# Effective and Accurate Bootstrap Aggregating (Bagging) Ensemble Algorithm Model for Prediction and Classification of Hypothyroid Disease

Awujoola Olalekan J.
Nigerian Defence Academy
Computer Science Department
Faculty of Military Science and Interdisciplinary Studies

Francisca Ogwueleka
Nigerian Defence Academy
Faculty of Military Science and Interdisciplinary Studies

P. O. Odion, PhD
Nigerian Defence Academy
Computer Science Department
Faculty of Military Science and Interdisciplinary Studies

## ABSTRACT

Accurate diagnose of diseases prior to their treatment is a challenging task for the modern research, therefore it becomes necessary and important to use modern computing techniques to design an efficient and accurate prediction systems. Thyroid is one of the most common diseases found in human body with many side effects the accuracy for thyroid diagnosis system may be greatly improved by considering an ensemble algorithm technique. In this paper, an effective and accurate thyroid disease prediction model is developed using an ensemble of Bagging with J45 and ensemble of Bagging with SimpleCart to extract useful information and diagnose diseases. The performances of the two ensemble model were compared with single classifiers. The Bagging ensemble algorithm for thyroid prediction system promises excellent overall accuracy of 99.66% while other single selected classifiers like Bagging and SimpleCART has accuracy of 99.55% and J48 with accuracy of 99.60%.

## General Terms

Machine Learning.

## Keywords

Receiver operating characteristic (ROC), Ensemble, classification, hypothyroid diseases, Bagging, SimpleCART, J48.

## 1. INTRODUCTION

The advancement of computational biology is used in the healthcare industry. It allowed collecting the stored patient data for medical disease prediction. There are different intelligent prediction algorithms are available for the diagnosis of the disease at early stages. The Medical restorative framework system is wealthy of information sets, but there are no brilliant framework that can effortlessly analyze the disease [2].

Over some period of time, machine learning algorithms has played crucial role in solving the complex and nonlinear problems in developing a prediction model. In any disease and infection prediction, models are required to fundamental the features that can be chosen from the distinctive datasets which can easily be used as a classification in the healthy patient as precisely as possible. Otherwise, misclassification may result in a healthy patient that endures unnecessary treatment.

Prevention in wellbeing care is a continuous concern for the healthcare providers and the correct disease examination at the right time for a patient is highly important, as a result of the implied risk. Lately, the normal and usual medical report can be followed by an extra report provided by decision support system or other advanced diagnosis techniques based on symptoms [6]. Machine learning is a modern way of computing where knowledge along with a technique is used to build a model which imitates the behaviour of human being. Once the machine learning classification model is trained it will start predicting the class of a given feature set.

Thyroid disease is one among the common lifestyle disease. Thyroid organ is a butterfly-molded organ which is present in the neck underneath the mouth of human body. It release hormones that control metabolism like heart rate, body temperature etc. It produces two main hormones T3 and T4. The Thyroid disease may be broadly categorized i.e. hypothyroid and hyperthyroid. When the amount of hormones exceed the amount required by the human body, it causes hyperthyroidism. Hypothyroidism is the inverse of hyperthyroidism; it reduces body metabolism, cause drowsiness and pain in joints. These hormones are responsible for various metabolic activities like body weight, heart rate etc. These activities may get disturbed if the level of these hormones changes. So the diagnosis of thyroid disease is important before its treatment [3].

The thyroid gland can also be referred to as an endocrine gland located in the neck. It is usually builds in the lower part of the human neck, mostly beneath the Adam's apple which causes the secretion of thyroid hormones and that basically influences the rate of metabolism and protein synthesis. To control the metabolism in the body, thyroid hormones are useful in many ways, counting how briskly the heart beats and how quickly the calories are burnt. The composition of thyroid hormones by the thyroid gland helps in the domination of the body's metabolism. The thyroid glands are composed of two active thyroid hormones, levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3).To regulate the temperature of the body these hormones are imperative in the fabrication and also in the comprehensive construction and supervision [2]. Their functions include stabilizing body temperature, blood pressure and regulating the heart rate. People suffering from thyroid gland tend to fall sick due to under or over production of hormones from this gland [8].

Hypothyroidism and hyperthyroidism are a result of an imbalance of thyroid hormone. Hypothyroidism is simply not enough thyroid hormone and hyperthyroidism is too much. Either imbalance affects the metabolism in the body. Hypothyroidism causes a reduction in stroke volume and heart rate causing lowered cardiac output with a decrease in heart sounds. Hypothyroidism is condition that underlies most chronic degenerative diseases and hormone irregularities and results in a weakened immune system [8].

Data mining is becoming strategically important tool for many organizations including healthcare sector having huge amount of data. Data mining, the extraction of hidden predictive and descriptive information from large databases, is a powerful new technology with great potential to help healthcare sector to focus on the most important information in their data warehouses. Data mining will be the cornerstone in detecting disease. Data mining is a technique which can be used to develop expert system for the classification of medical data [8].

The remaining part of this paper is organised as follows. The Section 2 of the paper presents a brief background of various related life style disease prediction systems and a brief study of thyroid disease prediction system. The Section 3 explains about the machine learning based framework and algorithm of the proposed machine learning model of thyroid prediction system. The training and prediction accuracy of the proposed thyroid system at various levels is computed in the Section 4. The Section 5 of the article provides brief findings and future scope of the presented research.

## 2. LITERATURE REVIEW

Various researchers have used different pattern classifiers for developing lifestyle disease prediction systems. Many authors have used various kinds of data mining technique. The authors proved to obtain an adequate approach and certainty to find out the diseases analogous to the thyroid by the work that includes various datasets and algorithms linked with the work that is to be done in the future perspective to accomplish effective and better results. The intent of their paper interprets various techniques of data mining mechanisms and the statistical attributes that is been popularized in the latter years for interpretation of thyroid diseases with the certainty by various authors to attain various prospects and for various approaches. There are various algorithms of machine learning counting random forest, decision tree, naïve Bayes, SVM and ANN that are extensively used in the frequent diseases and in the prognostic problems. In this section a brief study of thyroid disease prediction system have been presented

### 2.1 Review of Related Work Done Using Machine Learning Approach

[5] presents thyroid data analysis, by performing classification and prediction using Zero R on dataset after applying Info Gain attribute Eval Method and obtained accuracy of 92.2853 %

[7] compared the performance of three selected classification algorithms J48, Random forest and Naïve Bayes in prediction of hypothyroid diseases. He obtained accuracy of 99% from J48 classifier using 0.02s to build the model while Random forest yielded accuracy of 99.3% but in 1.17s in building the model. Therefore J48 classifier was considered best in predicting the hypothyroid disease.

[10] in their bid to predict thyroid diseases divided their experiment into three parts: pathological observations, serological tests and combination of both. The first model, achieved the highest accuracy of 98.56% with bagging while in the second model they achieved 99.08% with SVM. Then the highest accuracy of 92.07% was obtained by J48 classifier on the serological tests.

[9] compared Naive Bayes Classifier, Support Vector Machine (SVM), AdaBoost tree, Artificial Neural Networks (ANN), to find a powerful model for breast cancer prediction. They implemented PCA for dimensionality reduction

[1] used SVM classification technique on two different benchmark datasets for breast cancer which got 98.80% and 96.63% accuracies.

[2] proposed different machine earning techniques for diagnosis and the prevention of thyroid. Machine Learning Algorithms such as , support vector machine (SVM), K-NN, Decision Trees were used to predict the estimated risk on a patient’s chance of obtaining thyroid disease. However, SVM yielded the highest accuracy result of 99.63%.

[4] used Linear Discriminant Analysis data mining technique for the prediction of thyroid disease (LDA) Algorithm and obtained accuracy of 99.62% with cross validation k=6.

[8] applied various data mining classification algorithms like Mutlilayer perceptron , RBF Network, Bayes net , C4.5 , CART , Decision stump , REP tree techniques to develop classifier for diagnosis and classification of hypothyroid disease with various k-fold cross validation for C4.5 classifier. He obtained accuracy for different k- fold, however k =6 yielded 99.60 % against 99.575% with k=10 as the highest accuracy obtained.

### 2.2 Algorithm Description

#### *2.2.1 Ensemble*

The thought of ensemble classification algorithm is to discourage use of one single classifier but combining set of classifiers called an ensemble of classifiers, then combine their predictions or forecast for the classification of unseen data.

One of the basic major tasks of machine learning classification algorithms is to build a reasonable model from a dataset. The method of building a model from the dataset is referred to as the training or learning. The trained model are sometimes referred to as hypothesis or mostly called learner. Therefore, the learning algorithms that build the set of classifiers and after that classify the new data are known as Ensemble method [11].

Ensemble contains quite numbers of learners that are usually produced from the training set with the assistance of base learner classifier. Though many of the ensemble methods uses single base learning algorithm to generate what is referred to as an homogenous ensemble or base learner. There are also another methods where multiple learning algorithms are used and in this way create heterogeneous ensembles. Generally, ensemble technique are well known for their capacity to boost weaker learners [11].

From many research works, ensembles are much more often accurate and precise than the individual or single algorithm. The ensemble techniques that is also known as committee-based learning systems, train multiple hypotheses to solve the same problem. One of the most foremost common cases of ensemble modelling is the random forest trees where a number of decision trees are utilized to forecast the outcomes.

#### *2.2.2 J48 Decision Tree*

J48 is one of the most popular classification algorithm that is simple and easy to use and implemented. However, It is also called a Decision Tree with reduced error. It is based on Hunt’s algorithm and requires no domain knowledge or parameter setting. It can handle high dimensional data [13]. It handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. J48 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification [14].

### *2.2.3 Bagging*

Bagging is an ensemble method used to classify the data with good accuracy. It is also called as Bootstrap Aggregation [14]. Here first the decision trees are derived by building the base classifiers c1, c2,…, cn on the bootstrap samples D1, D2, .., Dn with replacement from the data set D. Later the final model or decision tree is derived as a combination of all base classifiers c1, c2,…, cn with the majority votes. It can be applied on any classifier such as REP Tree, random forest, C4.5 and J48 etc. Bagging plays an important role in the field of medical diagnosis [13].

## 3. RESEARCH METHOD

This work uses three data mining classification algorithms techniques which are Bagging, J48 and SimpleCart. However, we decided to build an ensemble of Bagging with J48 and Bagging with SimpleCart. Then compare and evaluate the performance accuracy results of the two ensembles.

All the classification algorithms were selected because of their properties, very often used for research purposes and have potential to yield good results. Moreover, they use different approaches for generating the classification models, which increases the chances for finding a prediction model with high classification accuracy. The dataset is publicly available from the University of California Irvine (UCI) Machine Learning Repository [12]. This data set was chosen because of the prevalence of nominal features and their predominance in the literature.

In this work k=6 was used because of its high prediction accuracy as obtained in [8] & [4].

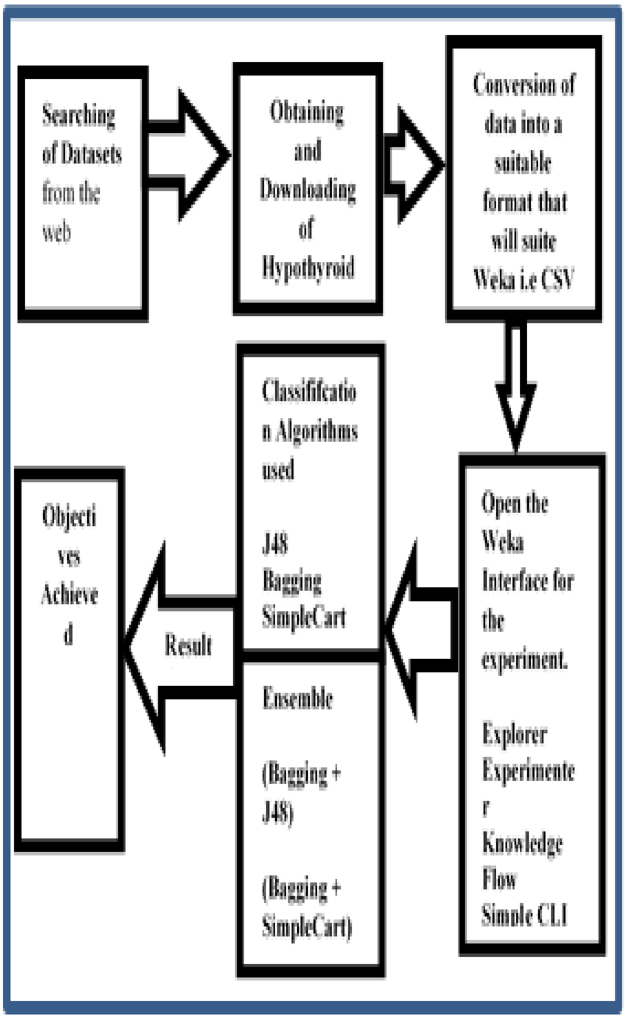
### 3.1 Experimental Setup

Weka popular, open-source data mining tool version 3.8.3 was adopted to use for this research work analysis. It is a collection of data mining algorithms designed in Java for solving real time data mining applications which can be used to perform a wide variety of tasks like regression, clustering, association, classification and visualization.

The analysis will be performed on a HP Windows 10 system with Intel® Core ™ i7 CPU, 2.30 ghz Processor and 8.00 GB RAM.

Weka (Waikato Environment For Knowledge Analysis) WEKA is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. Weka Waikato Environment for Knowledge Analysis (Weka) is a data mining tool available free of cost under the GNU General Public License. The version used in this study is 3.8.3 that has many state of the art machine learning tools and algorithms for data analysis and predictive modeling. This tool accepts the data file either in comma separated value (csv) or attribute-relation file format (arff) file format.

Figure1 shows the flow of methodology used in this research study, we first search for the hypothyroid dataset on the web and we have found suitable one on the UCI (University of California) repository which is famous for maintaining datasets for all the research related work. We then converted the data into suitable formats that can be used in our data mining tool i.e. CSV (comma separated values) and then convert all the data into nominal during the experiment. Then save in ARFF (attribute relation file format). Then we introduced our dataset in a suitable file format to our chosen data mining tool WEKA, it contains four interfaces namely Explorer, Experimenter, Knowledge flow and Simple CLI.


Searching of Datasets from the web
Obtaining and Downloading of Hypothyroid
Conversion of data into a suitable format that will suite Weka i.e CSV
Open the Weka Interface for the experiment.
Explorer
Experimenter
Knowledge Flow
Simple CLI
Classififcation Algorithms used
J48
Bagging
SimpleCart
Ensemble
(Bagging + J48)
(Bagging + SimpleCart)
Result
Objectives Achieved

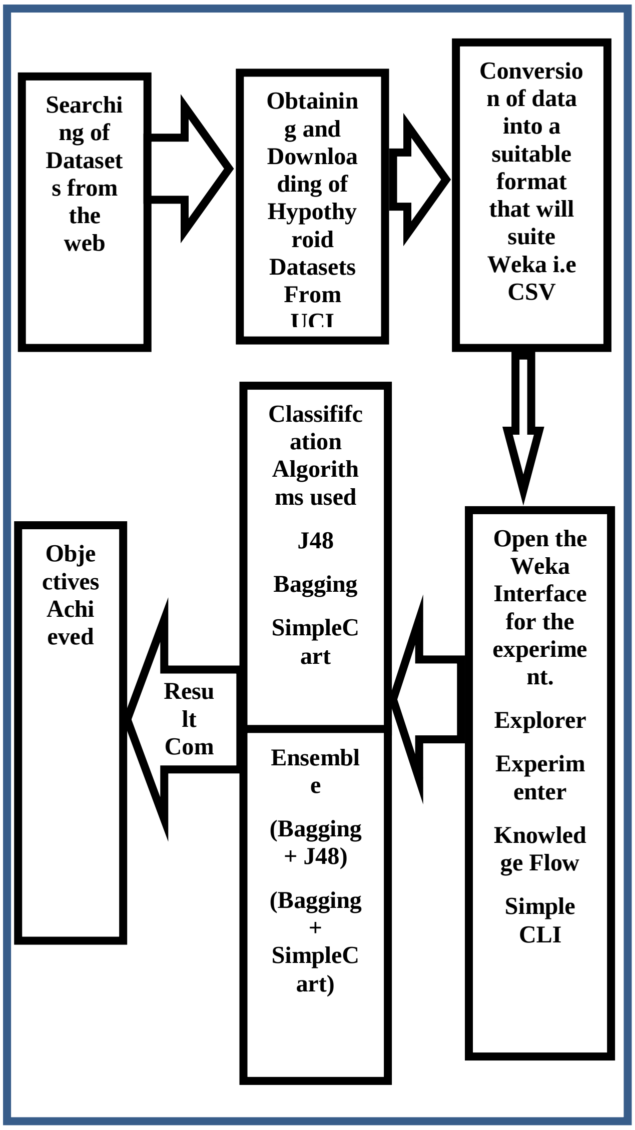

Searching of Datasets from the web
Obtaining and Downloading of Hypothyroid Datasets From UCI
Conversion of data into a suitable format that will suite Weka i.e CSV
Open the Weka Interface for the experiment.
Explorer
Experimenter
Knowledge Flow
Simple CLI
Classififcation Algorithms used
J48
Bagging
SimpleCart
Ensemble
(Bagging + J48)
(Bagging + SimpleCart)
Result Com
Objectives Achieved

**Figure 1: Flow of Methodology**

# 4. SURVEY RESULT ANALYSIS

## 4.1 Measures of Performance Evaluation

The actual and predicted classification done by a classification matrix is generated and represented by a confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

Once the confusion matrix is generated for each implemented algorithm the following metric values Accuracy, Sensitivity, Specificity and Error rate are calculated from the confusion matrix using the formulas listed below. The table 1 shows the confusion matrix for a two-class classifier [15].

**Table 1: Confusion Matrix for two class classifier.**

| ACTUAL | | PREDICTED | |
|---|---|---|---|
| | | Positive | Negative |
| | Positive | A (TP) | B (FN) |
| | Negative | C (FP) | D (TN) |

Where: A is the number of True Positives

B is the number of False Negatives

C is the number False Positive

D is the number of True Negatives

The experimental comparison of classification algorithms are done based on the performance measures of classification accuracy, specificity, sensitivity, error rate, Kappa statistics ROC and execution time.

1 Accuracy: It is the percentage of accurate predictions i.e the ratio of number of correctly classified instances to the total number of instances and it can be defined as:

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + FN + FP + TN}} \quad (1)$$

= (A + D) / (A+B+C+D)

Where TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

$$\frac{\textit{True Positive +True Negative}}{\textit{True Positive +False Negative +False Negative +True Negative}}$$

2. False Positive rate (FPR). This measures the rate of wrongly classified instances. A low FP-rate signifies that the classifier is a good one.

$$\text{FPR} = \frac{\text{FP}}{\text{FP + TN}} \quad (2)$$

3 Sensitivity: It is the proportion of positives that are correctly identified

Sensitivity = TP / TP + FN (3)

= D/ (D + C)

4. Precision. Precision is the ratio of positively predicted instances among the retrieved instances

$$\text{Precision} = \frac{\text{TP}}{\text{TP +FP}} \quad (4)$$

5 Specificity SP: It is the proportion of negatives that are correctly identified. It is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate. The worst is 0.0 while the best is 1.0.

Specificity = TN / TN + FP (5)

= A / (A + B)

6. . Recall is the ratio of positively predicted instances among all the instances

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FP}} \quad (6)$$

7 Error Rate: It is equivalent to 1 minus Accuracy.

= (B + C) / (A+B+C+D) (7)

8. Root mean square error (RMSE). This is the standard deviation of the predicted error. Predicted error is the error between the training and testing dataset. A low RMSE indicates that the classifier is an excellent one

$$\text{RMSE} = \sqrt{1 - r^2}\ x\ SD \quad (8)$$

**Where** SD = Standard Deviation, r = Predicted error

**9.** Receiver Operating Characteristic (ROC) curve. The true positive rate is constructed against the false positive rate.

10. ROC Curve is Plot of FPR(x) vs TPR **Where** TPR is True Positive Rate

## 4.2 Experimental Results and Discussion

The experiment was carried out in order to evaluate the performance and usefulness of different classification algorithms and ensembles for predicting hypothyroid diseases.

Figure 2 shows the visualization of the hypothyroid dataset while the results of the experiments were shown in table 2, 3, 4, 5 and 6. the tables shows the accuracy of all our selected classifiers, the time taking to build the model and other metrics that measure their accuracy as applied on each of the dataset. Figures 3, 4, 5, 6, 7, 8 and 9 showed their graphical representation. Table 5 shows the summary of the performance accuracy of the classifiers and the ensembles while figures 6 showed its graphical representation.

**Table 2: Performance Evaluation Accuracy of single algorithms**

| Classifiers | Correctly Classified instances | Incorrectly Classified instances | Time taken to Build Mode/l(s) | MAE | RMSE | TP Rate | FP Rate | Kappa statistic | ROC | Confusion Matrix | Accuracy | Sensitivity | Specificity | Error rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 99.6023 | 0.3977 | 0.22 | 0.003 | 0.0405 | 0.996 | 0.019 | 0.9725 | 0.994 | a b c d<br>3476 2 3 0<br>1 192 1 0 3<br>3 89 0<br>2 0 0 0 | 0.997 | 0.980 | 1.002 | 0.003 |
| Simple Cart | 99.5493 | 0.4507 | 4.55 | 0.005 | 0.0462 | 0.995 | 0.007 | 0.9692 | 0.994 | a b c d<br>3474 2 5 0<br>0 193 1 0<br>0 7 88 0<br>2 0 0 0 | 0.998 | 0.993 | 0.998 | 0.002 |
| Bagging | 99.5493 | 0.4507 | 1.04 | 0.0052 | 0.0452 | 0.995 | 0.007 | 0.9692 | 0.994 | a b c d<br>3473 3 5 0<br>0 194 0 0<br>0 7 88 0<br>2 0 0 0 | 0.996 | 0.993 | 0.998 | 0.004 |

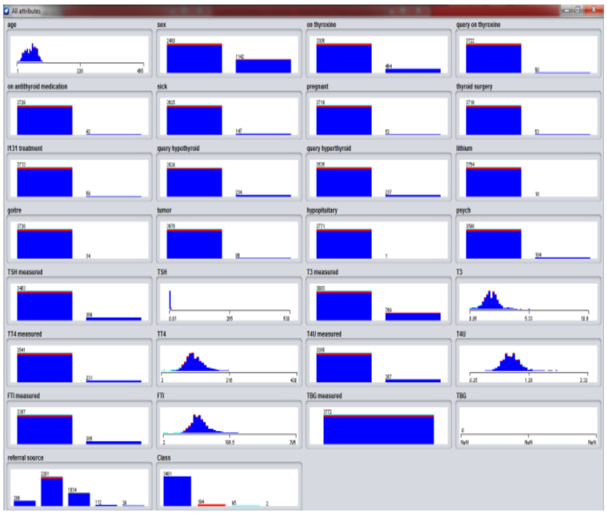


**Figure 2 : Visualization Of The Hypothyroid Disease Dataset**

### *4.2.1 Results Discussion*

In table 2, J48 algorithm has the best accuracy of 99.6023% when k=6 where k is the k fold cross validation and considering the time taking to build the model is very low (0.02s) compared to other classifiers above. SimpleCart and Bagging classifiers had same accuracies of 99.5492% but with different time taking to build the model which are 4.55s and 1.04s respectively. Therefore, J48 classification algorithm can be consider as the best classifier for prediction of hypothyroid.

However, Kappa statistics shows that all the classifiers used actually predicted well with 0.9722, 0.9692 and 0.9692 respectively, which are almost equals to 1. Also ROC result showed that the performances of the classifiers are very good as they are almost equal to 1

From the Confusion matrix column in the table 2, it can seen that the misclassified data's are few while most of the classes are well predicted with few or minor errors. Figure 3 shows the graphical representation of the performance accuracy with time of the classifiers prediction.

Table 2: Performance Evaluation Accuracy of single algorithms

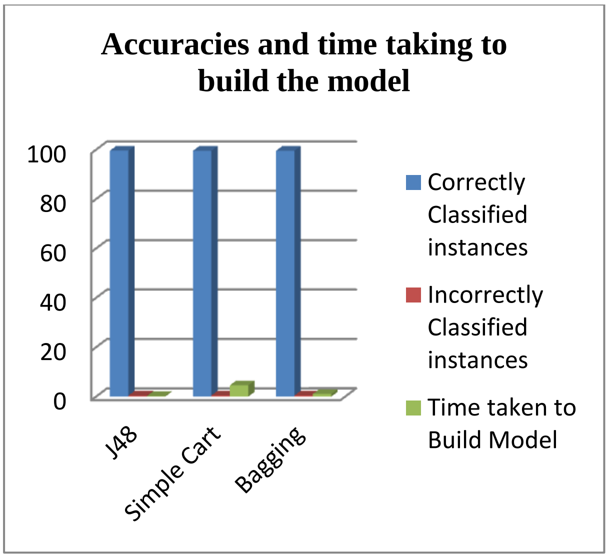


**Figure 3 : Accuracies and time taking to build the model**

Ensemble of SimpleCart algorithm correctly classified the instances with accuracy of 99.6554 while an Ensemble of J48 algorithm classified correctly with accuracy of 99.6023% as shown in table 3,. Therefore it will be a good practice to consider predicting with an Ensemble classification due to their excellent performance in predicting accuracy. However, Ensemble of SimpleCart algorithm outperformed the Ensemble of J48 in term of accuracy in the prediction of hypothyroid disease but in terms of time taken to build the model, Ensemble of J48 will be most preferred. Figure 4 shows the graphical representation of the prediction accuracies of the ensemble of both SimpleCart and J48 algorithm.

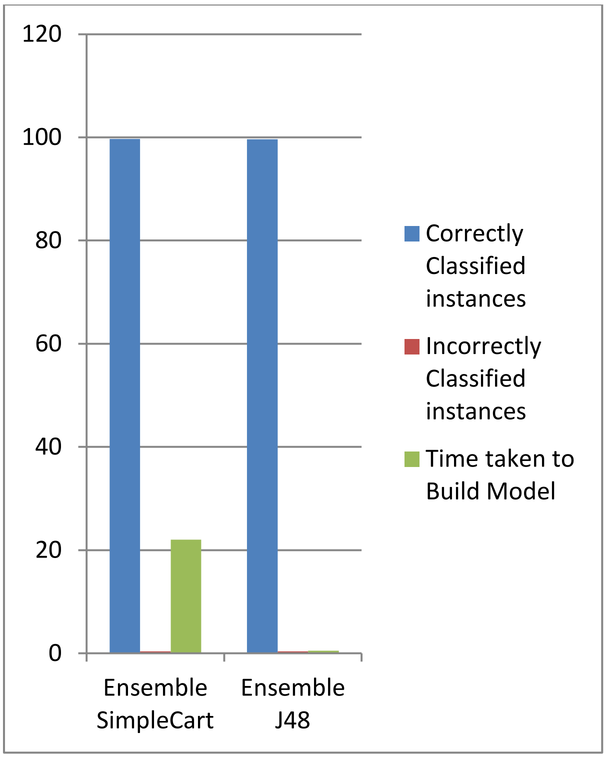


**Figure 4: prediction accuracies of the ensemble algorithms**

In table 4, ensemble of SimpleCart with Bagging algorithm predicted better with an accuracy of 99.6554% while only Simple cart, J48 and Bagging classified with accuracies of 99.5493%, 99.6023% and 99.5493% respectively. Figure 5 shows the graphical comparison accuracy of table 4

**Table 3: Accuracies of Ensemble Predictive Algorithms**

| **Accuracies of Ensemble Predictive Algorithms** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifiers | Correctly Classified instances | Incorrectly Classified instances | Time taken to Build Model /(s) | MAE | RMSE | TP Rate | FP Rate | Kappa statistic | ROC | Confusion Matrix | Accuracy | Sensitivity | Specificity | Error rate |
| Ensemble<br>Simple Cart | 99.6554 | 0.3446 | 22 | 0.0036 | 0.0404 | 0.997 | 0.006 | 0.9764 | 0.995 | a b c d<br>3475 2 4 0<br>0 194 0 0<br>0 5 90 0<br>2 0  0 0 | 0.998 | 0.993 | 0.998 | 0.02 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ensemble<br>J48 | 99.6023 | 0.3977 | 0.52 | 0.0033 | 0.0393 | 0.996 | 0.010 | 0.9727 | 0.995 | a b c d<br>3475 2 4 0<br>0 193 1 0<br>1 5 89 0<br>2 0 0 0 | 0.9976 | 0.989 | 0.998 | 0.02 |

**Table 4: Comparison of prediction accuracy of single and ensemble algorithm classification result**

| | J48 | Simple Cart | Bagging | Ensemble SimpleCart + Bagging | Ensemble J48 + Bagging |
|---|---|---|---|---|---|
| Correctly Classified instances | 99.6023 | 99.5493 | 99.5493 | 99.6554 | 99.6023 |

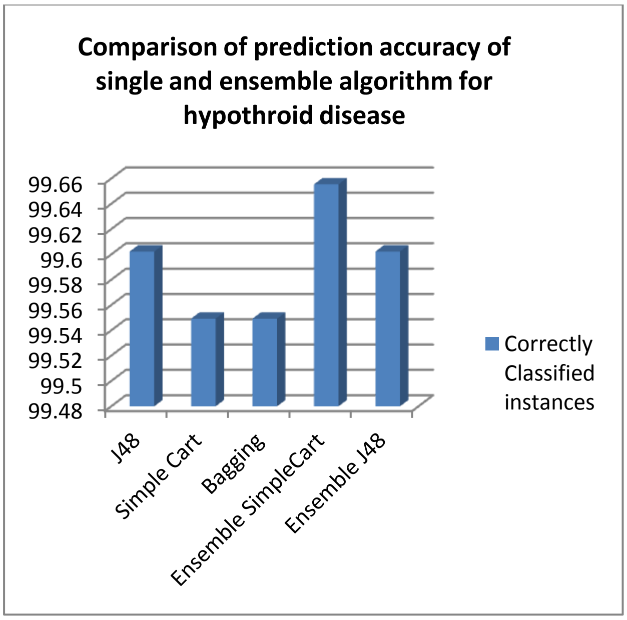


**Figure 5: Comparison of prediction accuracy of single and ensemble algorithm for hypothyroid disease**

**Table 5: Metrics performance**

| Classifiers | Kappa statistic | Accuracy | Sensitivity | Specificity | Error rate |
|---|---|---|---|---|---|
| Ensemble SimpleCart | 0.9764 | 0.998 | 0.993 | 0.998 | 0.02 |
| Ensemble J48 | 0.9727 | 0.997 | 0.989 | 0.998 | 0.02 |

Table 5 shows the metric performance of the two ensembles. The Kappa statistics is approximately equals to 1 on both sides of the models. This means that its in agreement with the result of the prediction. However, Ensemble with SimpleCart has the highest Kappa statistics result.

The accuracy of ensemble SimpleCart is also the highest with 0.998 while ensemble with J48 has the accuracy of 0.997. Figure 6, 7, 8 and 9 shows the graphical representation of the metrics accuracy of the model.

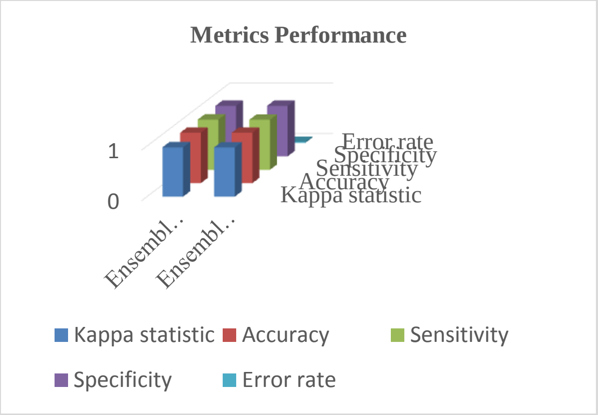


**Figure 6: Metrics performance**

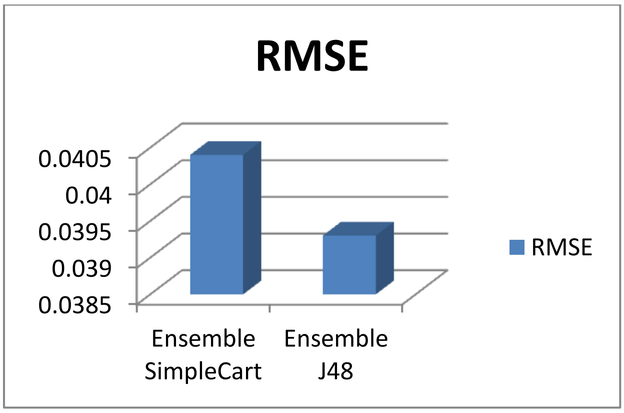


**Figure 7: Root Mean Square Error performance of the two ensemble algorithms for hypothyroid disease**

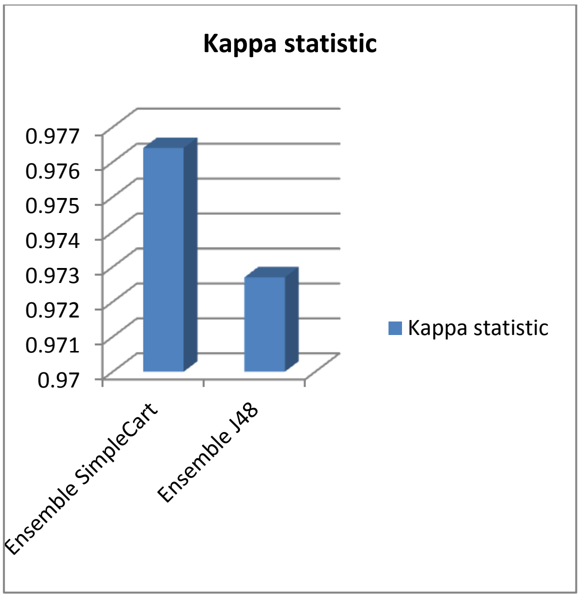


**Figure 8: Kappa statistic metric performance of the two ensemble algorithms for hypothyroid disease**

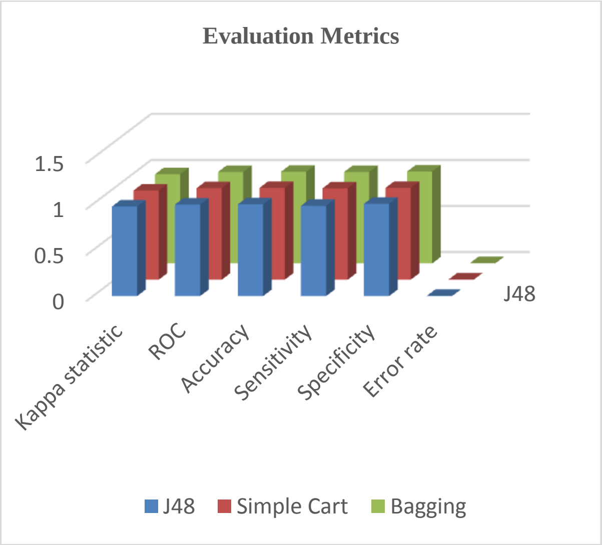


**Figure 9: Metrics Accuracies**

## 4.3 Comparison with the results of related work

Table 6 shows a brief comparison of the result of the proposed system with the results of other related works. Therefore, from the result table 6 this research model outperformed other models

## 5. CONCLUSION

In this paper, an approach for hypothyroid disease prediction using Ensemble classification machine learning algorithm has been discussed. An Ensemble model of Bagging with J48 and ensemble of bagging with SimpleCart model on the hypothyroid dataset were created and experimented, the models result output actually achieve effective and accurate predictions. Their performance accuracies were compared also with single selected algorithms. However, evaluation of the prediction accuracy of the model was done in two ways. One is prediction accuracy and the second is time taken for the prediction to build the model. Based on the obtained results, ensemble of bagging with simplecart algorithm has the highest predictive accuracy of 99.6554% while ensemble of bagging with J48 yielded an accurate result of 99.6023% in less time to build the model.

**Table 6: Comparison of proposed system with the results of other related works**

| Authors | Algorithms | Datasets | Results |
|---|---|---|---|
| Mrs.K.Sindhya, 2020 | Naïve Bayes, J48, Random forest | Hypothyroid (UCI) | 95%,99%,99.3% |
| Arvind Selwal & Ifrah Raoof, 2020 | Multilayer perceptron (MLP) | Hypothyroid (SKIMS Hospital, Jammu and Kashmir) | Approx 99.8% |
| Suwarna Gothane, 2020 | ZeroR | Hypothyroid (UCI) | 92.2853 % |
| Yasir Iqbal Mir, Dr. Sonu Mittal, 2020 | Support Vector Machine, Naïve Byes, J48, Bagging, Boosting | Sawai Man Singh (SMS) hospital. India | Svm- 99.08%, J48- 92.07%<br>Bagging 98.56% |
| Ankita Tyagi, Ritika Mehra & Aditya Saxena, 2019 | ANN, KNN, SVM, DT | Hypothyroid (UCI) | 97.50%, 98.62%, 99.63%, 75.76% |
| Shivanee Pandey, Rohit Miri, & S. R. Tandan, 2013 | multilayer perceptron, RBF network, Bayes Net, C4.5 CART, Decision stump, REPtree | Hypothyroid (UCI) | 94.035%, 95.228%, 98.59%, 99.57%, 99.54%, 95.38%, 99.57% |
| Irina IoniŇă & Liviu IoniŃă,2016 | SimpleCART, J48, MLP, RBF Network, Naive Bayes | Hypothyroid (UCI) | 89.58%,89.68%, 77.08%, 79.16% 70.83% |
| G. Rasitha Banu, PhD, 2016 | Linear Discriminant analysis (LDA) | Hypothyroid (UCI) | 99.62% |
| **Proposed** | J48, SimpleCart, Bagging, Ensemble (Bagging+J48) and Ensemble (Bagging+SimpleCart) | Hypothyroid (UCI) | 99.6023%, 99.5493%, 99.5493%, 99.6023%, 99.6554% |

## 6. FUTURE WORK

More attributes in a medical dataset means that patients has to undergo much numbers of clinical tests which might not be cost effective and as well time consuming. Thus, there is a need to develop a model with thyroid disease predictive models which require minimum number of parameters or test attributes required by patient for the diagnose of thyroid disease and saves both money and time of the patient.

# 7. REFERENCES

[1] AlirezaOsarech, & BitaShadgar. (2011). A Computer Aided Diagnosis System for Breast Cancer. International Journal of Computer Science Issues , 8 (2).

[2] Ankita Tyagi, R. M. (2019). Interactive Thyroid Disease Prediction System Using Machine Learning Technique. 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), (pp. 689-693). Solan, India: IEEE.

[3] Arvind Selwal, I. R. (2020). A Multi-layer perceptron based intelligent thyroid disease prediction system. Indonesian Journal of Electrical Engineering and Computer Science , 17 (1), 524-533.

[4] Banu, G. R. (2016). Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique. Communications on Applied Electronics (CAE) , 4 (12), 1-6.

[5] Gothane, S. (2020). Data Mining Classification on Hypo Thyroids Detection: Association Women Outnumber Men. International Journal of Recent Technology and Engineering (IJRTE) , 8 (16), 601-604.

[6] Irina IoniŃă, L. I. (2016). Prediction of Thyroid Disease Using Data Mining Techniques. BRAIN. Broad Research in Artificial Intelligence and Neuroscience , 7 (3), 115-124.

[7] Mrs.K.Sindhya. (2020). EFFECTIVE PREDICTION OF HYPOTHYROID USING VARIOUS DATA MINING TECHNIQUES. EPRA International Journal of Research and Development (IJRD) , 5 (2), 311-317.

[8] Shivanee, P., Rohit, M., & Tandan. (2013). Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques. International Journal of Engineering Research & Technology (IJERT) , 2 (6), 3188-3193.

[9] Haifeng Wang and Sang Won Yoon (2019) – Breast Cancer Prediction using Data Mining Method, IEEE Conference paper

[10] Yasir, I. M., & Sonu, D. M. (2020). Thyroid Disease Prediction Using Hybrid Machine Learning Techniques: An Effective Framework. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH , 9 (2), 2868-2874.

[11] Akshaya Asokan. (2020, May). Basics of Ensemble learning in Classification Techniques Explained. Retrieved may 3rd, 2020, from analyticsindiamag: https://analyticindiamag.com/basics-of-ensemble-learning-in-classification-techniques-explained/

[12] Lichman M (2017). UCI Machine Learning Repository : Breast Cancer Wisconsin (Diagnostic) DataSet.2014. http://archive.ics.uci.edu/ml.Accessed 8 june 2020

[13] Payal Dhakate; K. Rajeswari;& Deepa Abin (2015): International Journal of Computer Applications (0975 – 8887) Volume 111 – No 5, February 2015.

[14] Vikas Chaurasia & Saurabh Pal (2013): International Journal of Advanced Computer Science and Information Technology (IJACSIT) .Vol. 2, No. 4, Page: 56-66, ISSN: 2296-1739. © Helvetic Editions LTD, Switzerland www.elvedit.com

[15] A. K. Santra, C. Josephine Christy," Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI International Journal of Computer Science Issues, Vol.9, Issue 1, No 2, January 2012