A Comparative Study of Thyroid Dysfunction Prediction Models using Machine Learning

Ajayi A. F.
Department of Physiology
Faculty of Basic Medical
Sciences, Ladoke Akintola
University of Technology,
Ogbomoso, Nigeria.
aajayi22@lautech.edu.ng

Akindele A. T.
Department of Computer Science
Open and Distance Learning Center
Ladoke Akintola University of
of Technology, Ogbomoso,
Nigeria.
atakindele@lautech.edu.ng

Adepoju O.
Department of Physiology
Faculty of Basic Medical
Sciences, Ladoke Akintola
University of Technology
Ogbomoso, Nigeria.
helenopetunde@gmail.com

Wahab O.
Department of Computer Science
and Engineering, Obafemi Awolowo
University, Ile-Ife, Nigeria.

Adebayo O. I.
Department of Physiology, Faculty of Basic
Sciences, Ladoke Akintola University of
Technology, Ogbomoso, Nigeria

Corresponding email: atakindele@lautech.edu.ng

ABSTRACT

The prevalence of Thyroid Dysfunction (TD) is now alarming worldwide, particularly in Africa due to environmental and increased poor nutritional factors. The treatment of TD is valid only when it is detected and diagnosed accurately at early stages. The diagnosis of Thyroid Dysfunction requires experience and sound knowledge to analyze test results, however, the current manual method of interpreting test results in most developing countries is subjective and error-prone, however, the best scenario is to predict and detect the disease as early as possible. Data Mining techniques have been explored in the literature to automatically predict diseases based on patients' data in hospitals and clinics however the features used were less than adequate in modern methods context. Hence this work will explore the use of Machine Learning to evaluate twelve (12) elements or features of patient blood test data. Machine Learning (ML), a known subset of the field of Artificial Intelligence (AI) employs different statistical, probabilistic, and optimization operating rules that let the computer "learn" from earlier cases and then detect challenging to recognize patterns of event from massive, noisy or compound datasets. The drive for the current research is towards developing an efficient model to predict thyroid dysfunction at the early stages. These models are less expensive to build; thereby making sure that qualitative healthcare is affordable and accessible to the marginalized population in most developing and third world countries. In this research, the datasets used were acquired from the UCI (University of California, Irvine) Public Machine Learning Repository Database which contains three thousand, seven hundred and seventy-four (3774) patients' records. These records include levels of Free Triiodothyronine (FT3), Stimulating Thyroid Hormone (TSH), Triiodothyronine (T3) and Thyroxine (T4) amongst others. The Machine Learning Models used are Gradient Boosting, Decision Tree and Logistic Regression, and their accuracy, precision, and recall values were compared. The best accuracy (0.981), precision (0.827) and recall (0.727) were obtained in the Logistic Regression model. Therefore, integrating the Logistic Regression based model into a real-time Hospital Management System can enable medical experts to use the T3, T4, FTI, TSH levels gotten from blood test results to predict whether the patient has thyroid dysfunction or not.

Keywords: ***Decision Tree, Gradient Boosting, Logistic Regression, Machine Learning, Thyroid Dysfunction, Confusion Matrix, Early Detection.***

I. INTRODUCTION

Dietary deficiency of iodine happens to be the primary determining factor of various pathological events of thyroid in Africa. These events result in a wide range of iodine insufficiency disorders such as; goitres, mental deficiency, and hypothyroidism (Okosieme, 2006). In places with a daily iodine intake of <50 μg goitre is typically endemic, and in a situation where the daily intake of <25 μg occurs, inborn hypothyroidism results. Goitre prevalence in locations with extreme iodine deficiency could be increased by up to 80% (Vanderpump, 2011). Majority of these cases if undetected and diagnosed early, can eventually lead to complications and death. In a world where automation and hospital management system are used in the diagnosis of ailments and diseases, Thyroid dysfunction detection from patient's records via Machine Learning (ML) is also possible.

Iodine deficiency is a major public health problem throughout Africa and is the commonest cause of thyroid disorders in the continent (Tsegaye & Ergete, 2003). The scope of thyroid diseases that are frequently noted in Africa include hypothyroidism, thyrotoxicosis (which could be from hyperthyroidism or non-thyroid causes), thyroid malignancies, and iodine deficiency disorders. While the prevalence of thyroid disorders depends on a large number of factors, of which, the most important include: age, sex, geographic factors, ethnicity (Sulejmanovic et. al, 2019), the populations at most risk have tendency to be remote and live in mountainous areas in southeast Asia, Latin America and Central Africa and their thyroid dysfunctions are often attributed to environmental and nutritional causes (Ogbera & Kuku, 2011).

II. LITERATURE REVIEW

Machine Learning algorithms have been used to achieve significant advancements in disease prediction. This breakthrough is due to its ability to predict the occurrence of diseases through classification techniques that classify the data (patients' records) into predetermined categories.

Machine Learning Algorithms are categorized into three forms (Supervised Learning, Unsupervised Learning, and Reinforcement Learning). There are several models of Machine Learning Algorithms. Some of the popular ones include Decision Tree, Linear Regression, Multilayer Feed Forward Neural Network, Support Vector Machine (SVM), Gradient Boosting Algorithms (GBM, XGBoost, LightGBM, CatBoost), Logistic Regression, Naïve Bayes, amongst others.

Decision Tree

Decision Tree (DT) is a class of Supervised Learning algorithm that is frequently in classifying complications. In a real sense, this model is Classification and Regression Trees (CART), which is one the implementation of Decision Trees, as there are other forms. Its label has a tree-type structure that supplies it with stability and pronounced accuracy. DT employs the use of an easy "if-else rules" to build the trees. DT algorithm makes use of various techniques such as Gini Index, Information Gain, Chi-Square, and depletion in the variable to do a calculated split.

Logistic Regression

Logistic Regression (LR) is a powerful method used in diagnostic analysis. It is crucial mainly in estimating distinct values (Binary values like 0/1, true/false, yes/no) according to a specified group of a variable(s) that is independent. LR interprets the data efficiently by analyzing the relationship between one dependent binary variable and one or more nominal variables that are not independent. Logistic regression projects the possibility of the development of an episode by connecting data to a logit function. Therefore, referred to as logit regression. whereas, it predicts the probability, its output values often located between 0 and 1.

Related Works

Several researchers have explored the use of Machine Learning algorithms in the classification and prediction of diseases. Rajam *et al*. (2016) surveyed the various data mining methods used in diagnosing thyroid dysfunction. These researchers suggested the usage of algorithms such as Naïve Bayes, Decision Tree, backpropagation, and Support Vector Machine.

Lui and Pappas (2015) used an exploratory approach to compare different models in determining the presence of hyperthyroidism or hypothyroidism in previously undiagnosed patients who were presumed healthy. The conclusion was that thyroid dysfunction can be predicted to good accuracy using TSH, FTI, and TT4, using a simple decision tree model. They also recommended:

i. the use of more recent datasets, as the set used was from the mid-1980s;

ii. complete data is needed, to see if other classification methods can perform better than decision trees; and

iii. a more extensive collection of data from an enormous variety of sources should be obtained.

Based on the recommendations of Lui and Pappas (2015), this research employed a structured analytical method (Machine Learning) to design predictive models. New datasets, whose features are more substantial, were used in the prediction of thyroid dysfunction at its early stages via three (3) models designed with different ML algorithms. Their results were after that compared.

III. Methodology

The research methods are grouped into five main steps. Some of the measures have several processes involved. Figure 2 outlines the pipeline (the mode of operation) for building the model used, while Figure 3 shows the flowchart of the processes involved. Dataset used in the research was first acquired from the UCI repository. It was pre-processed by removing the empty features or replacing the new numerical elements with the median. The datasets (the pre-processed data) were grouped into the training dataset and the testing dataset. The dataset for training was employed in training the models, after which the model was tested with testing data to obtain the predicted output.

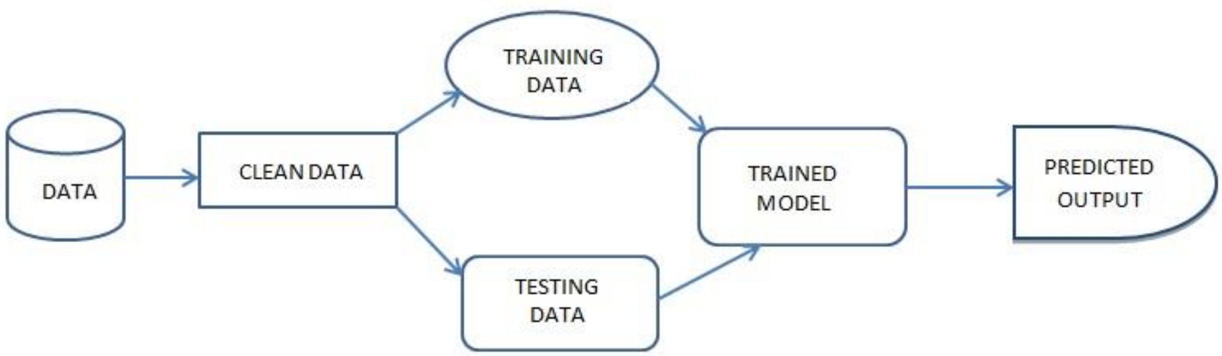

Figure 1: Model Pipeline

Summarily, the main steps are listed thus.

i. Dataset Acquisition

ii. Datasets Cleaning and Pre-processing

iii. Model Development

iv. Model Training

v. Model Testing

a. Dataset Acquisition

The dataset used was obtained from the Machine Learning data Repository of the University of California, Irvine (UCI) (http://archive.ics.uci.edu/ml/datasets/thyroid+disease). The database contains 3774 patients' data with features such as FTI, TSH, TT4, T4U levels, age, health status at the time of test (sick or not sick), gender, and pregnancy condition for women, etc. The dataset mostly contained binary annotations such as the presence of pregnancy, goitre, and other diseases that can affect thyroid stimulation in the body.

b. Dataset Cleaning and Preprocessing

In the dataset, there were some missing data values. Thus, data cleaning to either replace the missing values with median or drop those with missing values took place. As there are also related quantitative variables, a Correlation plot was used to analyze the relationship between them as shown in figure 2. From the correlation plot, it was realized that some features had an insignificant correlation to the target variable. An example is the TGB, referral source, and Iodine-131 treatment. These features were removed to enable the optimum performance of the model.

c. Model Development

In developing this model, python modules (Sci-Kit-Learn, Matplotlib, Numpy, and Pandas) were used. Matplotlib Python library for visualization - to visualize the correlation between the features (age, sex, pregnancy status) and the target (thyroid dysfunction). Numpy for mathematical computation (the mean, median, and standard deviations of the numerical features contained in the dataset). Scikit-learn in building the models (logistic regression, gradient boosting, and decision trees). The type of learning algorithms used is Supervised Learning.

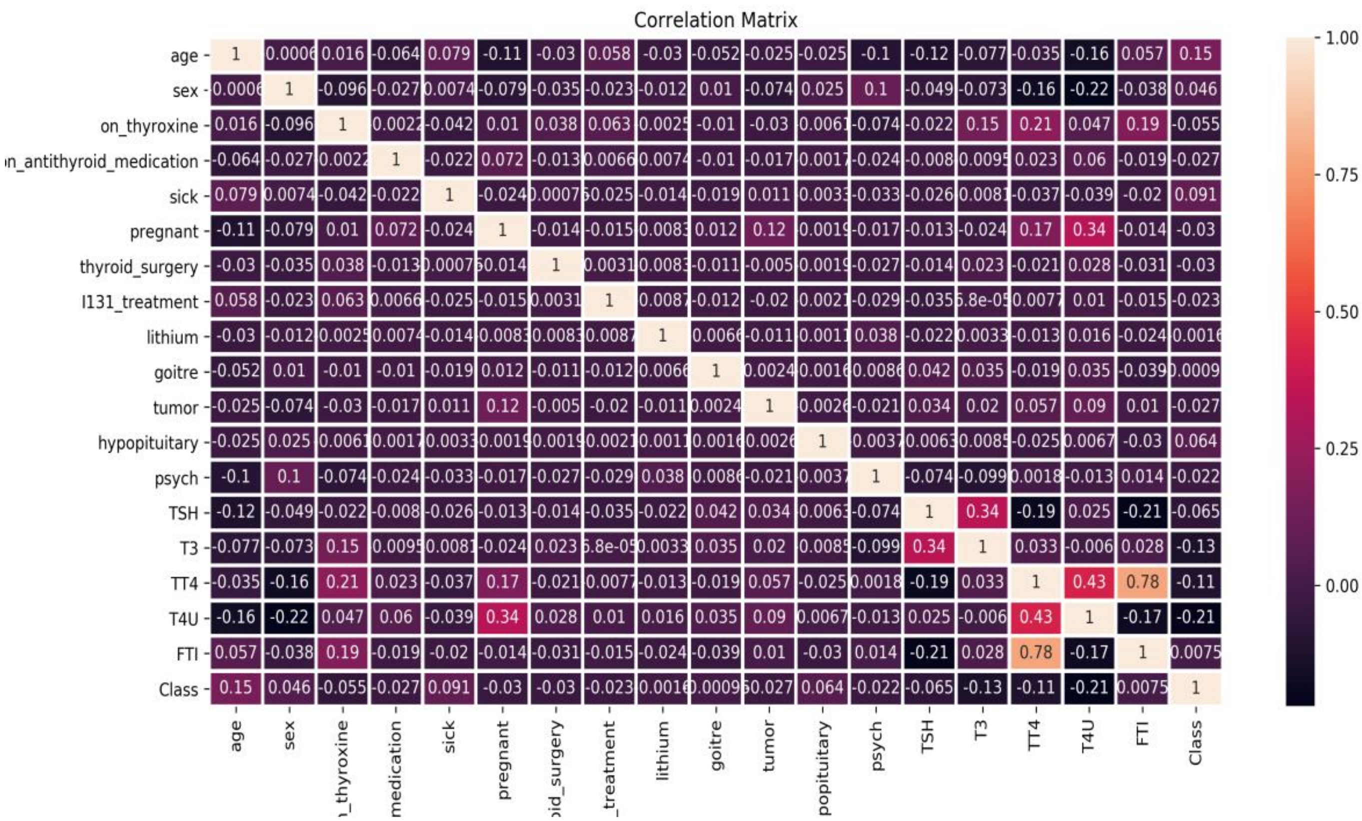

Figure 2: Correlation Plot of Class against features

Table 1: Features used for the models

	ATTRIBUTES	VALUE TYPE
1	Age	Continuous
2	Sick	False, True
3	Sex**	Male, Female
4	Thyroid surgery**	False, True
5	Pregnancy **	False, True
6	Iodine-131 treatment**	False, True
7	query hypothyroid	False, True
8	query hyperthyroid	False, True
9	Lithium **	False, True
10	Goitre**	False, True
11	Hypopituitary**	False, True
12	Psych**	False, True
13	Tumor **	False, True
14	T3 measured**	False, True
15	T3	Continuous
16	FTI measured **	False, True
18	FTI	Continuous
19	TSH measured**	False, True
20	TSH	Continuous
21	T4U	Continuous
22	T4U measured	False, True
23	TT4	Continuous
24	TT4 measured**	False, True

Model Flowchart

The flowchart employed in this research work is shown in figure 3. The flowchart denoted the flow of processes in all steps involved in the modelling, training and testing.

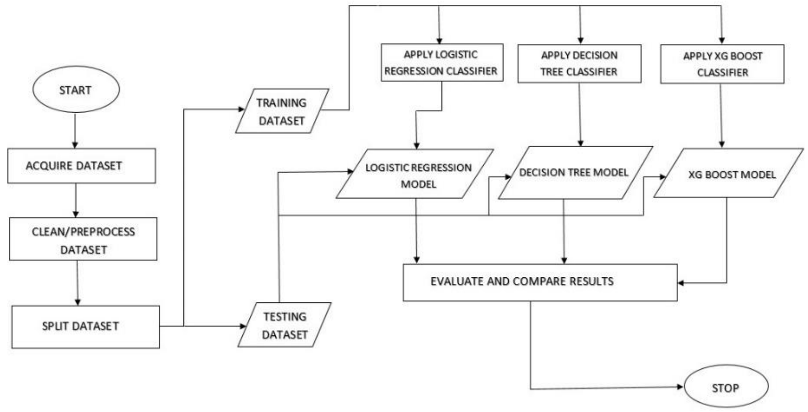

Figure 3: Flowchart of the Machine Learning Modelling and Testing

d. Model Training & Testing

The dataset acquired was split into two parts, 80% of the datasets were used in the training and 20% for testing. The frequently used patient's features in diagnosing thyroid disorders are shown in Table 1. Researchers have selected one or more of these features as inputs variables to the thyroid dysfunction prediction model. While some researchers used four (4) or five (5) components, this research employed the usage of twelve (12) attributes. Less discriminatory features were eliminated based on their correlation plot to thyroid dysfunction, leaving a subset of the original features that still retain sufficient information needed to discriminate well among the classes. The correlation plot in Figure 2 shows the relationship between thyroid dysfunction against each of the features used. Out of the 18 features plotted in the correlation plot, the 13 attributes or features employed in the research are double-asterisked in Table 1.

III. RESULTS AND DISCUSSION

a. Confusion Matrix Results of the models used for Thyroid Prediction

The complete dataset in the database used is 3774. Eighty percent (80%) of the dataset was used for training the models (3019), while the remaining 20% (755) for testing the models by supplying the test data to the classifier of Decision Tree, XGboost, and Logistic Regression algorithms. The Prediction results obtained from each model are shown in Tables 3 through 5

Table 2: Confusion Matrix Variables

N = Test Data Size	Predicted NO	Predicted YES
Actual NO	TN	FP
Actual YES	FN	TP

Table 3: Confusion Matrix Result of Decision Tree

	N = 755	Predicted NO	Predicted YES
DECISION TREE	Actual NO	666	51
	Actual YES	37	1

Table 4: Confusion Matrix Result of Logistic Regression

	N = 755	Predicted NO	Predicted YES
LOGISTIC REGRESSION	Actual NO	714	0
	Actual YES	38	3

Table 5: Confusion Matrix Result of Gradient Boosting Model

	N = 755	Predicted NO	Predicted YES
XG BOOST	Actual NO	650	49
	Actual YES	53	3

b. Performance Evaluation and Comparison

Based on the frequencies of TP, TN, FP, and FN, the performance of each model in the expression of Accuracy, Precision, and Recall was estimated. Table 6 shows the confusion matric variables of the three models as compared.

$$Recall = \frac{TP}{TP + FN}$$

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

$$Precision = \frac{TP}{TP + FP}$$

Where TP represents True Positive,
TN means True Negative,
FP is False Positive and
FN is False Negative

Table 6: Performance Evaluation and Comparison of the three ML Models

Models	Accuracy	Precision	Recall
Logistic Regression	99.60	100	95
Gradient Boosting	92.58	94.23	48
Decision Trees	93.11	42	92.68

c. Result Discussion

In this study, three (3) models were used but of the three models; the Logistic Regression model outperforms the other two models in terms of Accuracy, Precision, and Recall.

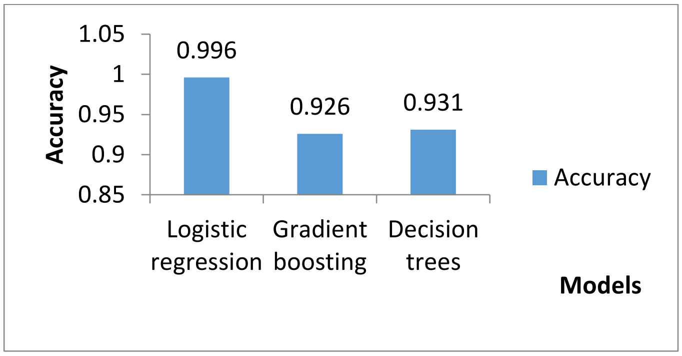

Figure 4: Accuracy of the three models compared

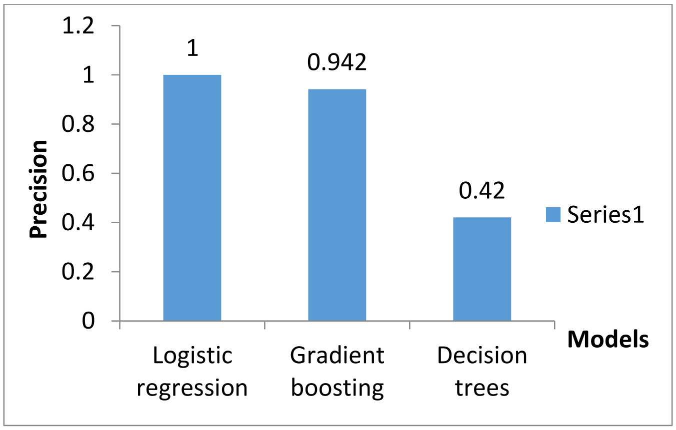

Figure 5: Precision of the three models compared

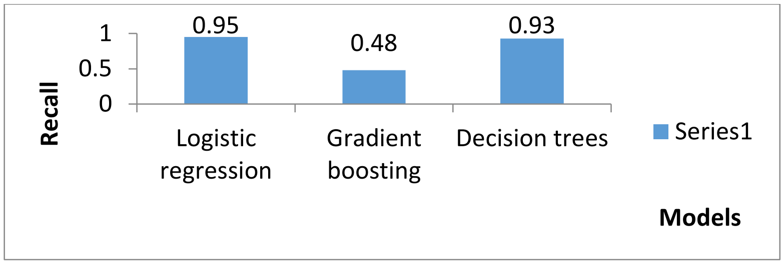

Figure 6: Recall of the three models compared

IV. CONCLUSION AND RECOMMENDATION

Currently, researchers globally have attained much progress in diagnosing thyroid disorders, but decreasing the huge variables required for the diagnosis of thyroid diseases is suggested. More variables mean a patient has to carry out more clinical tests, which is financially demanding and laborious. Therefore, there is a need to develop such types of algorithms based on thyroid disease predictive models that require collecting the least parameters from a patient needing diagnosis of thyroid disease. As a result of this, they are conserving both money and time needed for a patient to undergo diagnosis (Razia and Narasinga, 2016).

From the work carried out and the previous related work, it was concluded that machine learning algorithms would complement the efforts of human experts at predicting and diagnosing thyroid dysfunctions. Other research areas related to this work might include the use of more advanced techniques called deep learning to build predictive models. Deep learning technique even provides higher accuracy than machine learning techniques and provides the ability for the data to learn by itself from an unfamiliar and unknown data. This concept is otherwise known as Reinforcement Learning.

REFERENCES

Fontes R, Coeli CR, Aguiar F, Vaisman M. Reference interval of thyroid-stimulating hormone and free thyroxine in a reference population over 60 years old and in ancient subjects (over 80 years): comparison to young items. Thyroid Res. 2013 Dec 24; 6(1):13.

Gurmeet, K, Er.Brahmaleen, Kaur S. Artificial Neural Networks for Diagnosis of Thyroid Disease. International Journal for Technological Research in Engineering.2014; 2 (1): 56-59.

Guttag JV. Introduction to Computation and Programming Using Python: *With Application to Understanding Data. MIT Press. 2016. ISBN 978-0-262-52962-4*.

Jazzar MM. and Muhammad G. Feature selection based verification/identification system using fingerprints and palm print. Arabian J. Sci. Eng. 2013, 38 (4), 849–857.

Liu S, Liu S, Cai W, et al. Early diagnosis of Alzheimer's disease with deep learning. In: International Symposium on Biomedical Imaging, Beijing, China. 2014:1015–18.

Lui AY, and Pappas AM. Thyroid Dysfunction: Prediction and Diagnostics. 2015. https://docplayer.net/55037518- Thyroid-dysfunction-prediction-and-diagnostics.html.

Mark P. J. Vanderpump; The epidemiology of thyroid disease, British Medical Bulletin, Volume 99, Issue 1, 1 September 2011, Pages 39–51, https://doi.org/10.1093/bmb/ldr030

Milmann KJ and Avaizis M. Scientific Python, volume 11 of Computing in Science & Engineering. IEEE/AIP, March 2011.

Murphy, KP. Machine Learning - *A Probabilistic Perspective. The MIT Press. 2012; Pp. 245pp. ISBN 978-0-26201802-9*.

Ogbera, A., & Kuku, S. (2011). Epidemiology of thyroid diseases in Africa. Indian Journal of Endocrinology and Metabolism, 15(6), 82. https://doi.org/10.4103/2230-8210.83331

Okosieme O.E. Impact of iodination on thyroid pathology in Africa. Journal of the Royal Society of Medicine, 2006, 99(8): 396-401.

Rajam K, Jemin R, Priyadarsini A. A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques. IJCSMC, Vol. 5, Issue. 5 May 2016, pp.354–358.

Rougier NP. Scientific visualization and matplotlib tutorial. Euroscipy 2012 & 2013. Available: http://www.loria.fr/~rougier/teaching/matplotlib/matplotlib.html.

Smith MR and Martinez T. "Improving Classification Accuracy by Identifying and Removing Instances that misclassified". *Proceedings of International Joint Conference on Neural Networks, IJCNN 2011). Pp. 2690–2697.*

Srinivasan B and K. Pavya Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study. International Research Journal of Engineering and Technology (IRJET), 2016, Volume: 03 Issue: 11; 1191-1194.

Sulejmanovic, M., Cickusic, A., Salkic, S., & Bousbija, F. (2019). Annual Incidence of Thyroid Disease in Patients Who First Time Visit Department for Thyroid Diseases in Tuzla Canton. Materia Socio Medica, 31(2), 130. https://doi.org/10.5455/msm.2019.31.130-134

Tsegaye B, Ergete W. Histopathologic pattern of thyroid disease. East Afr Med J. 2003;80:525–8.

Authors Brief Profile

Dr. Ajayi Ayodeji Folorunsho is a researcher with about sixty peer-reviewed articles in reputable journals. He is a veterinary Doctor, who have also obtained masters and Ph.D. degrees in Human Physiology. He is a staff of the Faculty of Basic Medical Sciences of the Ladoke Akintola University of Technology, Ogbomoso, Nigeria since 2006. A member of the Veterinary Council of Nigeria, Physiological Society of Nigeria, Bioinformatics Society of Nigeria, among others. His research interest includes Reproductive Physiology, Bioinformatics, and other areas of Artificial Intelligence. He can be reached by phone on +2348033834495, and through E-mail aajayi22@lautech.edu.ng

Akindele Akinyinka Tosin is an eLearning Consultant at Kampala International University, Uganda. He also serves as an instructional Designer/eTutor at the Department of Computer Science, LAUTECH Open and Distance Learning Center (LODLC), Nigeria. Mr. Akindele holds an M.Tech degree in Computer Science from LAUTECH and is presently pursuing a Ph.D. at the same institution. His research areas include mHealth, Fuzzy Systems, Computer Vision, NLP, eLearning, and application of Artificial Intelligence in Healthcare, Education, and other sectors. He can be reached by phone on +2347030408417, and through email: atakindele@lautech.edu.ng

Opetunde Adepoju is a B.Tech graduate of the Physiology Department, Ladoke Akintola University of Technology, Nigeria. She is a skilled, driven, focused, and self-taught Data Science Professional whose research interest is in the Application of AI to healthcare and Business. She is an ambassador for Women in Data Science, a writer for Becoming Human and Devcenter on the trends in Artificial Intelligence and she won Forbes 30 under 30 Fellow award in 2018 for her efforts and contributions to making Africa a better place through Data Science. She can be reached by phone on +2347037062556, and through email: helenopetunde@gmail.com .