Random Forest, Random Tree, REP Tree, LMT, MSPP, and J48.

In this research work, the comparison is made to find out the best tree classifier algorithm among the five algorithms such as namely Decision Stump, J48, Logistic Model Tress (LMT), Random Forest, and REPTree for the disease datasets. The structure of the study is as follows. Section 2 illustrates the literature review, Section 3 illustrates the methodology for the existing algorithms for the disease datasets and Section 4 illustrates the experimental results, and finally, Section 5 gives the conclusion and future work.

II. LITERATURE REVIEW

Decision Stump is the easiest case and the special case of a decision tree. For masquerading detection, a rules-based approach that compares n-grams of the command sequence using the decision stump algorithm (Jian et al., 2007). The Energy-Aware Decision Stump Linear Programming Boosting Node Classification based Data Aggregation (EADSLPBNCDA) Model is used to increase the data aggregation and energy consumption in Wireless Sensor Networks (WSN) (Kokilavani Sankaralingam et al., 2020). Logistic regression models adapt the idea of classification problems that uses logistic regression instead of linear regression. A ++stagewise adjustment process is used to make logistic regression models which can select appropriate attributes within the data in a normal (Landwehr et al., 2003). The Logistics Model Tree (LMT) is an element determination strategy to choose the most suitable situations from the V3 amino corrosive arrangements. Measured by ten-fold- cross-validation on 273 sequences [4], their approach achieves an accuracy of 97.8%, a specificity of 97.7%, and a sensitivity of 97.9%.

The most extreme probability rule is utilized as an assessment strategy for the assessment of tree logit models. The strategy is fit for clear documentation for the tree structure and is accepted to be unique (Shoombuatong et al., 2012). The accuracy of the classification is compared across nine decision tree methods, and that they are divided into two primary teams (Snousy et al., 2011). For the primary cluster, single decision tree C4.5, CART, Decision Stump, Random Tree, and REPTree are compared. The general decision tree for the second cluster is Bagging (C4.5 and REPTree), AdaBoost (C4.5 and REPTree), ADTree, and Random Forests. Decision tree (DT) classification methods such as C5.0, CART, CHAID, and Logistical Regression (LR) techniques are used to implement the monetary distress prediction model (Chen, 2011). The productivity of the LADTree and REPTree classifier for predicting credit risk and compares their adequacy with various measures (Profile, 2015). The decision tree algorithm is built utilizing a fast splitting attribute selection (DTFS) for large datasets. The algorithm builds a decision tree without storing the entire training set in main memory and having only one parameter, however, being terribly stable relating to it (Franco-Arcega et al., 2011).

The J48 calculation gives a strategy considered the ascription that manages missing qualities. Upgraded J48 characterization calculation is utilized for the irregularity interruption location frameworks. This algorithm helps to detect the probable attacks, which could jeoparadise the network confidentially (Aljawarneh et al., 2019). With the use of binary datasets and multiple class datasets on 13 performance measurements(Panigrahi & Borah, 2018), the three popular group classifiers J48, namely J48,

J48Consolidated, and J48Graft, are experiments. The effectiveness of the Random Forest (RF) variable is examined by means of importance measures to determine the true predictor among a wide range of candidate predictors (Archer & Kimes, 2008). The Random Forest Classification algorithm is used to provide experimental information on the yield of the Variable Significance Index based on Random Forests (Auret & Aldrich, 2011). The measurement of variable significance is associated with the conditional inference of forests, random forests, and stimulated trees and uses several simulated datasets to compare these methods (Genuer et al., 2010).

III. METHODOLOGY

In this research work, the comparison is made to find out the best tree classification algorithm among the five algorithms, namely Decision Stump, J48, Logistic Model Tress (LMT), Random Forest, and REPTree for the disease datasets such as breast cancer, Pima diabetes and hypothyroid. Figure 1 shows the flow diagram for the comparative analysis.

A. Dataset Description

The disease datasets such as breast cancer, Pima diabetes, and hypothyroid are collected from the UCI repository. The breast cancer dataset contains 286 instances and 10 attributes, the Pima diabetes dataset contains 768 instances and 9 attributes, and the hypothyroid dataset contains 3772 instances and 30 attributes. The data mining tool weka is utilized for examining the performance of the classification algorithms.

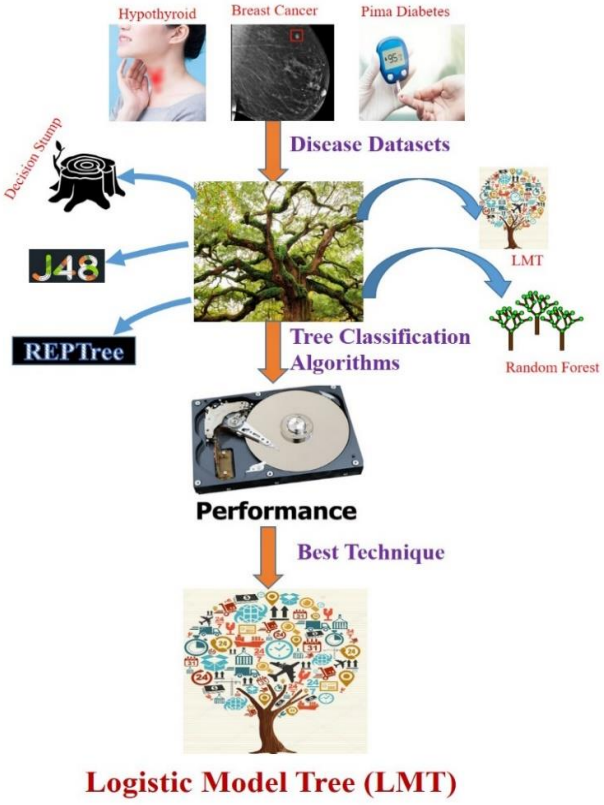

Figure 1: Flow diagram for comparative analysis of tree classifier algorithms

B. Classification

In data mining, the characterization strategy can be utilized to predict group participation for information instances. This is a case of supervised learning, where a set of properly identified observational training is available. The process by which ideas and objects are differentiated and understood may involve categorization or classification. This research work aims to find out the best tree classification algorithm among the five algorithms such as Decision Stump, J48, Logistic Model Tress (LMT), Random Forest, and REPTree for the disease datasets such as breast cancer, Pima diabetes, and hypothyroid.

C. Trees

The tree is one of the classification techniques, and this research article used five tree classification algorithms for the analysis of disease datasets and are as follows.

1. Decision Stump (DS)
2. J48
3. Logistic Model Tree (LMT)
4. Random Forest (RF)
5. REPTree

a) Decision Stump

The simplest of a decision tree is a decision stump algorithm, and the decision is made by checking the class as positive or negative. The decision stump algorithm comes under the machine learning model, and it consists of a one-level decision tree. Specifically, a decision tree with an internal node that is the root node is instantly connected to the terminal nodes known as leaves. A decision stump algorithm constructs a prediction based on the meaning of only one input function. Occasionally they are also referred to as 1-rules (Holte Holte, 1993).

b) J48

J48 algorithm is also termed as C4.5, and it is a broadly used machine learning algorithm that comes under the category of decision tree algorithms. J48 algorithm is a type of ID3 algorithm that differs from building a decision tree, and it accepts categorical and continuous attributes.

c) Logistic Model Tree

The Logistic Model Tree (LMT) algorithm is a decision-based algorithm that adopts logistic regression into the decision tree induction or combines these two methods in a single tree structure. The LMT algorithm creates a tree structure based on binary and multiclass target variables, numerical values, and missing values. The Logistic Model Tree generates a unique result in the form of a tree containing binary splits on numerical attributes (Landwehr et al., 2005).

d) Random Forest

The Random forests algorithm comes under the ensemble learning method, and it can be used for classification and regression. The randomization is presented in two ways. (i) random sampling of data for bootstrap samples (ii) random selection of input attributes for generating individual base decision trees. The Random Forest algorithmic rule was

initially developed by Leo Breiman, a statistician at the University of California at Berkeley (Genuer et al., 2010).

e) REPTree

The Reduced Error Pruning (REP) Tree algorithm is a fast decision tree learning algorithm based on C4.5algorithm that produces classification or regression trees. When preparing the model, the REPTree algorithm sorts the values of the numerical attributes once. The decision tree or regression tree is constructed using information gain or variance and pruned using a reduced-error size (Snousy et al., 2011). The REPTree applies the logic of the regression tree and to generates multiple trees in modified iterations.

IV. RESULTS AND DISCUSSION

This research work computes the experimental measures by utilizing the performance factors, classification accuracy, and error rate measures. Accuracy measurement, performance factors, and error rate are used to determine the best algorithm for the disease datasets. The accuracy measure and performance factors have compared with various tree classifiers for the breast cancer dataset are illustrated in Table 1.

Table 1 Comparison of accuracy and performance factors for breast cancer dataset

Algorithms	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	F Measure	ROC Curve
Decision Stump	69	31	0.685	0.466	0.677	0.681	0.588
J48	75	25	0.755	0.524	0.752	0.713	0.584
LMT	75	25	0.752	0.492	0.737	0.722	0.675
Random Forest	70	30	0.696	0.543	0.664	0.669	0.634
REPTree	71	29	0.706	0.572	0.669	0.664	0.621

Table 2 Comparison of accuracy and performance factors for Pima diabetes dataset

Algorithms	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	F Measure	ROC Curve
Decision Stump	72	28	0.719	0.348	0.716	0.717	0.684
J48	74	26	0.738	0.327	0.735	0.736	0.75
LMT	**77**	**23**	**0.775**	**0.325**	**0.770**	**0.766**	**0.831**
Random Forest	76	24	0.758	0.310	0.754	0.755	0.820
REPTree	75	25	0.753	0.328	0.747	0.748	0.766

The accuracy measure and performance factors have compared with various tree classifiers for the Pima diabetes dataset are depicted in Table 2. The accuracy measure and performance factors have compared with various tree classifiers for the hypothyroid dataset are depicted in Table 3. Fig.2 shows the comparison of accuracy for the breast cancer dataset with various tree classifier algorithms. From Fig.2, it is inferred that the J48 and LMT algorithm has the highest accuracy of the breast cancer dataset.

Table 3 Comparison of accuracy and performance factors for the hypothyroid dataset

Algorithms	Correctly Classified	Incorrectly Classified	TP Rate	FP Rate	Precision	F Measure	ROC Curve
Decision Stump	95	5	0.954	0.009	0.765	0.841	0.981
J48	99	1	0.996	0.019	0.975	0.974	0.993
LMT	**99.5**	**0.5**	**0.995**	**0.022**	**0.995**	**0.995**	**0.994**
Random Forest	99	1	0.993	0.038	0.955	0.924	0.999
REPTree	99	1	0.996	0.007	0.946	0.936	0.993

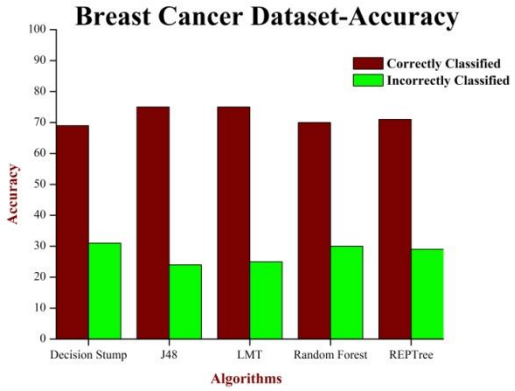

Figure 2 Comparison of Accuracy for the breast cancer dataset with various tree classifier algorithms

Fig. 3 shows the comparison of performance measures for the breast cancer dataset with various tree classifier algorithms. Fig.3, it is implied that the LMT and J48 algorithms have the highest performance metric values than the other existing algorithms. For the breast cancer dataset, the LMT algorithms perform 8% better than the Decision Stump algorithm, 6.66% better than the Random Forest algorithm, and 5.33% better than the REPTree algorithm.

Fig.4 shows the comparison of accuracy for Pima diabetes with various tree classifier algorithms. Fig.4, it is implied that the LMT has the highest accuracy of the Pima diabetes dataset. Fig. 5 shows the comparison of performance measures for the Pima diabetes dataset with various tree classifier algorithms. Fig.5, it is implied that the LMT has the highest performance metric values than the other existing algorithms. For the Pima diabetes dataset, the

LMT algorithms perform 6.49% better than the Decision Stump algorithm, 3.89% better than the J48 algorithm, 1.29% better than the Random Forest algorithm, and 2.59% better than the REPTree algorithm.

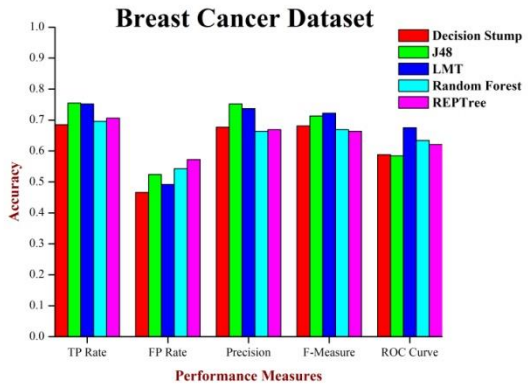

Figure 3 Comparison of Performance Measures for breast cancer dataset with various tree classifier algorithms

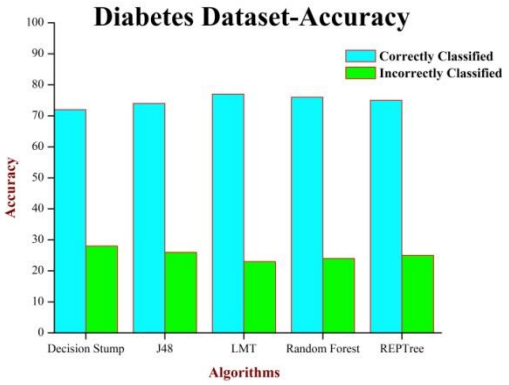

Figure 4 Comparison of Accuracy for the Pima diabetes dataset with various tree classifier algorithms

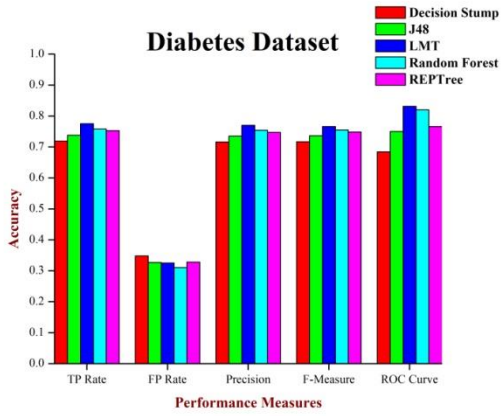

Figure 5 Comparison of Performance Measures for the Pima diabetes dataset with various tree classifier algorithms

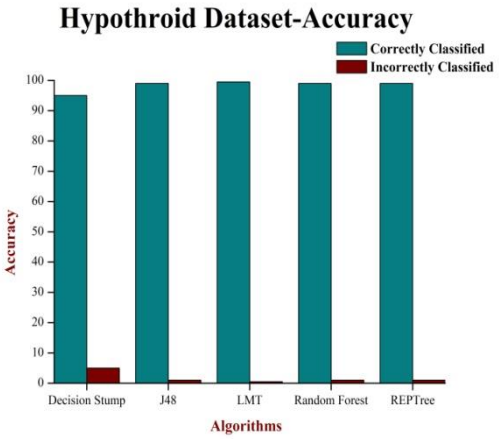

Figure 6 Comparison of Accuracy for the hypothyroid dataset with various tree classifier algorithms

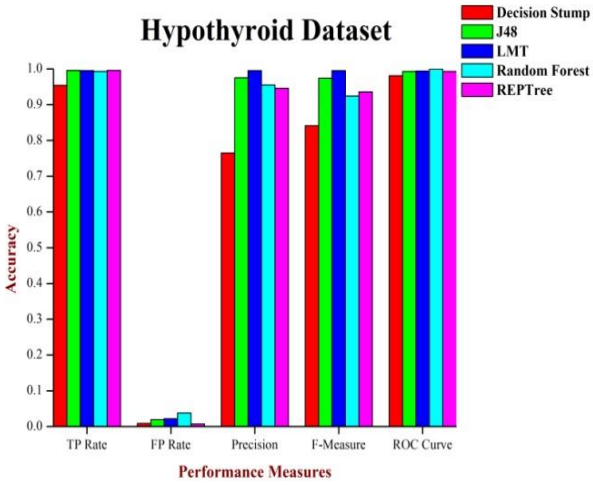

Figure 7 Comparison of Performance Measures for the hypothyroid dataset with various tree classifier algorithms

Fig.6 shows the comparison of accuracy for the hypothyroid dataset with various tree classifier algorithms. Fig.6, it is implied that the LMT has the highest accuracy of the hypothyroid dataset. Fig.7 shows the comparison of performance measures for the hypothyroid dataset with various tree classifier algorithms. From Fig.7, it's implied that the LMT algorithm has the highest performance metric values than the other existing algorithms. For the hypothyroid dataset, the LMT algorithms perform 4.52% better than the Decision Stump algorithm, 0.5% better than the J48 algorithm, 0.5% better than the Random Forest algorithm, and 0.5 % better than the REPTree algorithm. The kappa and error rate measures such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have compared with various tree classifiers for the disease datasets are shown in Table 4.

Table 4 Comparison of kappa and error rate measures for various disease datasets

Algorithms	Breast Cancer			Pima Diabetes			Hypothyroid		
	Kappa Value	MAE	RMSE	Kappa Value	MAE	RMSE	Kappa Value	MAE	RMSE
Decision Stump	0.2257	0.3801	0.4404	0.3745	0.3802	0.4418	0.7147	0.03	0.1225
J48	0.2826	0.3676	0.4324	0.4164	0.3158	0.4463	0.9707	0.003	0.0414
LMT	**0.3042**	**0.3589**	**0.4291**	**0.4756**	**0.3175**	**0.3963**	**0.9654**	**0.0025**	**0.0488**
Random Forest	0.1736	0.3727	0.4613	0.4566	0.3106	0.4031	0.9523	0.015	0.0642
REPTree	0.1601	0.3797	0.4652	0.438	0.3272	0.4289	0.971	0.0041	0.0445

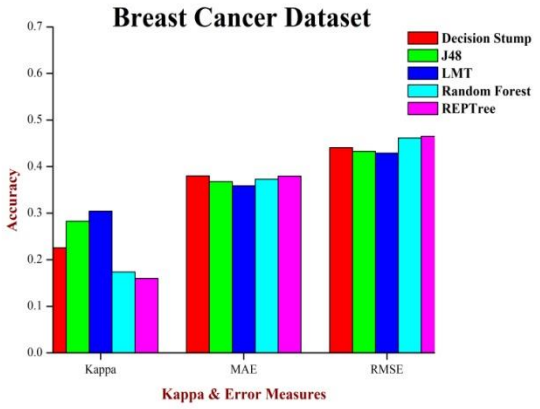

Figure 8 Comparison of Kappa & Error Rate measures for breast cancer dataset with various tree classifier algorithms

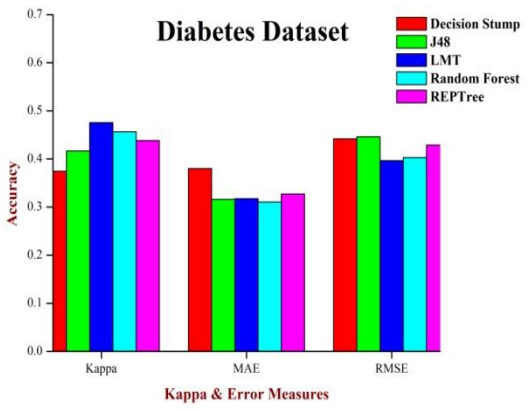

Figure 9 Comparison of Kappa & Error Rate measures for the Pima diabetes dataset with various tree classifier algorithms

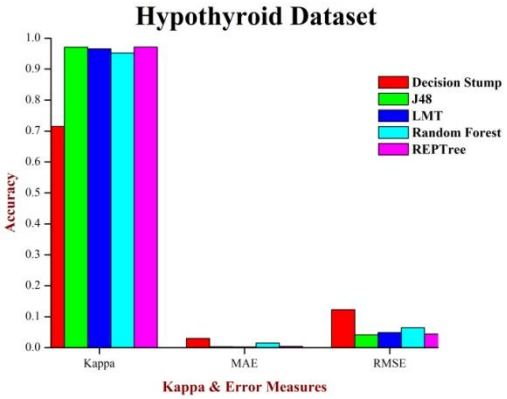

Figure 10 Comparison of Kappa & Error Rate measures for the hypothyroid dataset with various tree classifier algorithms

From the experimental results, it's implied that the LMT algorithmic rule achieves well as compared to the other existing algorithms like Decision Stump, J48, Random Forest, and REPTree. The LMT algorithm gives more correctly classified instances comparable to other existing techniques for all the disease datasets such as breast cancer, Pima diabetes, and hypothyroid. Also, the error rate for the LMT is less compared to other existing methods, as shown in Fig. 8, 9, and 10.

V. CONCLUSION

This research work analyzed the performance of five tree classifier algorithms, namely Decision Stump, J48, Logistic Model Tree, Random Forest, and REPTree. Various disease datasets, namely breast cancer, Pima diabetes, and hypothyroid, are utilized for calculative the performance of the tree classifier algorithms by using the 10-fold cross-validation parameter. Lastly, this research work analyzed the algorithms by using the performance factors, classification accuracy, and error rate measures. From the outcomes, it is recognized that the LMT performs better than other existing algorithms for all the disease datasets. In the future, the classification trees can experiment on other datasets also. Also, the LMT algorithm can modify in the future to get more effective outcomes. The tree classification algorithms can be analyzed using other specifications like the training set, percentage split, and supplied test set.

ACKNOWLEDGMENT

Non-funded
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] ZhouJian et al., Masquerade detection by boosting decision stumps using UNIX commands, Computers & Security, 26(4) (2007) 311-318.

[2] S.Kokilavani Sankaralingam, N.Sathishkumar Nagarajan, A.S.Narmadha, Energy-aware decision stump linear programming boosting node classification based data aggregation in WSN, Computer Communications, 155(1) (2020) 133-142.

[3] Niels Landwehr, Mark Hall, Eibe Frank, Logistic Model Trees, Machine Learning: ECML 2003, 2837(2003) (2003) 241-252.

[4] Watshara Shoombuatong, Sayamon Hongjaisee, Francis Barin, Jeerayut Chaijaruwanich, HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees, Computers in Biology and Medicine, Volume 42(9),http://dx.doi.org/10.1016/j.compbiomed.2012.06.011.

[5] Andrew Daly, Estimating "tree" logit models, Transportation Research Part B: Methodological, 21(4), (1987) 251–267,http://dx.doi.org/ 10.1016/ 0191-2615(87) 90026 - 9.

[6] Mohmad Badr Al Snousy, Hesham Mohamed El-Deeb, Khaled Badran, Ibrahim Ali Al Khlil, Suite of decision tree-based classification algorithms on cancer gene expression data, Egyptian Informatics Journal 12, (2011) 73–82.

[7] Mu-Yen Chen, Predicting corporate financial distress based on the integration of decision tree classification and logistic regression, Expert Systems with Applications, 38(9) (2011) 11261-11272.

[8] Lakshmi Devasena C, Proficiency_Comparison of LADTree and REPTree Classifiers for Credit Risk Forecast, International Journal on Computational Sciences & Applications (IJCSA) 5(1) (2015) 39-50.

[9] A. Franco-Arcega, J.A. Carrasco-Ochoa, G. Sánchez-Díaz, J.Fco. Martínez-Trinidad, "Decision tree induction using a fast splitting attribute selection for large datasets", Expert Systems with Applications, 38(11) (2011) 14290-14300.

[10] Aljawarneh, S., Yassein, M.B. & Aljundi, M. An enhanced J48 classification algorithm for the anomaly intrusion detection systems. Cluster Comput 22, (2019) 10549–10565.

[11] Panigrahi, Ranjit & Borah, Samarjeet. Rank Allocation to J48 Group of Decision Tree Classifiers using Binary and Multiclass Intrusion Detection Datasets. Procedia Computer Science. 132. (2018) 323-332. 10.1016/j.procs.2018.05.186.

[12] Kellie J. Archer, Ryan V. Kimes, Empirical characterization of random forest variable importance measures, Computational Statistics & Data Analysis, 52 (4) (2018),http://dx.doi.org/10.1016/j.csda.2007.08.015.

[13] Lidia Auret, Chris Aldrich, Empirical comparison of tree ensemble variable importance measures, Chemometrics and Intelligent Laboratory Systems 105 (2011) 157–170.

[14] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot, Variable selection using random forests, Pattern Recognition Letters 31 (2010) 2225–2236.

[15] Holte, Robert C.. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. CiteSeerX: 10.1.1.67.2711. (1993).

[16] N. Landwehr, M. Hall, and E. Frank, Logistic model trees, Machine Learning, 59(1-2), (2005) 161-205.

[17] L. Breiman, Random Forests, Machine Learning, 45 (1) (2001) 5-32.

[18] Snousy, Mohmad & El-Deeb, Mohamed & Badran, Khaled & Khlil, Ibrahim. The suite of decision tree-based classification algorithms on cancer gene expression data. Egyptian Informatics Journal. 12. 73-82. 10.1016/j.eij.2011.04.003. (2011).