International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.41, September 2023 – www.ijais.org

Development of Thyroid Disease Prediction Model in Nigeria

Terlumun Emmanuel Togor
Department of Computer Science, Faculty of Computing, Federal University of Lafia P.M.B 146 Lafia, Nigeria

Joshua Abah
Department of Computer Science, Faculty of Computing, Federal University of Lafia P.M.B 146 Lafia, Nigeria

Dekera Kenneth Kwaghtyo
Department of Computer Science, Faculty of Computing, Federal University of Lafia P.M.B 146 Lafia, Nigeria

ABSTRACT
Thyroid Diseases are often diagnosed in Nigeria's endocrinology clinics and are reported as the second most common endocrine disease encountered amongst Nigerians, especially women. This led to a vast volume of thyroid data. When applied to machine learning algorithms, the data can greatly benefit endocrinologists, patients, and health organizations. However, diverse machine-learning methods have been in place for thyroid disease prediction leveraging the UCI thyroid data for experimental purposes. Conversely, this study differs by utilizing indigenous thyroid data to predict the thyroid conditions: hypothyroidism, hyperthyroidism and euthyroidism. The dataset was pre-processed using pandas and NumPy libraries. Random Forest classifier and Support Vector Machine classifier were trained using the indigenous dataset. Experimentally, the model was evaluated using Accuracy, Precision, F1-measure, Sensitivity and the Receiver Operating Characteristic (ROC) curve – Area Under the Curve AUC. The classification results show that the Random Forest classifier obtained the best accuracy of 99.30%, while the Support Vector Machine classifier achieved an accuracy of 98.60%. The study achieved the goal of turning the data comprising of thyroid conditions gathered from the Federal Medical Centre, Yenagoa, Bayelsa State, Nigeria into a powerful tool to support endocrinologists, patients, and health organizations.

General Terms
Thyroid-disease prediction, Nigeria endocrinological data

Keywords
Thyroid disease, Hypothyroidism, Hyperthyroidism, Euthyroidism, SVM, Random-Forest, Endocrinology

1. INTRODUCTION
Thyroid problems are also common in Nigeria as it is in other parts of the world. According to Ogbera et al. [1], in Nigeria's endocrinology clinics, thyroid problems are often detected and reported as being the second most common endocrine disease encountered amongst Nigerians, especially women. By definition, thyroid disease is a benign or malignant illness that harms or damages the structure and normal functioning of the thyroid gland [2]. There are several kinds of thyroid diseases, however, the commonly known types include hyperthyroidism and hypothyroidism. The earlier is the over-production of thyroid hormone while the latter is the under-production of the thyroid hormone. Hyperthyroidism is caused by Graves' disease, nodules, thyroiditis, and excessive iodine found in some food or medications. Hypothyroidism is caused by thyroiditis, Hashimoto's thyroiditis, postpartum thyroiditis, iodine deficiency, and non-functioning of the thyroid gland [3].

Generally, the major cause of any thyroid disease whether hyperthyroidism or hypothyroidism is the occurrence of an imbalance in the production and circulation of thyroid hormones, inflammation or physical damage to the thyroid gland. The thyroid gland is a soft reddish organ that looks like H in terms of structure. It consists of left and right lobes with a single isthmus that joins them in the anterior neck. Essentially, the thyroid gland takes iodine contents which are deposited in the food which the body consumes and secretes two primary active hormones, namely Triiodothyronine (T3) and thyroxine (T4) [4]. These hormones regulate breathing, muscle strength, Heart rate, cholesterol levels, body weight, body temperature, and menstrual cycles [5].

To diagnose hypothyroidism and hyperthyroidism conditions, examination of clinical parameters such as Total Triiodothyronine Test (T3), Total Thyroxine Test (T4), Thyroid-stimulating Hormone Test (TSH), Free Total triiodothyronine test (FT3), Free Total Thyroxine Test (FT4) and Thyroid-stimulating Hormone (TSH) is needed [6]. These test parameters generate a high volume of data in healthcare which requires the application of ML techniques for quick detection and treatment. Consequently, several authors have employed both supervised and unsupervised ML techniques to diagnose thyroid problems [7]–[10]. The unison goal of these techniques is to support endocrinologists with efficient means to carry out thyroid disease prediction seamlessly for early and proper management.

Despite the good performance of the existing models, there are intrinsic limitations associated with the models in addition to geographic and cultural variations. Specifically, the UCI thyroid dataset used in the works of [5], [7]–[9], and several other studies may not yield a realistic result for thyroid prediction in Nigeria. The dataset hosted by the UCI is possible to have genetic disparities as a result of choices of food, lifestyle, culture and geographic locations. Besides, the UCI data may be outdated and have many missing values. It was collected as far back as 1987 by Ross Quinlan which new symptoms may not be captured in the old data available online. Thus, need a region-specific thyroid dataset for want of a reliable model in Nigeria.

This study is therefore motivated by the need to utilize a locally collected thyroid dataset in Nigeria to build an ML model for the diagnosis of thyroid problems in Nigeria. The study leveraged Multi-class SVM and Polynomial Random Forest algorithms to develop the thyroid prediction model. This research aims to build an ML model with an indigenous

thyroid dataset to classify and predict the thyroid conditions as whether hypothyroidism, hyperthyroidism and euthyroidism (normal) with high precision. The main contribution of the study is the development of a valuable decision-supporting system for endocrinologists in Nigeria to accelerate the prediction of thyroid disease for better management at an early stage.

The remainder of the paper is structured thus: Segment (2) centres on reviewing existing works. Segment (3) explains the materials and methods adopted. Segment (4) dwells on the experimental setup. Segment (5) gives the experimental results and the discussion of the study. Segment (6) concludes and provides suggestions for further research.

2. REVIEW OF RELATED WORKS

In recent years, much research has been done towards using machine learning models for the diagnosis of thyroid disorders. Most of these models have proved to produce good predictive results. Consequently, the performance of LR and SVM algorithms was analyzed by [6] to classify thyroid disease. UCI thyroid dataset was analyzed using WEKA version 3.6. The performance analysis of the thyroid disease was compared based on Precision, Recall, F-measure, ROC curve, RMS Error, and accuracy. The comparative analysis reviewed that logistic regression performed better than SVM based on both classification accuracy and all the performance parameters. Also, the research suggested that SVM is more stable for binary classification than logistics regression, but for multiclass classification, logistics regression performed more than SVM at 96.84%. Also, an interactive ML model for thyroid disease diagnosis was developed in a study conducted by [7]. The model provides comparative diagnostic and prediction of thyroid disease analysis with four basic algorithms: ANN, SVM, K-NN, and DT. For performance evaluation, Accuracy and Mean Absolute Error matrices were used. UCI thyroid dataset was leveraged to train the algorithms to carry out prediction. It was reported that ANN gave the highest performance with 97.50% accuracy.

The authors in [11], proposed fuzzy logic and LR algorithms for diagnosing hypothyroidism. The thyroid dataset was gathered from Imam Khomeini Clinic and Shahid Beheshti Hospital of Iran. In terms of predictive performance, the model was compared on Accuracy, Sensitivity, Specificity, and the area under the curve of the receiver operating characteristic metrics. The fuzzy logic classifier gives the highest predictive accuracy of 97%. In a separate study, the authors [12] proposed a framework for diagnosing hypothyroidism with seven different classification algorithms. The proposed algorithms were KNN, NN, NB, DT, RF, SVM and LR. Five metrics, namely accuracy, Kappa, F-score, precision, and Recall were employed for performance comparisons and analysis of all the algorithms. UCI thyroid dataset was used. The comparative study for using different metrics to judge the performance of all the classifiers shows kappa as the highest priority metric for evaluating the accuracy of the classifiers. In conclusion, the random forest classifier produces the highest accuracy of 96.5% prediction. Dharmarajan et al, in [13], proposed an integrated model to predict thyroid disease with three machine-learning algorithms. The proposed algorithms are SVM, NB, and DT. The thyroid data was collected from 500 patients in India. The integrated model was evaluated using Accuracy, Precision, Recall, and F-score matrices. According to the results of all the classifiers, the decision tree demonstrated the best classification accuracy result of 97.35%.

The study conducted in [14], designed a neural network model to carry out a reliable prognosis of thyroid cancer patients. The model was designed with a combination of ANN and MLP algorithms using thyroid cancer data collected from the United States Surveillance Epidemiology and End Results database. The two unique classifiers were built with the thyroid cancer data consisting of clinical variables and lymph nodes to differentiate patients who survived for more than ten years after diagnosis from patients who didn't live or survived within five years after diagnosis of thyroid cancer conditions. The Area under the curve of the receiver operating characteristic curve was the metric used to evaluate the performance of the classifiers. An accuracy of 94.5% cancer prognoses was achieved in this research. On the other hand, in the study [8], Reddy et al. designed a model for a quicker and more efficient technique for diagnosing thyroid problems. Six supervised learning techniques were experimented with using the UCI thyroid datasets. However, the XGBoost model outperformed with 99.57% accuracy in detecting thyroid disease. The authors, therefore, made recommendations for using the XGBoost model for thyroid disease detection in the medical field.

Borzouei et al, [15] compared LR and ANN techniques to diagnose hyperthyroidism and hypothyroidism as the most widespread thyroid disease. The thyroid data collected at Imam Khomeini Clinic and Shahid Beheshti Hospital in Hamadan was used to experiment with the model. Evaluation metrics such as the mean of the accuracy and area under the curve (AUC) were leveraged to ascertain the performance of the model. In terms of predicted results, the ANN performed better having recorded 96.3 % accuracy. Similarly, the authors in [9], developed a system that can detect all possible types of thyroid disease. This study created a Web App such that users can enter data on the interface and thyroid disease-predicted results would display on the screen. The Web App was designed using Python Flash, HTML5, and CSS, while the machine learning model was trained with 5 algorithms: SVM, decision tree, logistic regression, KNN, and ANN. The UCI machine learning repository obtained the thyroid disease datasets for creating the system. All the implemented algorithms obtained the following accuracy scores: the KNN algorithm obtained an accuracy score of 93.84%, the SVM algorithm obtained an accuracy score of 95.38%, the ANN algorithm obtained an accuracy score of 75.38%, the decision tree algorithm obtained an accuracy score of 92.3%, and logistic regression obtained accuracy score of 96.92% respectively.

Again, Salman and Sonuc [16], created a comparative model to examine the proficiency of machine-learning approaches in the classification of thyroid disease. The learning techniques adopted include SVM, RF, DT, NB, LR, K-NN, MLP, and linear discriminant analysis. These algorithms were used in the study to diagnose 2 categories of thyroid disease: hypothyroidism, and hyperthyroidism. The dataset collected from Alkindi in Iraq was used to train and test the model using a Python environment. The findings revealed that the RF model yielded the highest prediction result of 98.93%. The authors Vadhiraj et al. [17] built a machine-learning model for an ultrasound image classifier for thyroid nodules prediction. The model classified the thyroid nodules disease to ascertain whether the condition is benign or malignant. For the preprocessing and segmentation of the thyroid nodules image, the study used a median filter and image binarization techniques. For the classification of benign or malignant

conditions, the study compared the performance of SVM and ANN. These algorithms were compared using the accuracy, sensitivity, and specificity metrics. The concluded comparison and analysis revealed that the SVM algorithm achieved the highest classification accuracy of 96%.

Also, to classify thyroid disease [18] focused on the treatment trend of LT4 for people affected with hypothyroidism. The study leveraged the thyroid data generated from the Federico II hospital in Naples. To obtain the best result among the proposed classifiers, several classifiers were compared with different characteristics. The first sets of classifiers belong to Boosting Algorithms which include AdaBoost, Gradient Boosting, XGBC, and CatBoost. The second set of classifiers belongs to the Decision Trees classifier's family such as Decision Tree, Extra-Tree and Random Forest. The third group of classifiers include Naïve Bayes, KNN, ANN and MLP. Metrics like Accuracy, Precision, Recall and F-Score were employed in validating the performances of the classifiers. The extra-tree classifier produced the highest result on all the metrics. Accuracy was 84%, Precision obtained 85%, Recall record 84% and F-Score 84% respectively. In [10], predicted hypothyroidism and hyperthyroidism conditions with seven (7) machine learning algorithms. Algorithms such as DT, RF, GBoost, NB, K-NN, LR and SVM. The prediction of the two types of thyroid conditions was carried out on UCI thyroid datasets. Also, the performance evaluation of the seven (7) classifiers was assessed using precision, recall, accuracy, and F1-measure. In terms of success achieved in this study, the Logistics Regression classifier recorded the highest predictive accuracy of 84.48%.

Relatedly, the author in [19], diagnosed thyroid disorder with many machine-learning classification models. The number of algorithms utilized in this study are Random Forest, KNN, SVM, Linear Analysis, NB, Decision Tree, MLP and LR. The authors of this work directed their effort toward classifying thyroid disease into two categories which are hypothyroidism and hyperthyroidism based on thyroid data collected from 1250 Iraqi. The performance of all eight (8) classifiers is measured on a single Accuracy metric, which Random Forest found to produce the highest classification value of 98.97%. To accurately predict the risk of thyroid problems Islam et al, in [20], utilized ten (10) machine learning algorithms. The aim of comparing such multiple algorithms is to investigate which of the algorithms can produce the best performance in terms of prediction accuracy. Algorithms utilized are Decision Tree, Randon Forest, Light GBM, XGBoost, GaussianNB, KNN, ANN, SVC, CatBoost and Extra-Trees. The thyroid datasets used in this research were downloaded from the UCI repository Precision, F1-measure, Accuracy, and Recall were employed to evaluate the distinct performance of algorithms. The outcome of all the classifiers indicates that ANN came out with the best accuracy of 0.9587.

3. MATERIALS AND METHODS

3.1 Dataset Collection

The dataset used to experiment with this study was collected by going through each thyroid patient's medical records. Hence, picking out the laboratory test results (parameters) of Thyroid Hormones Screening for all thyroid patients that have been diagnosed in the hospital. The dataset consists of Laboratory Test Reports for T3, T4, FT3, FT4, and TSH of thyroid patients, which are manually kept in the thyroid repository of the Internal Medical Department at Federal Medical Center Yenagoa (FMCY), Bayelsa State of Nigeria. The data collection was conducted with the approval of the hospital ethics committee, File Number: FMCY/REC/ECC/2022/JANUARY/417. Figure 1 depicts the sample of the thyroid dataset collected.

	Age	Gender	T3	T4	FT3	FT4	TSH	DIAGNOSIS
0	53	F	0.93	6.60	0.00	0.00	0.8900	Hyperthyroidism
1	49	F	3.5	160.90	4.42	7.19	0.1000	Hypothyroidism
2	48	F	?	NaN	14.40	32.90	0.0046	Hyperthyroidism
3	40	F	3	3004.00	5.09	11.49	0.0075	Hyperthyroidism
4	43	F	3.2	1.57	0.00	0.00	0.4600	Hypothyroidism
5	29	F	0	0.00	0.36	0.25	94.3700	Hyperthyroidism
6	42	F	3.71	11.50	3.13	2.31	0.8900	Hyperthyroidism
7	30	F	0	0.00	8.50	33.90	0.1000	Hyperthyroidism
8	48	F	#?	0.00	1.01	0.50	4.9000	Hypothyroidism
9	42	F	0	0.00	8.30	3.70	0.1400	Hyperthyroidism
10	30	M	6.7	13.90	0.00	0.00	0.3600	Hyperthyroidism
11	69	F	2.3	0.45	0.39	0.80	0.4800	Hypothyroidism
12	30	M	2.6	217.30	2.10	7.60	0.1000	Hypothyroidism
13	30	M	11.1	52.10	0.00	0.00	0.1800	Hyperthyroidism
14	58	F	NaN	0.00	2.10	1.60	1.2000	Hyperthyroidism

Figure 1: Sample of the thyroid dataset

The data collection was monitored by an endocrinologist consultant Dr Gesiye E.L. Bozimo. The dataset consists of 716 instances with 8 features. The target class (diagnosis feature) comprise Hypothyroidism, Hyperthyroidism and Euthyroidism (normal functioning of the thyroid gland). The condition can be hyperthyroidism if the laboratory outcome of T3, T4, FT3, and FT4 levels is above the NORMAL RANGES and the TSH parameter is below the NORMAL RANGE. Hypothyroidism condition is confirmed if the T3, T4, FT3, and FT4 parameters are below the NORMAL RANGES or the TSH parameter exceeds the NORMAL LIMIT. The Euthyroidism is ascertained from the sample parameters when the T3, T4, FT3, FT4 and TSH parameters all fall in the NORMAL RANGE. Table 1 illustrates the parameters which are used to determine hypothyroid, hyperthyroid or euthyroid conditions.

Table 1. Parameters for Determining Thyroid Conditions

T3/FT3	T4/FT4	TSH	INTERPRETATION
High	High	Low	Hyperthyroidism
Low	Low	High	Hypothyroidism
Normal	Normal	Normal	Euthyroidism

3.2 Data Pre-processing

The manually collected data is possible to consist of anomalies like missing values or empty cells. Specifically, the dataset involved categorical labels such as F or M, and hyperthyroidism or hypothyroidism. The features comprised of categorical labels are "Gender", and "Diagnosis". Thus, to use the dataset for experimental purposes, the categorical labels were required to be encoded to numeric values. To achieve this, the pandas and NumPy libraries were utilized. The libraries assigned integer values to the categorical labels. Data points that contained empty or missing values were dropped. Also, the dataset feature named “Diagnosis” was renamed to 'Outcome' where the labels Euthyroidism, Hypothyroidism, and Hyperthyroidism were assigned to 0, 1 and 2 respectively. The "Gender" feature was dropped because it was not needed to determine any of the labels in the Outcome variable.

3.3 Feature Selection

The Randomized Logistic Regression Feature Extraction was employed to select features in the dataset utilized in the study. It is a wrapper feature selection approach that leverages logistic regression and randomization to choose suitable features. It arbitrarily chooses a subsection of data features using logistic regression. This procedure is recursively performed and certain features which have been constantly elected as the most suitable features are set aside. This method can be built using any ML algorithm and can engage any evaluation metrics like recall, accuracy etc. for model assessment. The main advantage of the randomized logistic regression approach is its ability to perform well even on large and complex datasets. Additionally, it gives room for the selection of poorly correlated features while still enhancing the functioning of the model. Although, this feature selection method is time-consuming when it comes to training and assessing several classifiers, yet provides great performance results compared to several other feature selection techniques in existence. The aforementioned benefits made the randomized logistic regression method for feature selection more suitable for this study.

3.4 Classification Models

ML classification algorithms are of many types. Determining which type of classifier to use depends on the type of problem you want to solve. In this research, the problem presented to solve is a multi-class problem (three-class problem). The thyroid dataset collected in Nigeria is of hypothyroid, hyperthyroid and euthyroid (normal) conditions, which necessitate using multiclass classifiers to achieve the classifications. The study utilized two learning models: multi-class SVM and polynomial features random forest algorithms. The main objective for choosing two classifiers is to compare their classification performance on a single collected thyroid dataset and choose the best-performing model. These classifiers are briefly explained in the following subsections.

3.4.1 Multiclass SVM

SVM is a supervised learning model often used for both regression and classification [21]. The principle behind the functioning of SVM is the hyperplane, a line or decision boundary, that helps to differentiate n-features according to input labels ensuring accurate prediction of data in the future. The algorithm separates all data instances in a feature space into two or more classes. SVM treats all data objects individually in feature space such that the object belongs to a binary class (0 and 1) or multiclass (Euthyroidism (0), Hypothyroidism (1), or Hyperthyroidism (2) {Formatting Citation}. The algorithm is considered in this study due to its ability to handle computational complexities and high dimensional data because of its kernel trick function. In the past, the traditional SVM was developed to perform binary classification problems. However, researchers have developed techniques like One-to-One Class, One-to-Rest Classes, Directed Acyclic Graph, Multi-class Objective function, and divide and conquer support vector machine (DCSVM), to solve multi-class problems. Therefore, the One-to-One multi-class SVM approach is deemed suitable to handle the multiclass classification problem in this study.

3.4.2 Polynomial Features Random Forest Classifier

The Random Forest is a supervised machine learning algorithm in which the forest it builds is an ensemble of decision trees that are usually trained with the bagging method, which combines many classifiers to provide solutions to complex problems. The difference between the decision tree algorithm and the random forest algorithm is that the random forest establishes root nodes. Segregating nodes is done randomly and the outcome is based on the predictions of the decision trees. It predicts by taking the average or mean output from various trees [24]. The idea of the random forest using the bagging method is that a combination of learning models increases the overall result. The random forest can be used for classification and regression problems, forming most current machine learning systems [25]. The Random Forest algorithm is built on the mathematical probability of random

variables with standard deviation and mean. To enhance the performance of the random forest algorithm, this study employed the Polynomial Features Transform (PFT). PFT generates new features to enhance the correlation of the dataset features for better performance. Moreover, transforming the data input features to an exponent can help to expose better the important relationships between input features and the target output. For instance, if a dataset had two input features say $D_2 = [x_1, x_2]$, then a polynomial feature would be the addition of three new features (columns) with a bias of 1 where values were calculated by $1, x_1, x_2, x_1x_2, x_1^2, x_2^2$. This process can be repeated for each input feature in the dataset, thus creating a transformed version of each feature. The degree for the polynomial features transforms increases the number of dataset features and caution has to be taken in choosing the polynomial degree as more features may result in more overfitting and in turn, worsen the performance [20]. Thus, the polynomial features transformation was applied to the input thyroid dataset to enhance its performance in the identification of the relationship involving the input and the labels. The architectural diagram describing the entire methodology is depicted in Figure 2.

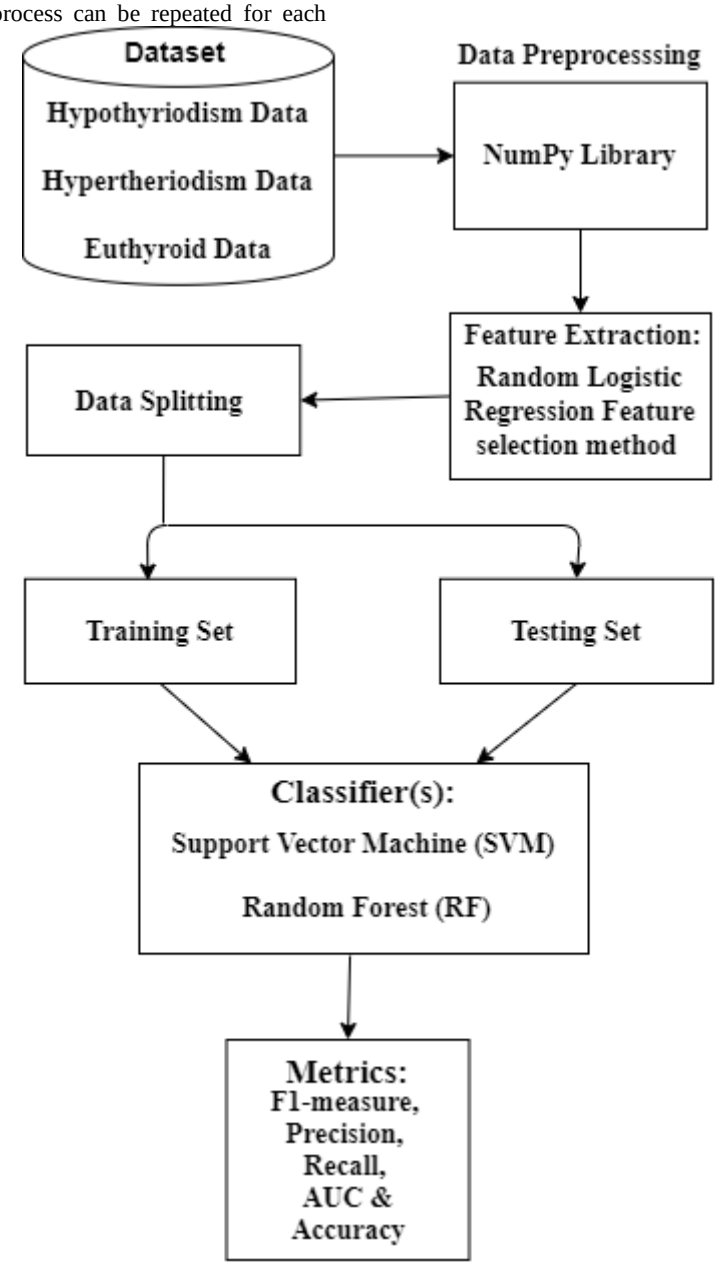

Figure 2: The proposed model architecture

4. EXPERIMENTAL SETUP

This section describes the empirical implementation steps engaged to develop the model aimed at predicting hypothyroidism and hyperthyroidism conditions. First, the indigenous thyroid dataset was manually collected using Microsoft Excel and saved as a .csv file to be used in Python's pandas and NumPy dependencies for preprocessing. Splitting of the preprocessed data into train and test sets was achieved with the Scikit-learn library. The train set was utilized to teach the Multi-class SVM and Polynomial Feature Random Forest algorithms. The test set was utilized to validate the classifiers for predicting euthyroidism, hypothyroidism and hyperthyroidism conditions. The minimum hardware requirements expected for the implementation of this model

are; a personal computer (PC) with a 32-bit Windows 7 operating system or above, x64-based processor, 4.00GB RAM and Intel Pentium (R) Dual-Core CPU T4500 @ 2.30GHz. The software platforms used for the implementation of this model include the Microsoft Excel 2016 worksheet for data collection. Google Collaborator simply called Collab was used to host the Jupiter Notebook via Python 3 programming environment online for model development, visualization of the dataset, and training of algorithms. The evaluation of the effectiveness of the model was carried out using several metrics described in the next section.

4.1 Performance Evaluation Measures

The performance evaluation metrics play an important role in assessing machine learning models. To ascertain the effectiveness of classification algorithms, one metric might not be satisfactory. Hence, this study considered the following metrics to measure the classifiers' performance. These include Accuracy, Precision F1-score, Sensitivity, and graphical performance known as the AUC-ROC curve. The metrics are clearly described in the subsequent sections.

4.1.1 Accuracy

The Accuracy gives the percentage ratio of the rightly predicted outcome over the total number of observations. It is a standard classification metric in machine learning and artificial intelligence that assesses models' performance by giving the percentage of the correct predictions [13]. Thus, it is deemed suitable in this study to ascertain the number of correctly predicted results out of the whole dataset. Moreover, accuracy metrics can easily be implemented for both binary and multi-class classification problems. Consequently, this metric helps to know how accurately the model can classify the three thyroid cases from the thyroid data sample. Accuracy can be obtained mathematically by:

$$\text{Accy} = \frac{\text{Num of Correct pred}}{\text{Total num o f pred}} \quad (1)$$

Where; Accy = Accuracy, Num = Number, and Pred = Prediction(s).

4.1.2 Sensitivity (Recall)

The ratio of correctly predicted outcomes to the sum of correctly predicted and misclassified outcomes is determined by recall. The sensitivity or recall metric is opined as one of the important metrics to apply in medical research [20]. It is obtained using the mathematical relationship:

$$\mathbf{R} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

Where; R = Recall (Sensitivity), TP = True Positive, and FN = False Negative.

4.1.3 Precision

The measure of the correctly predicted data to the sum of the correctly predicted and wrongly predicted positive data is ascertained by the precision metric. It describes how many times the positive prediction is positive. This metric was used in this study to measure the patients that are correctly identified as having hypothyroidism, or hyperthyroidism cases from the thyroid sample dataset. Precision matric is expressed mathematically as:

$$\mathbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

Where; FP = False Positive.

4.1.4 F1-Measure

The F1-Measure provides the ratio of the product of precision and sensitivity to the sum of precision and sensitivity of the learning model [12]. F1-Measure is denoted mathematically as:

$$\mathbf{F1} = \frac{\text{Precison} \times \text{Sensitivity}}{\text{Precision} + \text{sensitivity}} \quad (4)$$

4.1.5 The AUC ROC Curve Metric

ROC is a probabilistic graph upon which the AUC indicates the actual accuracy level at which a model can separate data at various thresholds. An increase in threshold yields fewer positive classes as it decreases TP and FP. The AUC-ROC metric is also deemed suitable in this study as it works fine on an imbalanced dataset [15]. The results of the various evaluation metrics are presented in the next section.

5. RESULTS AND DISCUSSION

The experimental performance results obtained in this study are presented and discussed in this section. Confusion matrix and the AUC ROC curve were employed to comprehend the best-performing algorithm. The experimental results are also further contrasted with those of previous research. The test's findings are detailed in the various sections and subsections.

5.1 Results of the SVM Algorithm

In this section, the confusion matrix and the AUC-ROC curve for the SVM model are presented. Also included in the section is the tabular data as well as the graphical visualization of the SVM performance.

5.1.1 Confusion Matrix for the SVM Classifier

The graphical visualization of the performance of the SVM model in terms of the confusion matrix is depicted in Figure 3.

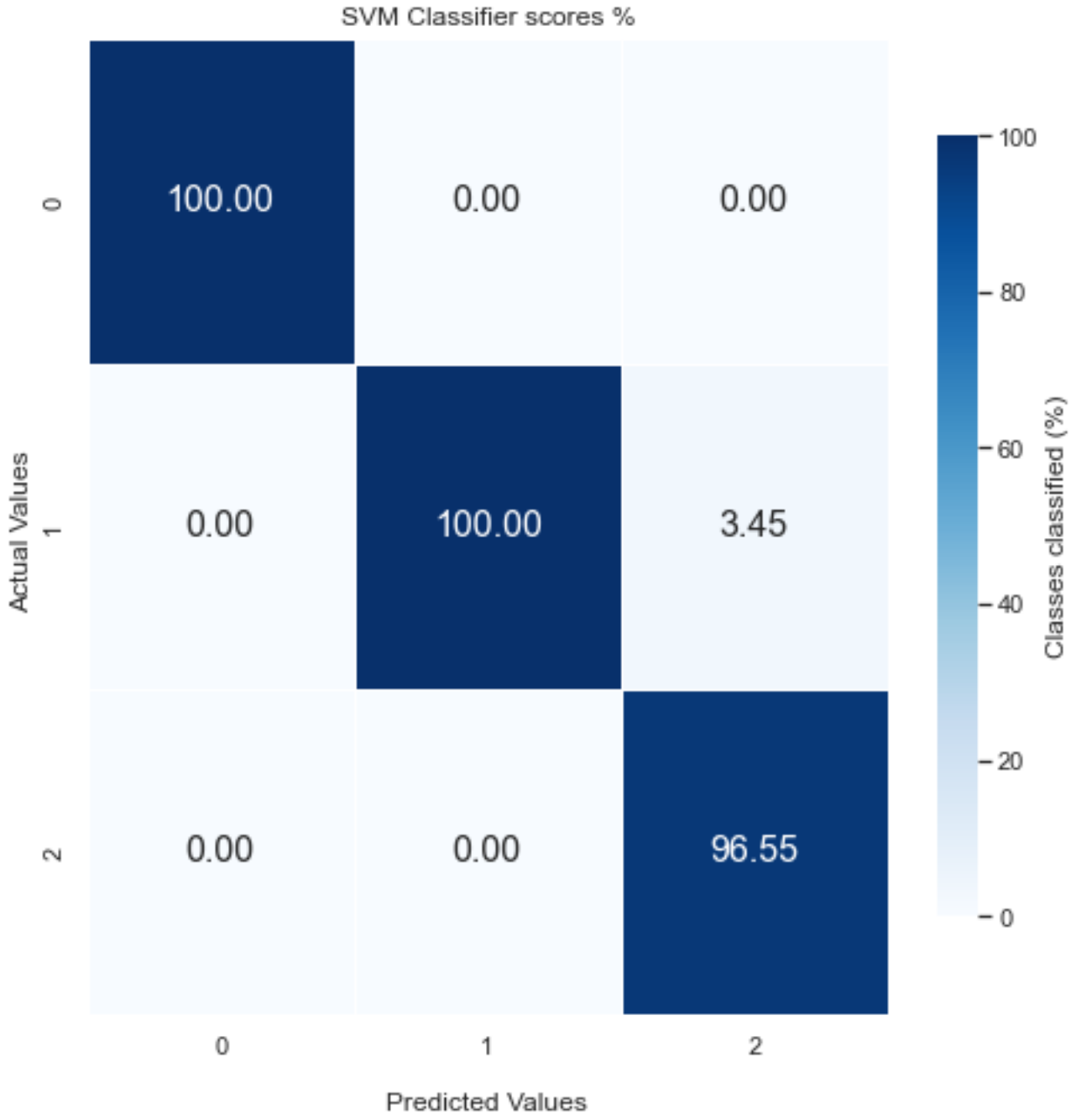

Figure 3: The SVM Confusion Matrix

Figure 3 shows that the SVM model was able to correctly learn from the dataset and classify whether a patient is hypothyroid, hyperthyroid or euthyroid (free from thyroid disease). The dataset was encoded such that class 0, 1, and 2 represents euthyroidism, hypothyroidism and hyperthyroidism accordingly. This shows that the SVM model classified class 0 and class 1 without missing any data instance. However, class 2 was slightly misclassified with about 3.45 per cent. Overall, the SVM classifier performed well on the dataset.

The following subsection presents the performance of the SVM model using the AUC-ROC curve metric.

5.1.2 ROC and AUC of the SVM Classifier

ROC curve and AUC were also used to further validate the performance of the SVM classifier. The AUC and ROC curve demonstrates how good or poor the utilized classifiers learnt from the Multi-dimensional thyroid dataset used in the study. Figure 4 presents the performance of the support vector machine classifier using the ROC curve.

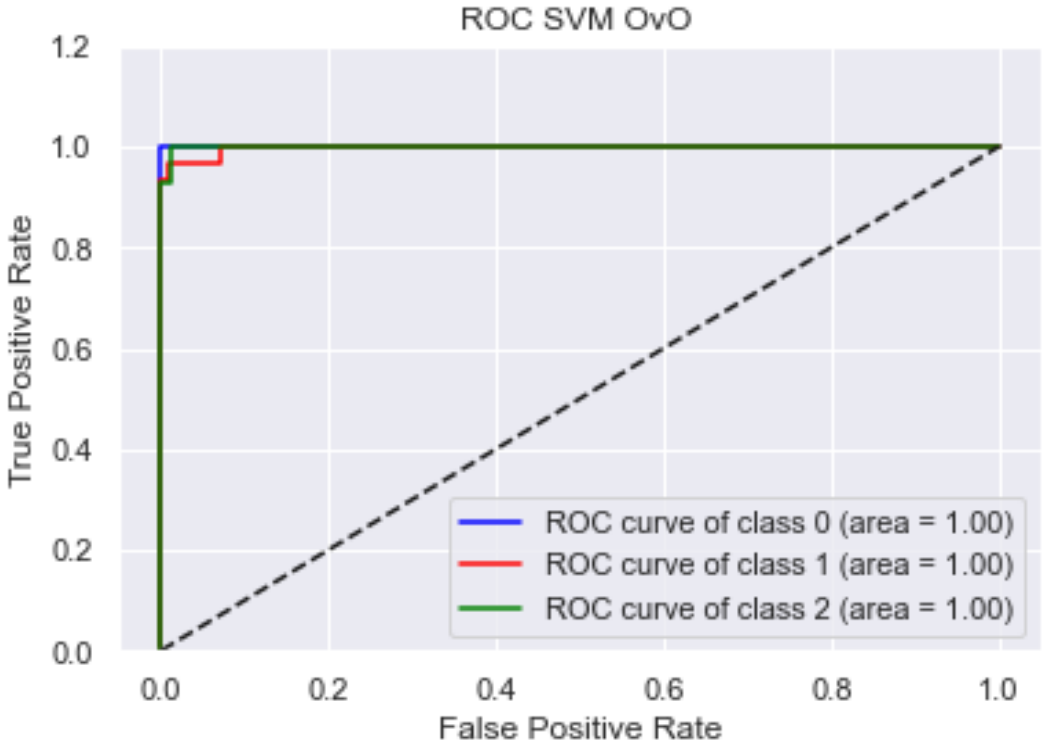

Figure 4: The ROC Curve for the SVM Classifier

As shown in Figure 4, the support vector machine classifier was also evaluated using the ROC curve metric. When the ROC curve fits closer to the top left angle of the plot it yields a superior performance at classifying data into various class labels. Thus, the SVM model categorizes the various classes of thyroid conditions. This reveals the strength of the model's separability using the ROC curve metric. To quantify the ROC curve plot, AUC was simultaneously implemented. The AUC showing the micro and macro average performance of the SVM classifier is presented in Figure 5.

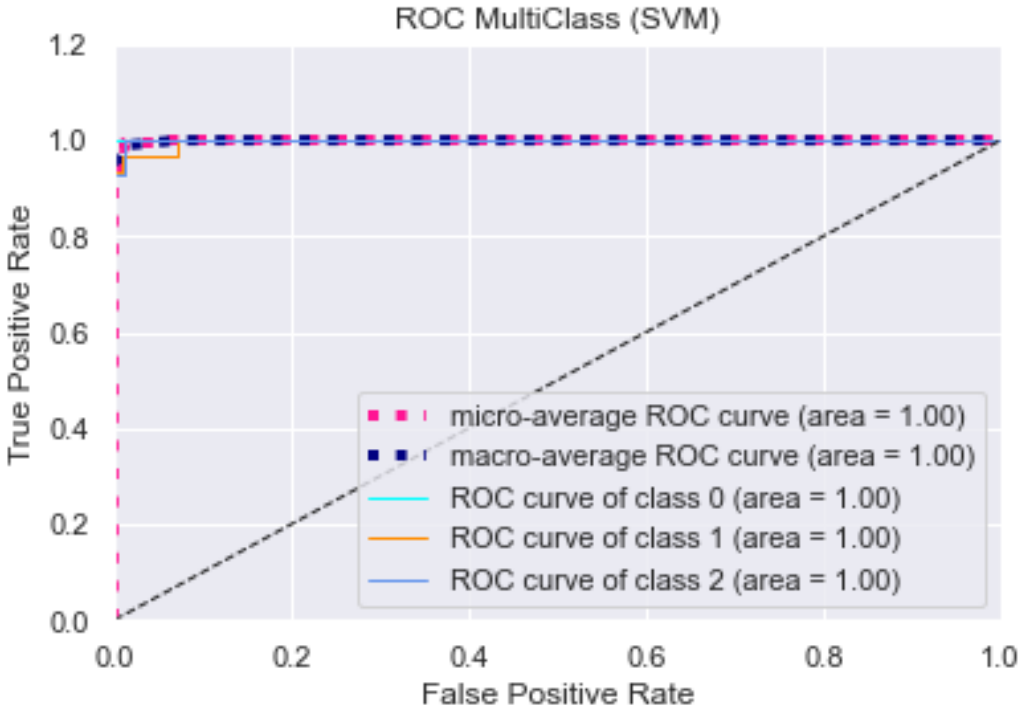

Figure 5: The AUC-ROC Curve for the SVM Classifier

Figure 5 clearly shows the micro and macro average of best-fit at the top left corner of the plot which indicates the performance of the SVM model using the AUC-ROC curve metric. In AUC, the nearer a model performance is to 1, the

better the model is in terms of classifying data points. It should be noted that, if the AUC model is less or equal to 0.5, there is more likelihood of the model to make random classifications. Numerically, the summary of the performance of the SVM algorithm is presented in Table 2.

Table 2. SVM Performance Summary

Performance of the SVM Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	98.60
Precision	99.00
Recall or Sensitivity	98.00
F1-score	98.00
AUC Score	99.88

It is obvious in Table 2 that the SVM model attained an overall accuracy of 98.60 per cent, with an average score of 98.00 for F1-score and Recall while precision achieved 99.00 per cent. Figure 6 depicts the SVM model's performance results graphically.

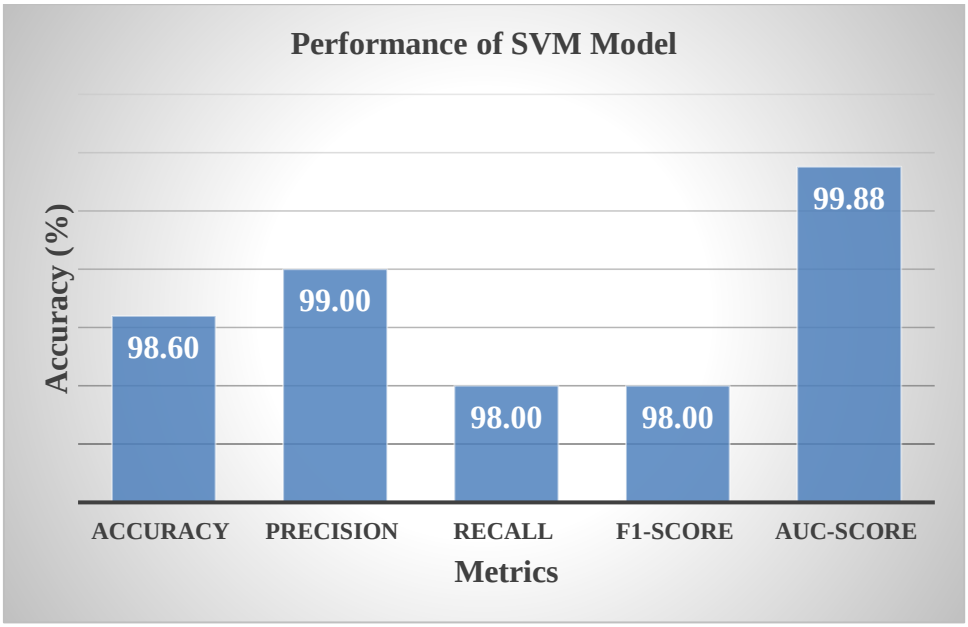

Figure 6: The SVM Model Performance

5.2 The Random Forest Confusion Matrix Performance

This section presents the confusion matrix and the AUC-ROC curve for the RF model. Also included in the section is the tabular data as well as the graphical visualization of the RF performance.

5.2.1 Confusion Matrix for the RF Classifier

The confusion matrix as well as the classification scores of the Polynomial Features-Random Forest classifier are presented in Figure 7.

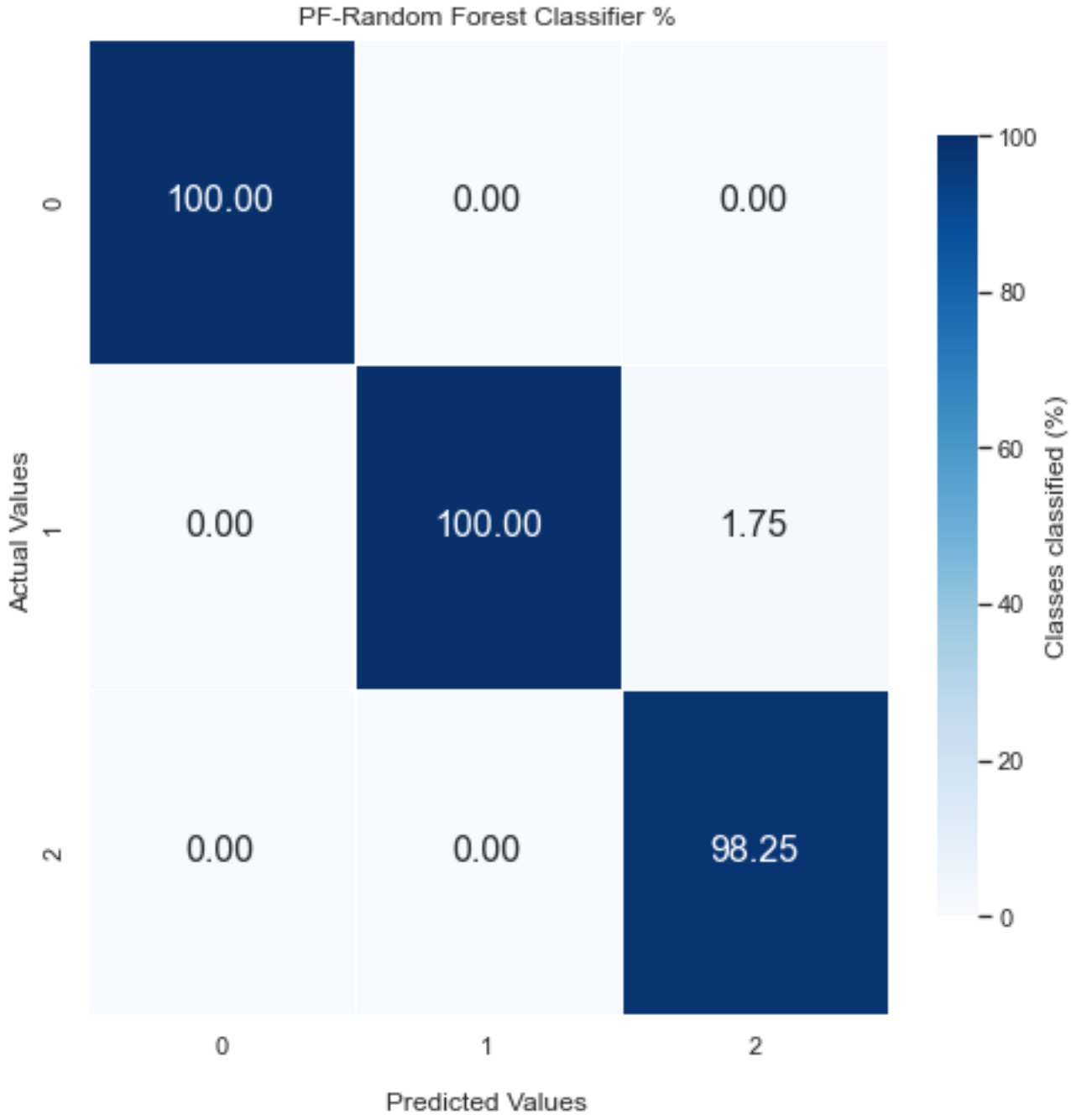

Figure 7: The Random Forest Confusion Matrix

It is clear in Figure 7 that the RF Classifier perfectly learnt from the dataset and categorized it into the thyroid conditions: euthyroid (free from thyroid disease), hypothyroid or hyperthyroid. The RF classifier grouped class 0 and class 1 without any misclassification. It however faintly misclassified class 2 with only 1.75 per cent. It is noteworthy that the RF classifier understood the dataset at very high accuracy. In the next subsection, the performance of the RF model using the AUC-ROC curve is presented.

5.2.2 ROC and AUC of Random Forest Classifier

The ROC-AUC evaluation metric was also employed to assess the performance of the RF classifier. Experimental results show that the model classified the thyroid conditions perfectly. The result of the polynomial-random forest classifier using the ROC curve is shown in Figure 8.

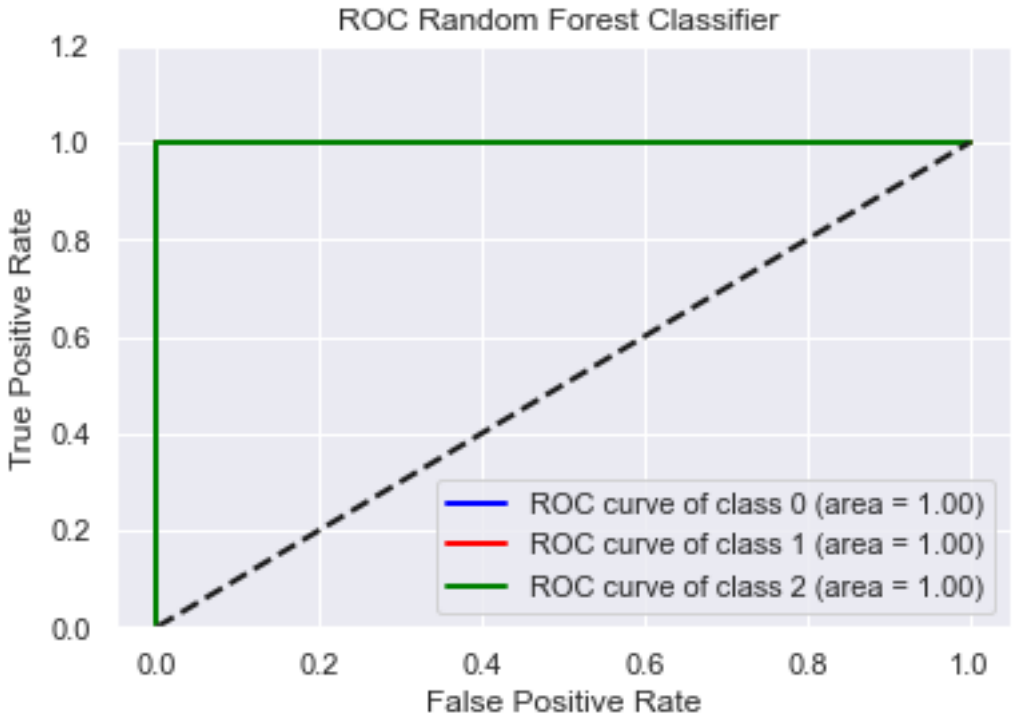

Figure 8: The ROC Curve for the RF Classifier

The classification of the thyroid conditions into classes 0, 1, and 2 for Euthyroidism, Hypothyroidism and Hyperthyroidism respectively was successfully evaluated using the ROC curve as earlier shown in Figure 8. The AUC was computed concurrently during implementation to quantify the ROC curve plot. This indicates the ROC curve's degree of separability. The closer the AUC is to 1, the better the model is at classifying data points. It should be highlighted that if the AUC is less than or equal to 0.5, the model is more likely to make accurate classifications. Figure 9 depicts the micro and macro average performance measure of the RF using the AUC evaluation metric.

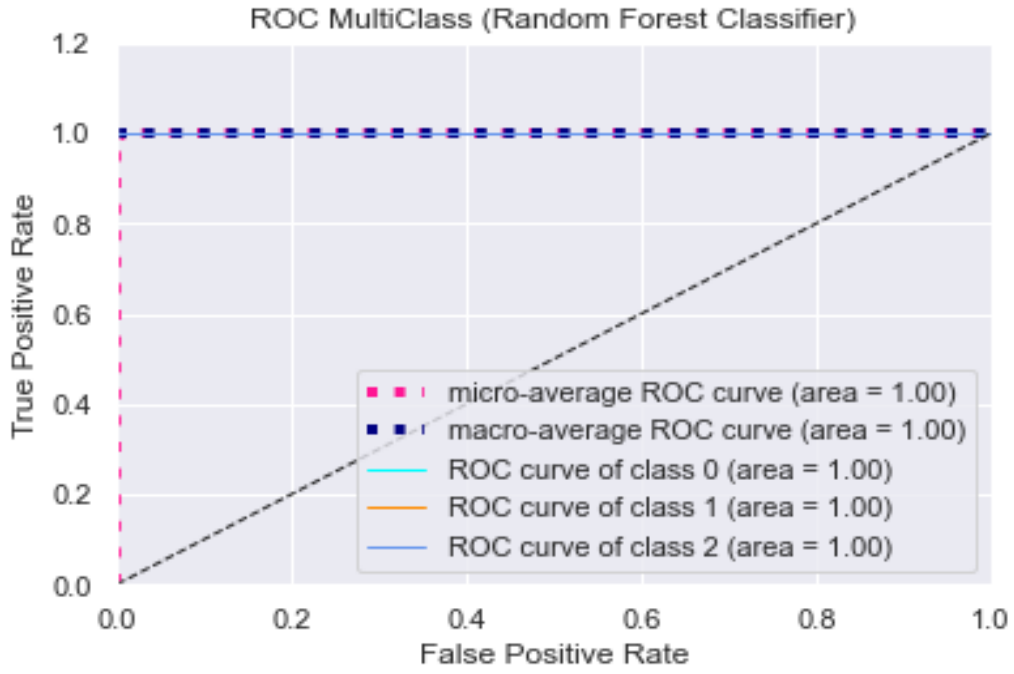

Figure 9: The AUC-ROC Curve for the RF Classifier

In Figure 9, it is clear that the AUC's micro and macro average lines of best fit embrace one another at the top left corner of the plot. This is approximately 100 per cent average which means that the RF model performed well at classifying the thyroid conditions. The summary of the performance of the RF algorithm using various metrics is presented in Table 3.

Table 3. The RF Performance Summary

Performance of the RF Model	
Evaluation Metrics	**Performance Results (%)**
Accuracy	99.30
Precision	99.00
Recall or Sensitivity	99.00
F1-score	99.00
AUC Score	100.00

As can be seen in Table 3, the RF model achieved an overall average accuracy of 99.30 per cent, with an average score of 99.00 for F1-score, precision and Recall. The performance outcomes of the RF model are visually shown in Figure 10.

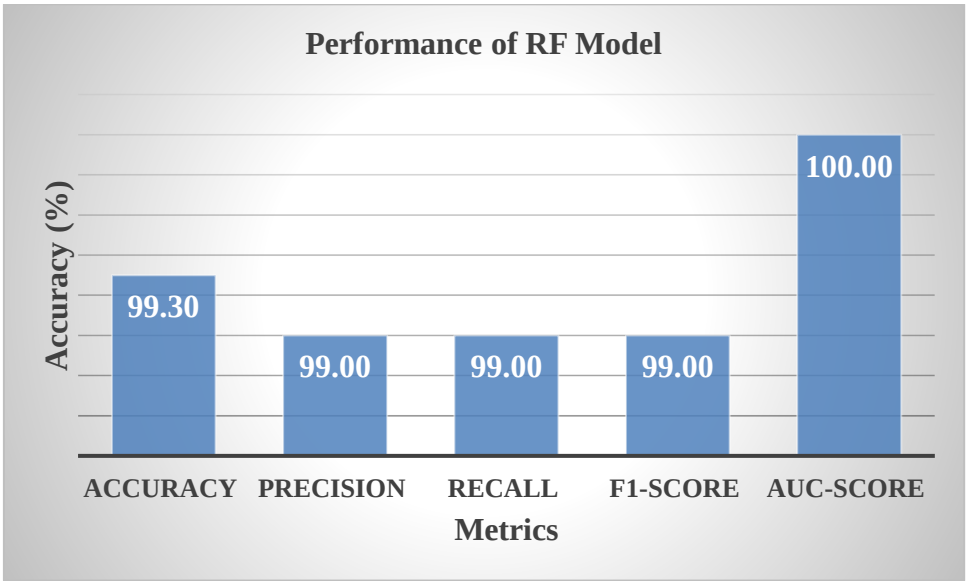

Figure 10: Performance of the RF Model

5.3 Result Comparison of the Utilized Models

Choosing the best learning model for a certain dataset is a difficult task and must be done experimentally. As a result, current studies employ two or more machine learning algorithms and closely monitor them to discover a single algorithm capable of producing the best results [18]. The selection of ML algorithms employed in this work was inspired by the literature review. As a result, the model was built using SVM and RF. Thus, comparative research is necessary to determine the best-performing ML model among the two utilized models. Consequently, the findings of the comparative performance of both the SVM and the RF are summarized in Table 4.

Table 4. Results Summary of the Utilized Classifiers

Evaluation Metrics	Classifiers	
	One-Vs-One SVM (%)	**Polynomial RF (%)**
Accuracy	98.60	99.30
Precision	99.00	99.00
Recall	98.00	99.00
F1-score	98.00	99.00
AUC-score	99.88	100.00

All of the ML models used in the study, as shown in Table 4, had good performance accuracy, ranging from 98 per cent and above. This performance result is an indication that all the ML models used in the study comprehended the thyroid dataset relevant for the classification of the various kinds of thyroid conditions under consideration. The findings further demonstrate that the RF algorithm outperformed the SVM model having achieved the best performance accuracy of 99.30 per cent. The performance of the utilized algorithms is visualized in Figure 11.

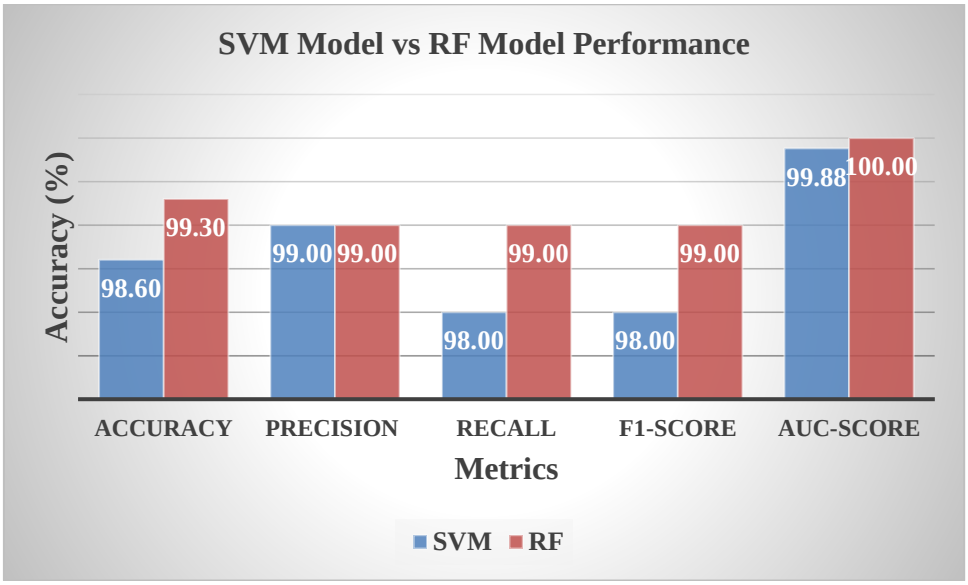

Figure 11: Comparison of the Utilized ML Algorithms

It can be observed in Figure 11 that the AUC metric yielded the best performance outcome. This affirms the assertion that the AUC ROC curve performs great on imbalance data which assesses a model's performance based on the different thresholds of the data instances. Whereas, the role of accuracy in measuring the exact proportion of correctly classified instances out of the total number of instances in the dataset makes it most preferred for this study [9].

5.4 Results Comparisons with the Existing Models

The study's performance results are then compared to five state-of-the-art models. This is to validate the enhanced model's success in classifying the thyroid conditions: hyperthyroidism, hypothyroidism or euthyroidism. The contrasted state-of-the-art studies are chosen based on their relevance to this study and the improved model's capacity to replicate or use their models, as well as some parameter tuning such as cross-validation (train and test set) and evaluation metrics. It was noted that the accuracy result of this study, which is 99.30%, beat the performance of all five (5) existing models. Table 5 provides a comparative summary of the performance of the improved model and the existing state-of-the-art models for the classification of thyroid conditions.

Table 5. Comparative Analysis of the Improved Model with Existing Model(s)

Author	Accuracy Results (%)
Chandan et al. [9]	95.38
Salman & Sonuc [16]	98.93
Shama et al. [10]	91.42
Singh et al. [19]	98.98
Islam et al. [20]	94.79
Proposed Model	99.30

As can be seen in Table 5, the performance of the developed thyroid classification model is provided in the last row. The model achieved 99.30 per cent accuracy, which is higher than the existing models. This demonstrates the significance of the improved model and its efficacy using the indigenous dataset. The graphical or pictorial representation of Table 5 is shown in Figure 12.

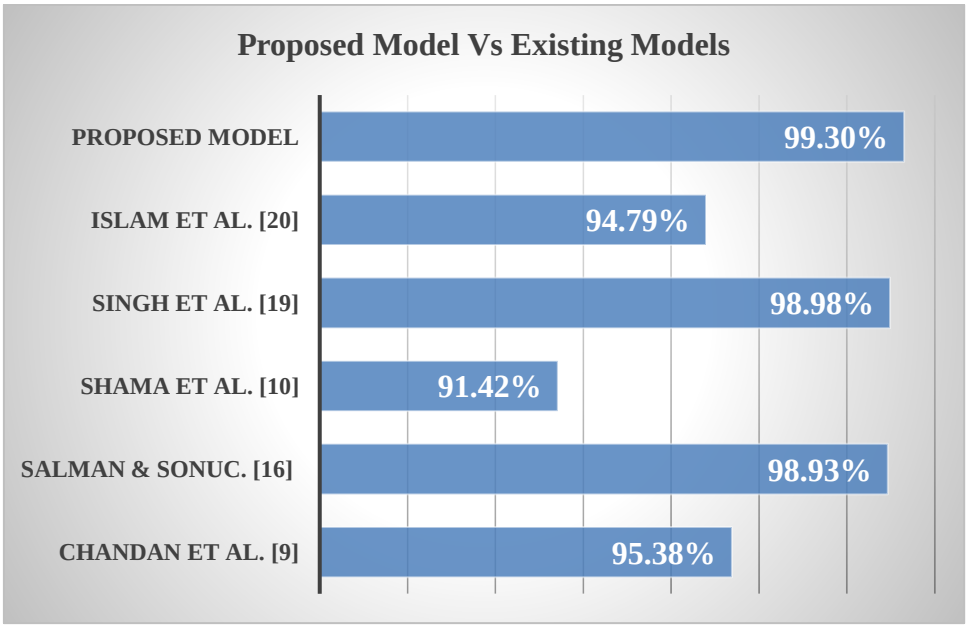

Figure 12: Graphical View of the Improved Model and the Existing Model(s)

This comparative analysis demonstrates how well the proposed thyroid scheme understood the local dataset collected from FMC, Yenagoa, Nigeria. The proposed model's effectiveness can be attributed to its capacity to address the flaws in the existing models such as the attention given to feature selection and data cleansing, which are occasionally overlooked in earlier studies. The richness of the dataset to include the parameters of hyperthyroidism, hypothyroidism, and euthyroidism conditions is one important enhancement over the existing works.

6. CONCLUSION

Efforts have been made in this study to develop a thyroid disease prediction model. For experimental purposes, the utilized thyroid sample data parameters were collected from the Federal Medical Center Yenagoa, Bayelsa State of Nigeria. Preprocessing of the data was done using the NumPy library in Python programming language. Various machine learning techniques such as One-Verse-one Multiclass SVM and Polynomial Features Random Forest Algorithms were trained using the locally collected data. To assess the performance efficacy of the utilized ML models and the dataset, several evaluation metrics were employed. The metrics employed include Accuracy, Precision, and Recall including F1-score and AUC-ROC curve. Experimental results revealed that the Polynomial Features Random Forest classifier achieved the highest classification accuracy of 99.30% than the One-Verse-One SVM classifier which recorded 98.60% accuracy. Due to the experimental results of the classifiers, this study recommends the Polynomial Features Random Forest Algorithm to researchers for the prediction or classification of thyroid diseases. Future research should consider generating more thyroid datasets or any other endocrine disease feature sets like diabetes in Nigeria. Hence, build a more robust thyroid disease prediction model to enhance quick diagnosis and treatment to support endocrinologists in Nigeria.

7. REFERENCES

[1] A. O. Ogbera, O. Fasanmade, and O. Adediran, “Pattern of Thyroid Disorders in the Southwestern Region of Nigeria,” Ethn. Dis., vol. 17, 2007.

[2] Anthonia and Sonny, “Epidemiology of thyroid diseases in Africa.”

[3] O.-O. N. Fidelis and A.-P. J., “A Multigene Genetic Programming Model for Thyroid Disorder Detection,” 2015.

[4] M. R. Obeidavi, A. L. I. Rafiee, and O. Mahdiyar, “Diagnosing Thyroid Disease by Neural Networks,” Biomed. Pharmacol. J., vol. 10, no. 2, pp. 509–524, 2017, doi: doi.org/10.13005/bpj/1137.

[5] I. Mofek and Z. Bozkuş, “Use of Machine Learning Techniques for Diagnosis of Thyroid Gland Disorder,” 2016.

[6] S. Godara and S. Kumar, “Prediction of Thyroid Disease Using Machine Learning Techniques,” Int. J. Electron. Eng. (ISSN, vol. 10, no. 2, pp. 787–793, 2018, [Online]. Available: www.csjournals.com%0APrediction

[7] A. Tyagi, R. Mehra, and A. Saxena, “Interactive Thyroid Disease Prediction System Using Machine Learning Technique,” 2018 Fifth Int. Conf. Parallel, Distrib. Grid Comput., no. March, pp. 689–693, 2019, doi: 10.1109/PDGC.2018.8745910.

[8] D. S. Reddy, O. S. Vaishnavi, J. Vidya, K. S. Sharan, and R. Subramanyam, “Literature Survey,” Int. J. Adv. Sci. Technol., vol. 29, no. 5, pp. 4752–4761, 2020.

[9] R. Chandan, C. Vasan, M. S. Chethan, and H. S. Devikarani, “Thyroid Detection using Machine Learning,” Int. J. Eng. Appl. Sci. Technol., vol. 5, no. 9, pp. 173–177, 2021, [Online]. Available: http://www.ijeast.com

[10] A. Shama, (B. H., A. Adhikary, A. Uddin, and M. A. Hossain, “Prediction of Hypothyroidism and Hyperthyroidism Using Machine Learning Algorithms,” Res. Sq., pp. 1–21, 2022, [Online]. Available: doi.org/10.21203/rs.3.rs-1486798/v2%0ALicense:

[11] N. A. Sajadi, S. Borzouei, H. Mahjub, and M. Farhadian, “Diagnosis of hypothyroidism using a fuzzy rule-based expert system,” Clin. Epidemiol. Glob. Heal., vol. 7, no. 4, pp. 519–524, 2019, doi: 10.1016/j.cegh.2018.11.007.

[12] L. Shalini and M. R. Ghalib, "A Hypothyroidism Prediction using Supervised Algorithm," no. 1, pp. 7285–7288, 2019, doi: 10.35940/ijeat.F7897.109119.

[13] K. Dharmarajan, K. Balasree, A. S. Arunachalam, and K. Abirmai, "Thyroid Disease Classification Using Decision Tree and SVM," Indian J. Public Heal. Res. Dev., vol. 11, no. 03, pp. 224–229, 2020.

[14] M. Mourad et al., "Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis," Sci. Rep., pp. 1–11, 2020, doi: 10.1038/s41598-020-62023-w.

[15] S. Borzouei, H. Mahjub, N. A. Sajadi, and M. Farhadian, "Diagnosing thyroid disorders: Comparison of logistic regression and neural network models," J. Fam. Med. Prim. Care, 2020, doi: 10.4103/jfmpc.jfmpc.

[16] K. A. Salman and E. Sonuc, "The Efficiency of Classification Techniques in Predicting Thyroid Disease," 2021.

[17] V. V. Vadhiraj, A. Simpkin, J. O. Connell, N. S. Ospina, S. Maraka, and D. T. O. Keeffe, "Ultrasound Image Classification of Thyroid Nodules Using Machine Learning Techniques," Medicina (B. Aires)., vol. 57, no. 527, pp. 1–18, 2021, doi: doi.org/10.3390/medicina57060527.

[18] L. Aversano et al., "Thyroid Disease Treatment prediction with machine learning approaches," Procedia Comput. Sci., vol. 192, pp. 1031–1040, 2021, doi: 10.1016/j.procs.2021.08.106.

[19] T. Singh, A. K. Sahu, S. D. Greater, M. P. Sharma, S. Verma, and C. Kumar, "Treatment of thyroid disease through machine learning predictive model," Int. J. Heal. Sci. ISSN, vol. 6, no. July, pp. 3176–3188, 2022, doi: doi.org/10.53730/ijhs.v6nS8.12813 Treatment.

[20] S. S. Islam, S. Haque, M. S. U. Miah, T. Bin, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," PeerJ Comput. Sci., pp. 1–35, 2022, doi: 10.7717/peerj-cs.898.

[21] M. Ramya and P. V. S. Kumar, "Prediction and Providing Medication for Thyroid Disease using Machine Learning Technique (SVM)," Turkish J. Comput. Math. Educ., vol. 11, pp. 1099–1107, 2020.

[22] R. Bridgelall, "Tutorial on Support Vector Machines," no. February 2022, doi 10.20944/preprints202201. 0232.v1.

[23] O. Mbaabu and Onesmus, "Introduction to Random Forest in Machine Learning," pp. 1–20, 2020.

[24] N. Donges, "What Is Random Forest_ A Complete Guide _ Built In."