Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

HYBRIDIZATION OF MACHINE LEARNING ALGORITHM FOR THE PREDICTION OF HYPOTHYROID

*Khadijat Lami Abdulwahab, Rabi Mustapha

Department of Computer Science, Kaduna State University, Nigeria

*Corresponding Author Email Address: lamiabdulwahab@gmail.com

ABSTRACT
Thyroid disease is one of the most progressive endocrine disorders in the human population today, and prediction of this disease is a very critical task in the field of clinical data analysis. Machine learning (ML) has shown effective results in the decision-making and predictions from the enormous data generated in the healthcare domain. However, there are limited studies that combined hybridized machine learning classifiers with a hybrid pre-processing technique of solving imbalance data class problem in a medical dataset. This research is aimed at hybridizing machine learning algorithm for the prediction of hypothyroid using dual preprocessing technique of SMOTE (Synthetic minority over sampling) and RESAMPLE. This study design a hybrid machine learning algorithm for the prediction of hypothyroid using dual filtering pre-processing technique of SMOTE and RESAMPLE to handle data class imbalance as one of its objectives and evaluate the hybrid algorithm with the dual pre-processing technique and without the dual pre-processing technique. This study used WEKA 3.8.3 as the tool of analysis. Four machine learning classification algorithms were compared (J48, Random Forest, Simple Logistic and AdaboostM1) both as a single and hybrid algorithm. Public dataset obtained from UCI repository was used for this work and AdaboostM1 combine with J48 achieved highest accuracy of 99.8% when pre-processed with the combination of SMOTE and Resample technique.

Keywords: Hybridization, Machine Learning, Algorithm, Prediction, Hypothyroid.

INTRODUCTION

Thyroid Disease (TD) has become one of the most widespread endocrine disorders worldwide. Although the cause of TD is still unknown, however the symptoms of TD can be reduced if the illness is identified at an early stage (Yasir & Sonu, 2020). TD is one of the most progressive endocrine disorders in the human population today and prediction of the endocrine disease is a critical task in the field of clinical data analysis. It is not easy to identify TD because of a variety of threatening factors such as high cholesterol, high blood pressure, unusual pulse rate and various other factors (Yasir & Sonu, 2020). However, machine Learning (ML) has shown effective results.

In healthcare and medical science, the applications based on data mining are very beneficial and important. The large amount of data gathered from health care organization has no organizational value unless transformed into most useful information and knowledge, which could be helpful in cost controlling, increasing the profits, and high-quality maintenance of patient healthcare (Srinidhi & Prabha, 2021). Disease diagnosis is the method of determining which disease explains a person's symptoms. Some symptoms and signs are non-specific, and thus the most challenging problem is the diagnosis (Ibrahim &Abdulazeez, 2021). Over some period of time, machine learning algorithms have played a crucial role in solving the complex and nonlinear problems in developing a prediction model. In any disease and infection prediction, models are required to identify the fundamental features that can be chosen from the distinctive datasets that can easily be used as a classification in the healthy patient as precisely as possible. Otherwise, misclassification may result in a healthy patient receiving unnecessary treatment (Awujoola et al. 2020). 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 (Irina, 2016). Machine learning is a modern way of computing where knowledge along with a technique is used to build a model which imitates the behavior of human being. Once the machine learning classification model is trained it will start predicting the class of a given feature set (Awujoola et al., 2020). Machine learning offers the ability for machines to learn without being specifically programmed. Developing a model by machine learning algorithms can predict an early-stage diagnosis of disease and provide solutions. An early diagnosis and effective treatment are the best way to minimize the death rates induced by any disease. Therefore, most medical scientists are drawn to new predictive model technologies for disease prediction based on machine learning algorithms (Battineni et al. 2020).

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 (Awujoola et al., 2020). 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 (Arvind, 2020).

Review of Related Literature

Banu (2016) 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. In the same vein, Shivanee et al. (2016) 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. Ankita (2019) 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%. More so, the authors Gu et al. (2019), conducted a research on a total of 103 patients (training cohort–to-validation cohort ratio, ≈ 3:1) with suspected thyroid nodules who had undergone thyroid ectomy and immune his to chemical analysis were enrolled. The immunohistochemical markers were cytokeratin 19, galectin 3, thyroperoxidase, and high-molecular-weight cytokeratin. All patients under went CT before surgery, and a 3D slicer was used to analyze images of the surgical specimen. The best performance of the cytokeratin 19 model yielded accuracy of 84.4% in the training cohort and 80.0% in the validation cohort. The thyroperoxidase and galectin 3 predictive models yielded accuracies of 81.4% and 82.5% in the training cohort and 84.2% and 85.0% in the validation cohort. The performance of the high-molecular weight cytokeratin predictive model was not good (accuracy, 65.7%) and could not be validated.

Abd-Elsalam et al. (2020) proposed to build a predictive Esophageal Varices diagnosing model that uses a minimum number of the most significant variables, trying to avoid unneeded endoscopy procedures. The dataset included more than forty individual clinical laboratory variables, only ten of them were found to be significant, which was achieved via merging the correlation coefficient and p-value filtering methods to acquire improved results.
To improve the overall performance of the predictive diagnosis model, this research introduced a novel algorithm that improves the traditional Naïve Bayes Tree by adding a boosting technique, dubbed "Boosted-Naïve Bayes Tree" (B-NBT). Applying B-NBT on the dataset revealed an improved performance in both AUROC (Area under Receiver Operating Characteristic Curve) of 0.865 and Accuracy of 79%.

Awujoola et al., (2020) developed an ensemble of Bagging with J48 and ensemble of Bagging with SimpleCart to extract useful information and diagnose for hypothyroid diseases prediction. The experiment carried out with ensemble of Bagging with Simple Cart yielded the highest predictive accuracy of 99.6554% while Bagging with J48 resulted in 99.6023% accuracy.
The study of Wang et al., (2020) carried out a study and enrolled consecutive patients who underwent neck MR scans and subsequent thyroid ectomy during the study interval. The diagnosis and aggressiveness of PTC were determined by pathological evaluation of thyroidectomy specimens. Thyroid nodules were segmented manually on the MR images, and radiomic features were then extracted. The experiment carried out with the combination of Least Absolute Shrinkage and Selection Operator for radiomic feature selection and Gradient Boosting Classifier for classifying PTC aggressiveness achieving the AUC of 0.92. In contrast, clinical characteristics alone poorly predicted PTC aggressiveness, with an AUC of 0.56.
Same study by Range et al. (2020) proposed on 6 categories used for the diagnosis of thyroid fine-needle aspiration biopsy (FNAB). Each category has an associated risk of malignancy, which is important in the management of a thyroid nodule. An MLA was designed to identify follicular cells and predict the malignancy of the final pathology. The test set comprised cases blindly reviewed by a cytopathologist who assigned a TBSRTC category. The area under the receiver operating characteristic curve was used to assess the MLA performance. Nine hundred eight FNABs met the criteria. The MLA predicted malignancy with a sensitivity and specificity of 92.0% and 90.5%, respectively.

Mourad et al. (2020), conducted a research and tested three separate neural network models to determine the outcomes of thyroid cancer patients after diagnosis from distilling the U.S. Surveillance Epidemiology and End Results (SEER) database. This study has led to the most accurate method to date utilized to predict thyroid cancer survival using data compiled from the SEER program registry. They validated their network through a direct comparison to an ANN generated using the AJCC TNM staging system. As a final step, in order to justify the use of machine learning-based algorithms to carry out this classifying task, PLS-DA models have been calculated for statistical comparison. PLS-DA is a classic mathematical approach based on creating linear regressions to estimate categorical variables. Three PLS-DA models have been calculated using the same independent variables and datasets as MLP-1, MLP-2, and MLP-3. All calculations performed for this manuscript have been completed via MATLAB.

Iqbal and Mittal (2020) proposed three novel models based on the primary dataset collected from1464 Indian patients. In these models, they compared top five ML algorithms (Support Vector Machine, Naïve Byes, J48, Bagging, and Boosting). The first model achieved the highest accuracy of 98.56% with bagging on both parameters, the second model that is based on pathological observations of the patient yielded the highest accuracy of 99.08 with SVM. 92.07% was obtained in the third model with J48 classifier on the serological tests.
Gothane (2020), presented 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%. In addition, Sindhya (2020) 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.

Yasir and Sonu (2020) 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.
The study by Park and Lee (2021), carried out a study and analyzed the prognostic significance of clinico-pathologic factors, including the number of metastatic lymph nodes (LNs) and lymph

node ratio (LNR), in patients with papillary thyroid carcinoma (PTC), and attempted to construct a disease recurrence prediction model using machine learning techniques. The experiment was carried out using decision Tree model which showed the best accuracy at 95%, and the light GBM and stacking model together showed 93% accuracy. Among those factors mentioned above, LNR and contralateral LN metastasis were used as important features in all machine learning prediction models. They confirmed that all machine learning prediction models showed an accuracy of 90% or more for predicting disease recurrence in PTC.

In the study of Ibrahim and Abdulazeez (2021) conducted a study using machine learning algorithms in the medical sector for diagnosing diseases from the medical database**,** for the early prediction of diseases and enhance medical diagnostics. The motivation of this paper is to give an overview of the machine learning algorithms that are applied for the identification and prediction of many diseases such as Naïve Bayes, logistic regression, support vector machine, K-nearest neighbor, K-means clustering, decision tree, and random forest. After comparing nineteen papers for different models that predicted diseases, it illustrated that many algorithms have good accuracy for predicting SVM, K-nearest neighbors, random forest, and the decision tree. Nevertheless, the accuracy of the same algorithm may differ from one dataset to another because many important factors affect the model's accuracy and performance, like datasets, feature selection, and the number of features.

Srinidhi and Prabha (2021), proposed a Random Forest Technique to identify and classify types of thyroids from the dataset. This study adopted the use of data mining classification and regression algorithms. However, both regression and classification are combined to produce accurate diagnosis of the thyroid diseases. From the result obtained, the logistic regression is more efficient and accurate compared to other classification techniques. Looking at the related work on Predictive Models with ranges of accuracies, an improved model can be developed based on data mining.

RESEARCH METHODOLOGY

This work used four data mining classification algorithms, which are J48, Random forest, simple logistic, and AdaBoostM1. However, a hybrid of AdaBoostM1 with J48, AdaBoostM1 with random forest, and AdaBoostM1 with simple logistic was developed. Then compare and evaluate the performance accuracy results of the three hybrids and with the single algorithm.

All the classification algorithms were selected because of their properties, very often used for medical data predictions and have potential to yield good results. Furthermore, the use of different approaches was used to generate the classification algorithm, which will increase the chances of finding a prediction algorithm with high classification accuracy. This research selected hypothyroid dataset which is publicly available from the University of California Irvine (UCI) Machine Learning Repository (Lichman, 2017). These dataset was chosen because of the prevalence of nominal features and their predominance in the literature.

Experimental Setup and Computer Program Used

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 is a data mining tool available free of cost. 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. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available. Originally written in C++ the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

MATERIALS AND METHODS

The experiment was carried out in order to evaluate the performance and usefulness of different classification algorithms and hybrid for predicting hypothyroid disease. The flow of the methodology is shown in Figure 1

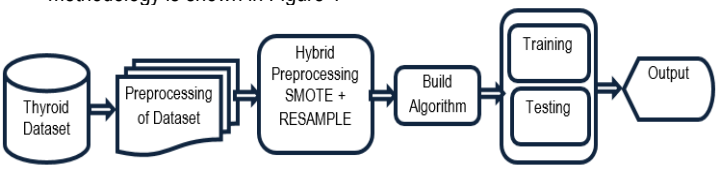

Figure 1: Flow of Methodology

Figure 1 shows the flow of methodology that was adopted in this research study. The dataset was sourced from University of California Irvin repository. The repository contained different types of datasets for related research work and most has been used as a benchmark for different domain. The Frame work of the methodology is shown in figure 2

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

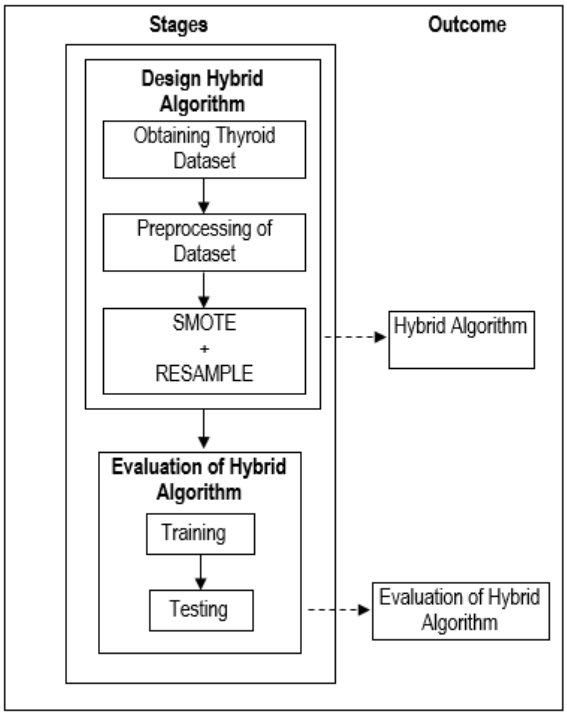

Figure 2: Methodology frame work
Adapted from Mustapha R. (2019) and Mustapha et al. (2020)

The figure 2 is the methodology frame work of this study. The frame work has two stages

i. The design stage
ii. The evaluation stage

The explanation of the frame work is in line with the explanation of figure 1.

Preprocessing

The datasets after collection will be pre-processed by converting them to appropriate file format that will suit the programming language for the development of the algorithm. Also the dataset will be subjected to feature selection procedure where some irrelevant attributes will be removed from the data. This will enhance the classifiers to obtained good accuracy in good time. Also the dataset will be further pre-processed by subjecting the dataset to resample and possibly invoke synthetic minority over sampling technique so as to balance the dataset.

Algorithm Evaluation and Performance

The experimental comparison of classification algorithms will be done based on the performance measures of classification accuracy, specificity, sensitivity, error rate, Kappa statistics ROC and execution time. The algorithm will be evaluated based on the following metrics.

Confusion Matrix

The actual and predicted classification done by a classification matrix is usually 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 (Santra& Josephine, 2012).

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

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

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:

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

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

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

$$\frac{TruePositive + TrueNegative}{TruePositive + FalseNegative + FalseNegative + TrueNegative}$$

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.

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

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

$$\text{Sensitivity} = TP / TP + FN \quad (3)$$

$$= D/ (D + C)$$

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

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

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.

$$\text{Specificity} = TN / TN + FP \quad (5)$$

$$= A / (A + B)$$

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

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

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

$$= (B + C) / (A+B+C+D) \quad (7)$$

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

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

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

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.

RESULTS AND DISCUSSION

Experiment 1: Classification/Prediction of Hypothyroid disease with single Algorithms without the Pre-processing technique.

Table 2 Hypothyroid Disease Prediction Results with Single Algorithms without Preprocessing

Classifiers	Correctly classified instances	Incorrectly Classified instances	TP Rate	FP Rate	Confusion Matrix	Accuracy	Sensitivity	Specificity	Error Rate
J48	99.5758	0.4242	0.999	0.021	a b c d 3476 3 2 0 1 191 2 0 3 3 89 0 2 0 0 0	0.9231	0.0000	0.9465	0.0769
Random forest	99.3107	0.6893	0.997	0.041	a b c d 3471 6 4 0 4 190 0 0 6 4 85 0 2 0 0 0	0.9234	0.0000	0.9457	0.0767
Simple logistic	96.6861	3.3139	0.996	0.320	a b c d 3466 7 8 0 86 105 3 0 6 13 76 0 1 0 1 0	0.9435	0.0000	0.9661	0.0565
AdaBoostM1	93.2131	6.7869	0.998	0.615	a b c d 3473 5 3 0 177 17 0 0 0 69 26 0 2 0 0 0	0.9682	0.0000	0.9757	0.03181

Table 2 shows the prediction results of hypothyroid disease with single algorithms without pre-processing the medical data of the disease. The result from the table shows, J48 has the highest prediction accuracy of 99.58%, followed by Random Forest classifier with 99.31% accuracy while Simple Logistic and AdaboostM1 has 96.69% and 93.21% respectively. Figure 3 express the prediction accuracy of single algorithms without preprocessing in a graphical form.

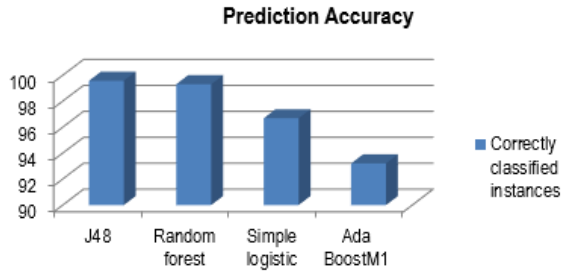

Figure 3: Prediction Accuracy of hypothyroid with single algorithms without preprocessing

Experiment 2: Classification/Prediction of Hypothyroid disease with Hybrid algorithm without the Pre-processing technique.

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Table 3: Hypothyroid Disease prediction Results with Hybrid Algorithms without Pre-processing

Classifiers	Correctly classified instances	Incorrectly classified instances	TP Rate	FP Rate	Confusion Matrix	Accuracy	Sensitivity	Specificity	Error Rate
(Hybrid) AdaBoostM1 +J48	99.5758	0.4242	0.997	0.024	a b c d 3472 3 4 0 4 190 0 0 1 0 94 0 2 0 0 0	0.9223	0.0000	0.9474	0.0771
(Hybrid) AdaBoostM1 + Random Forest	99.4963	0.5037	.998	0.024	a b c d 3473 5 3 0 1 193 0 0 4 4 87 0 2 0 0 0	0.9226	0.0000	0.9451	0.0774
(Hybrid) AdaBoostM1 + Simple Logistic	96.6861	3.3139	0.996	0.320	a b c d 3466 7 8 0 86 105 3 0 6 13 76 0 1 0 1 0	0.9435	0.0000	0.9661	0.0565

Table 3 reveals the prediction results of hypothyroid disease with hybrid algorithms without pre-processing the dataset as it was done in experiment 1. However, hybrid of AdaboostM1 with J48 has the highest classification accuracy of 99.58% followed by hybrid of AdaboostM1 + Random Forest with accuracy of 99.50% while AdaBoostM1 + Simple logistic have 96.69% accuracy. For clarity, figure 4 expresses the prediction accuracy of hybrid algorithms without pre-processing.

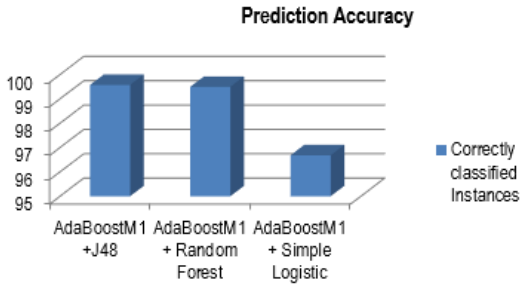

Figure 4: Prediction Accuracy of hypothyroid with hybrid algorithm without Preprocessing

Table 4: Summary of Hypothyroid Prediction Results with both Single Algorithms and Hybrid Algorithms

Single and Hybrid without preprocessing	
Classifiers	**Correctly Classified instances**
J48	99.5758
Random forest	99.3107
Simple logistic	96.6861
Ada BoostM1	93.2131
(Hybrid) AdaBoostM1 +J48	99.5758
(Hybrid) AdaBoostM1 + Random Forest	99.4963
(Hybrid) AdaBoostM1 + Simple Logistic	96.6861

The Table 4 shows the summary of the prediction result which contained both single and hybrid algorithms without preprocessing the dataset. Figure 5 expresses the prediction accuracy of Single and Hybrid without preprocessing. However, hybrid of (AdaboostM1 + J48) and J48 outperformed other algorithms

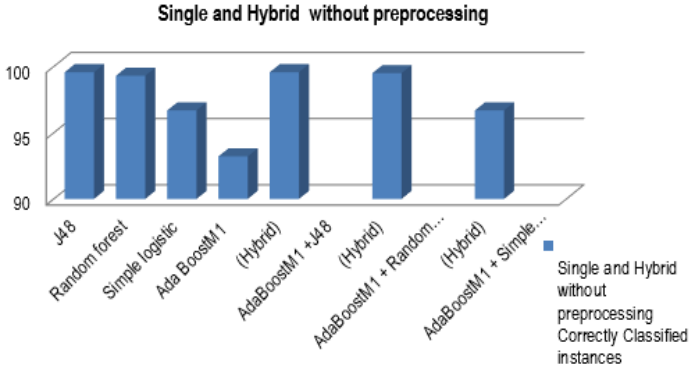

Figure 5: Prediction Accuracy of both Single and Hybrid Algorithm without preprocessing

Experiment 3: Classification/Prediction of Hypothyroid disease with Single Algorithms with the application of SMOTE and RESAMPLE as the Pre-processing technique

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Table 5: Hypothyroid Classification using single algorithms with SMOTE and Resample preprocessing Technique

Classifiers	Correctly classified Instances	Incorrectly classified Instances	TP Rate	FP Rate	Confusion Matrix	Accuracy	Sensitivity	Specificity	Error Rate
J48	99.7614	0.2386	0.999	0.003	a b c d 3479 2 3 0 0 193 0 0 0 3 9 1 0 1 0 0 0	0.9226	0.000	0.9462	0.0774
Random Forest	99.4963	0.5037	0.997	0.024	a b c d 3471 5 5 0 1 192 0 0 5 2 87 0 1 0 0 0	0.9229	0.000	0.9459	0.0771
Simple Logistic	96.9247	3.0753	0.996	0.292	a b c d 3469 6 9 0 78 110 5 0 5 12 77 0 1 0 0 0	0.9419	0.000	0.9652	0.0581
AdaBoostM1	93.6638	6.3362	0.989	0.427	a b c d 3447 30 7 0 122 71 0 0 0 79 15 0 1 0 0 0	0.9464	0.000	0.952	0.0536

Table 5 displayed the result obtained from the experiment 3 with single algorithms, where SMOTE and Resample were applied to handle the class imbalance in hypothyroid dataset. The result revealed that J48 algorithm outperformed others with 99.76% accuracy, followed by Random Forest with prediction accuracy of 99.50%. Simple logistic algorithm has 96.92% accuracy while AdaBoostM1 is the least with 93.66% accuracy. Comparing the result obtained in table 2 where there was no application of SMOTE and Resample, it's obvious that there is great difference in the performance of the classifier accuracy in table 5

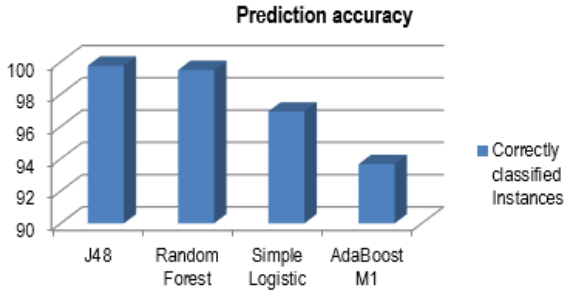

Figure 6: Hypothyroid Prediction Results of Single Algorithm with application of SMOTE and RESAMPLE

Table 6: Summary of Hypothyroid Experimental Results for both Pre-processed and without pre-processed using single algorithms

	Pre-Processed	Without Pre-processed
Classifiers	Accuracy	Accuracy
J48	99.7614	99.5758
Random Forest	99.4963	99.3107
Simple logistic	96.9247	96.6861
AdaBoostM1	93.6638	93.2131

The Table 6 shows the summary of the prediction results for both pre-processed and without pre-processed using single algorithms. However, from the table J48 outperformed other algorithms with an accuracy of 99.76% when pre-processed with the dual pre-processing technique. Figure 7 expresses the prediction accuracy of hypothyroid for both Pre-processed and without pre-processed using single algorithms.

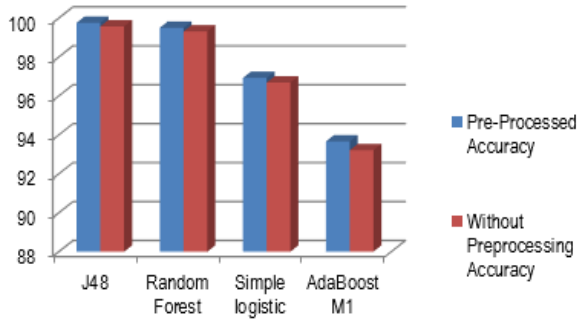

Figure 7: Summary of Hypothyroid Experimental Results for both Pre-processed and without pre-processed using single algorithms

Science World Journal Vol. 18(No 1) 2023
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Experiment 4: Classification/Prediction of Hypothyroid disease with Hybrid algorithm with the application of SMOTE and RESAMPLE as the Pre-processing technique

Table 7: Hypothyroid Classification using single algorithms with SMOTE and Resample preprocessing Technique

Classifiers	Correctly classified Instances	Incorrectly classified Instances	TP Rate	FP Rate	Confusion Matrix	Accuracy	sensitivity	Specificity	Error Rate
(Hybrid) AdaBoostM1 +J48	99.8409	0.1591	0.999	0.003	a b c d 34769 2 3 0 0 193 0 0 0 0 94 0 1 0 0 0	0.9223	0.000	0.9469	0.0774
(Hybrid) AdaBoostM1 + Random Forest	99.4698	0.5302	0.997	0.028	a b c d 3474 5 5 0 1 192 0 0 6 2 86 0 1 0 0 0	0.9231	0.000	0.9459	0.0769
(Hybrid) AdaBoostM1 + Simple Logistic	96.9247	3.0753	0.996	0.292	a b c d 3469 6 9 0 0 78 110 5 0 5 12 77 0 1 0 0 0	0.9419	0.000	0.9652	0.0581

Table 7 shows the prediction results of hypothyroid disease with preprocessing of the data with the construction of hybrid algorithm. SMOTE and Resample technique was applied to handle the imbalance in the dataset. However, hybrid of AdaboostM1+J48 has the highest classification accuracy of 99.84% followed by hybrid of AdaboostM1 + Random Forest with accuracy of 99.47% while AdaboostM1 + Simple logistic have 96.92% accuracy. For clarity, Figure 8 expresses the prediction accuracy of hybrid algorithm with pre-processing

Figure 8: Hypothyroid Disease Prediction Results of hybrid Algorithm with application of SMOTE and RESAMPLE

Table 8: Summary of Hypothyroid Experimental Results for both Pre-processed and without Pre-processed with hybrid

Pre-Processed with Hybrid		Without Pre-processed with Hybrid
Classifiers	Accuracy	Accuracy
AdaBoost M1 + J48	99.8409	99.5758
AdaBoost M1 + Random Forest	99.4698	99.4963
AdaBoost M1 + Simple Logistic	96.9247	96.6861

Table 8 shows the prediction results of the two experiments with both Pre-processing and without pre-processing of the disease data. However, from the table hybrid of AdaboostM1 + J48 outperformed other algorithms with an accuracy of 99.84% followed by AdaBoostM1 + Random Forest and AdaBoostM1 + Simple Logistic with 99.47% and 96.92% respectively. Figure 9 expresses the prediction accuracy of hypothyroid for both Pre-processed and without pre-processed with hybrid algorithm.
This revealed that hybrid actually predicts better with the application of SMOTE and RESAMPLE

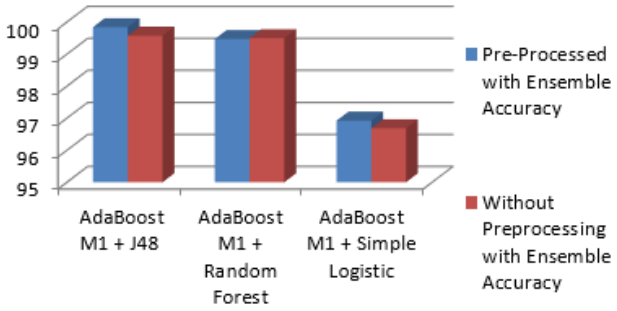

Figure 9: Summary of Hypothyroid Disease Experimental Results for both Pre-processed and without Pre-processed using hybrid algorithms

Table 9: Comparison of the developed hybrid algorithm with the results of other related work

Authors	Algorithms	Datasets	Result
Yasir et al., (2020)	Support Vector Machine, Naïve Byes, J48, Bagging, Boosting	Sawai Man Singh (SMS) hospital. India	Svm- 99.08%, J48- 92.07% Bagging 98.56%
Suwarna, (2020)	ZeroR	Hypothyroid (UCI)	92.2853 %
Awujoola et al., (2020)	J48,SimpleCart, Bagging, Ensemble (Bagging+J48) and Ensemble (Bagging+SimpleCart)	Hypothyroid (UCI)	99.6023%, , 99.5493%, 99.5493%, 99.6023%
Rasitha,(2016)	Linear Discriminant analysis (LDA)	Hypothyroid (UCI)	99.62%
Shimaa et al, (2020)	"Boosted-Naïve Bayes Tree	prospective cohort of 5013 chronic hepatitis C Egyptian patients collected	79%
Satish et al., (2020)	KNN, Random Forest	Egyptian patient's dataset	51.06% 54.56%
Young &Byung-Joo, (2021)	Decision tree Random forest XGBoost LightGBM	analyzed cl inico-pathologic data from 1040 patients	95%,91%,92%,93%
Ankita et al.,(2019)	ANN, KNN, SVM, DT	Hypothyroid (UCI)	97.50%,98.62%, 99.63%, 75.76%
Developed	**(Hybrid)** AdaBoostM1+J48, **(Hybrid)** AdaBoostM1 + Random Forest, **(Hybrid)** AdaBoostM1+Simple Logistic	Hypothyroid (UCI)	99.8409, 90.3226 99.4698

Conclusion and Future Work

In this paper, an approach for prediction / classification of medical disease hypothyroid based on hybrid machine learning algorithm was developed. This work compared results obtained from four single machine learning algorithms with hybrid algorithms, However, the hybridized algorithms and pre-processing techniques combining both SMOTE and Resample to solve imbalance data problem outperformed ordinary hybrid algorithm. The hybrid of AdaboostM1 + J48 achieved highest accuracy of 99.8% when pre-processed by hybrid technique of SMOTE and RESAMPLE. However, for future work, it is recommended that more medical datasets and classifiers be considered so as to test the generalization of the Hybrid Algorithm

REFERENCES

Arvind S 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.

Awujoola O, J, Ogwueleka, F., &Odion, P, O. (2020). Effective and accurate bootstrap aggregating (bagging) ensemble algorithm model for prediction and classification of hypothyroid disease. *International Journal of Computer Applications*, *176*(39), 40–48. https://doi.org/10.5120/ijca2020920542

Ankita, L., &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)*, 689-693.

Banu, G, R. (2016). Predicting thyroid disease using linear discriminant analysis (LDA) data mining technique. *Communications on Applied Electronics (CAE)*, 4(12), 1-6.

Battineni, G., Sagaro, G, G., Chinatalapudi, N., &Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine*, *10*(2), 21. https://doi.org/10.3390/jpm10020021

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.

Ibrahim, I., &Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, *2*(01), 10–19. https://doi.org/10.38094/jastt20179

Iswanto, I., Laxmi, L., Shankar, K., Phong, T, N., Wahidah, H. (2019). Identifying diseases and diagnosis using machine learning. *International Journal of Engineering and Advanced Technology*, *8*(2), 978–981. https://doi.org/10.35940/ijeat.f1297.0886s219

Irina, L., &IoniŃă, L, I. (2016). Prediction of thyroid disease using data mining techniques. brain. *Broad Research in Artificial Intelligence and Neuroscience*, 7 (3), 115-124.

Mustapha R. (2019).An Enhanced Computational Integrated Decision Model for Prime Decision Making in Driving. University utara Malaysia Pp. 7.

Mustapha, R.,Ahmad, M. A., Daniel, A., Ahmed, M. A., & Hussaini, M. (2020). Validating Measures of Driver Behavior's Training Factors for Prime Decision-Making.

Science World Journal, 15(1), 102–112.

Shivanee, P., Rohit, M., &Tandan, P. (2013). Diagnosis and classification of hypothyroid disease using data mining techniques. *International Journal of Engineering Research & Technology (IJERT)*, 2(6), 3188-3193.

Sindhya, K. (2020). Effective prediction of hypothyroid using various Data mining techniques. *EPRA International Journal of Research and Development (IJRD)*, 5(2), 311-317.

Santra, K, A., & Josephine, C. (2017). Genetic algorithm and confusion matrix for document clustering. *IJCSI International Journal of Computer Science Issues*, 9(1), 25-39.

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.