Thyroid Disorder Classification using Machine Learning Models

Vincent Peter C. Magboo, MD, MS, Ma. Sheila A. Magboo, MS

Department of Physical Sciences and Mathematics, University of the Philippines Manila
E-mail address: vcmagboo@up.edu.ph, mamagboo@u.edu.ph

ABSTRACT

Introduction:

Thyroid hormones are produced by the thyroid gland and are essential for regulating the basal metabolic rate. Abnormalities in the levels of these hormones lead to two classes of thyroid diseases – hyperthyroidism and hypothyroidism. Detection and monitoring of these two general classes of thyroid diseases require accurate measurement and interpretation of thyroid function tests. The clinical utility of machine learning models to predict a class of thyroid disorders has not been fully elucidated.

Objective:

The objective of this study is to develop machine learning models that classify the type of thyroid disorder on a publicly available thyroid disease dataset extracted from a machine learning data repository.

Methods:

Several machine learning algorithms for classifying thyroid disorders were utilized after a series of pre-processing steps applied on the dataset.

Results:

The best performing model was obtained by with XGBoost with a 99% accuracy and showing very good recall, precision, and F1-scores for each of the three thyroid classes. Generally, all models with the exception of Naïve Bayes did well in predicting the negative class generating over 90% in all metrics. For predicting hypothyroidism, XGBoost, decision tree and random forest obtained the most superior performance with metric values ranging from 96-100%. On the other end in predicting hyperthyroidism, all models have lower classification performance as compared to the negative and hypothyroid classes Needless to say, XGBoost and random forest did obtain good metric values ranging from 71-89% in predicting hyperthyroid class.

Conclusion:

The findings of this study were encouraging and had generated useful insights in the application and development of faster automated models with high reliability which can be of use to clinicians in the assessment of thyroid diseases. The early and prompt clinical assessment coupled with the integration of these machine learning models in practice can be used to determine prompt and precise diagnosis and to formulate personalized treatment options to ensure the best quality of care to our patients.

***Keywords**: thyroid disorders, machine learning, feature importance, SMOTE, XGBoost*

INTRODUCTION

The thyroid is a butterfly-shaped organ producing the thyroid hormones the levels of which play an important function of regulating the basal metabolic rate. Sufficient levels of these metabolic thyroid hormones are crucial for protein synthesis, for fetal and childhood tissue development and growth, for normal development of the nervous system in utero, in early childhood and continuing further to support neurological function in adults [1]. Derangement in the levels of these hormones lead to two classes of thyroid diseases namely: hyperthyroidism and hypothyroidism characterized by hyperfunction and hypofunction of the thyroid gland, respectively. In hypothyroidism, many patients complain of fatigue, weight gain and intolerance to cold temperature while anxiety, weight loss and sensitivity to heat are common symptoms of hyperthyroidism [2]. Needless to say, detection and monitoring of these two general classes of thyroid diseases require accurate measurement and interpretation of thyroid function tests [3].

The technological advancements in data mining techniques including processing and computation, machine learning (ML) approaches can also be applied to classify thyroid diseases [4]. Mollica et al., applied machine learning approach coupled with oversampling techniques and Bayesian networks framework on classification of thyroid tumors on histopathological images [5]. Authors concluded that integrating ML models in clinical practice could help reduce a pathologist's workload on top of improving disease diagnosis. In the study by Alyas et al., researchers applied several ML algorithms like decision tree, random forest algorithm, k-Nearest Neighbors (KNN), and artificial neural networks (ANN) on a thyroid disease dataset [6]. Results showed random forest with the highest classification accuracy at 94.8%. In [7], authors applied ML machine learning-based techniques to predict hypothyroidism namely: decision tree, random forest, naive Bayes, and ANN. Results showed decision tree and random forest generated the highest classification performance with an accuracy of 99.6% and 99.3%, respectively.

The objective of this study is to develop machine learning models that classify the type of thyroid disorder on a publicly available thyroid disease dataset extracted from a machine learning data repository. The clinical utility of ML models to predict a class of thyroid disorders has not been fully elucidated. The main contribution of this research is to find robust and reliable prediction models that will assist healthcare professionals in assessing the type of thyroid disease.

METHODOLOGY

The study was performed in several steps. The first step was loading of the dataset. Pre-processing techniques applied to the dataset include data cleaning, imputation method, and utilizing Synthetic Minority Oversampling TEchnique (SMOTE) for handling data imbalance. The next step was the application of several ML algorithms followed by assessment of classification performance. The machine learning pipeline for this study is shown in Figure 1.

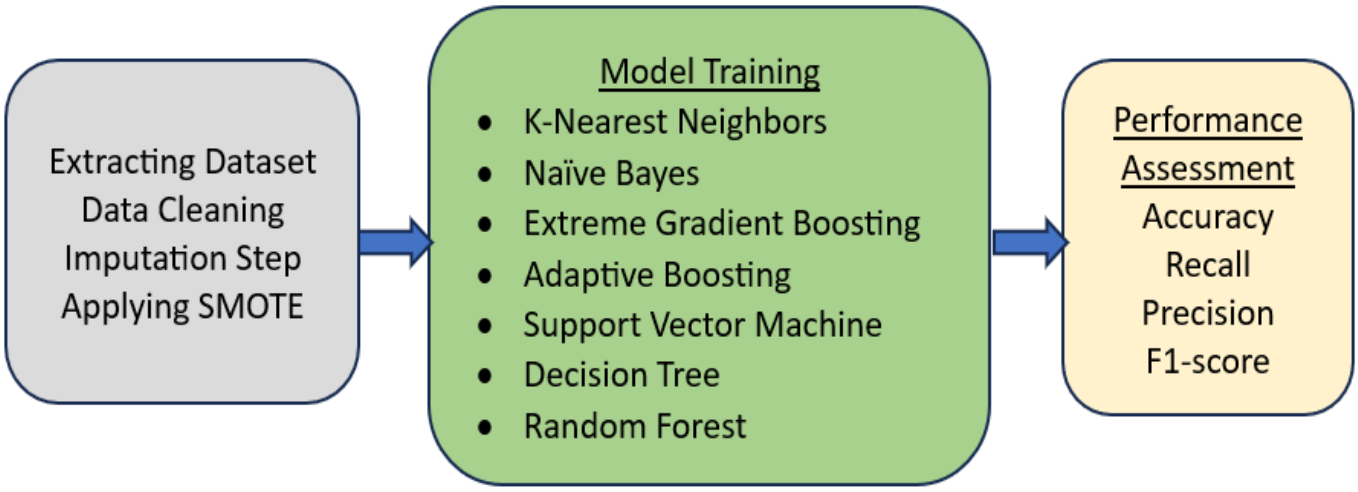

FIGURE 1. Machine Learning Pipeline for Thyroid Disorder Classification

TABLE 1. Independent attributes for the Thyroid Disorder Classification

Attribute	Data Type	Attribute	Data Type	Attribute	Data Type
patient_id	object	sick	object	FTI	float
age	integer	lithium	object	TBG	float
on_thyroxine	object	goitre	object	TSHmeasured	object
query_on_thyroxine	object	tumor	object	T3measured	object
on_antithyroid_meds	object	hypopituitary	object	TT4measured	object
query_hyperthyroid	object	psych	object	T4Umeasured	object
pregnant	object	TSH	float	FTImeasured	object
thyroid_surgery	object	T3	float	TBGmeasured	object
I131_treatment	object	TT4	float	referralsource	object
query_hypothyroid	object	T4U	float	binaryclass	object

Dataset Description

In this study, a publicly available thyroid disease extracted from a publicly available machine learning repository (University of California Irvine Machine Learning Repository) was used [8]. This dataset contains 9,172 anonymized thyroid disease cases from Garavan Institute, Sydney, Australia. The dataset consisted of 31 columns including the target variable, diagnosis. The listing of the independent attributes is seen in Table 1.

Pre-processing Steps

To prepare the dataset for machine learning, data cleaning and pre-processing methods were applied. Redundant and irrelevant variables such as 'TSHmeasured', 'T3measured', 'TT4measured', 'T4Umeasured', 'FTImeasured', 'TBGmeasured', 'patient_id', 'referralsource' were dropped from the dataset as they were mainly boolean variables with no predictive capability. Rows with inconsistent values, particularly those age over 100 years old, were likewise gnoses (negative, hypothyroid, and hyperthyroid) were retained for ML application as other diagnoses were deemed not relevant to the main focus of this research study. Hence, as a result of these data cleaning, the dataset was reduced to 7,142 records. The dataset has a severe imbalance with negative class comprising 89.4% (6,384 records) while the hypothyroid and hyperthyroid classes comprised 8.1% (582 records) and 2.5% (175 records) respectively. To address this severe imbalance, SMOTE was utilized.

Machine Learning Models

The dataset was split into 25% testing and 75% training with 10-fold cross validation. Python 3.8 and its ML libraries (scikit-learn, pandas, Matplotlib, seaborn, and NumPy) were used. Several ML models were utilized to predict thyroid diseases namely: k-Nearest Neighbors (kNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost).

Performance Metrics

Metrics such as accuracy, recall, precision, and F1 score were computed to assess classification performance. Accuracy refers to the ability of the ML model to predict the classes of the dataset correctly and assess how close or near the predicted value is to the actual or theoretical value [9]. Recall is the ratio of the correctly classified number of positive instances to the number of all instances whose actual class is positive [10]. Recall is also called the true positive rate or sensitivity rate. The precision, sometimes called the positive predictive value, denotes the proportion of the retrieved samples which are relevant and is calculated as the ratio between correctly classified samples and all samples assigned to that class [11]. F1—score is defined as the harmonic

TABLE 2. Performance Metrics of the ML Models for Thyroid Disorder Classification

ML Model	Accuracy	Target Class	Recall	Precision	F1-score
XGBoost	99	Negative	99	100	99
		Hypothyroid	100	97	99
		Hyperthyroid	89	74	80
Decision Tree	98	Negative	99	99	99
		Hypothyroid	99	98	98
		Hyperthyroid	80	73	76
Random Forest	98	Negative	99	100	99
		Hypothyroid	100	96	98
		Hyperthyroid	89	71	79
AdaBoost	98	Negative	99	99	99
		Hypothyroid	100	73	96
		Hyperthyroid	70	99	75
Support Vector Machine	94	Negative	94	99	96
		Hypothyroid	97	84	90
		Hyperthyroid	93	35	51
k-Nearest Neighbors	92	Negative	93	98	96
		Hypothyroid	86	77	81
		Hyperthyroid	80	34	47
Naïve Bayes	36	Negative	31	93	47
		Hypothyroid	67	46	54
		Hyperthyroid	84	4	7

mean of precision and recall and as such, to generate a high F1-score, necessarily require to have high values of recall and precision [12]. Additionally, the feature importance scores of the best performing model was also generated.

RESULTS AND DISCUSSION

The performance metrics of the various ML models are shown in Table 2. The best performing model is XGBoost with a 99% accuracy. XGBoost also generated the highest recall, precision, and F1-score for each of the three thyroid classes. Following XGBoost is random forest with an accuracy of 98% and showing very good recall, precision, and F1-scores for each of the three thyroid classes with XGBoost. AdaBoost and decision tree also obtained excellent accuracy rates of 98%. On the other hand, Naïve Bayes performed the worst with a measly accuracy rate of 36%. Additionally, its predictive capability for all the three thyroid classes were below par indicating its inability to predict thyroid disorders.

Generally, all ML models with the exception of Naïve Bayes did well in predicting the negative class generating over 90% in all metrics. This is expected as the negative class had the greatest number of instances in the dataset. For predicting hypothyroidism, XGBoost, decision tree and random forest obtained the most superior performance with metric values ranging from 96-100%. On the other end in predicting hyperthyroidism, all models have lower classification performance as compared to the negative and hypothyroid classes. Note that the hyperthyroid class only constituted 2.5% of the entire dataset. Nonetheless, XGBoost and random forest did obtain good metric values ranging from 71-89% in predicting hyperthyroid class. Likewise, AdaBoost and decision tree yielded fairly acceptable metric values in predicting hyperthyroid class. However, support vector machine, k-Nearest Neighbors and Naïve Bayes generated poor predictive capability in classifying hyperthyroidism more prominently with its very low precision and F1-scores. Nonetheless, our results highlight the importance of addressing the severe data imbalance to obtain a more reliable diagnostic performance.

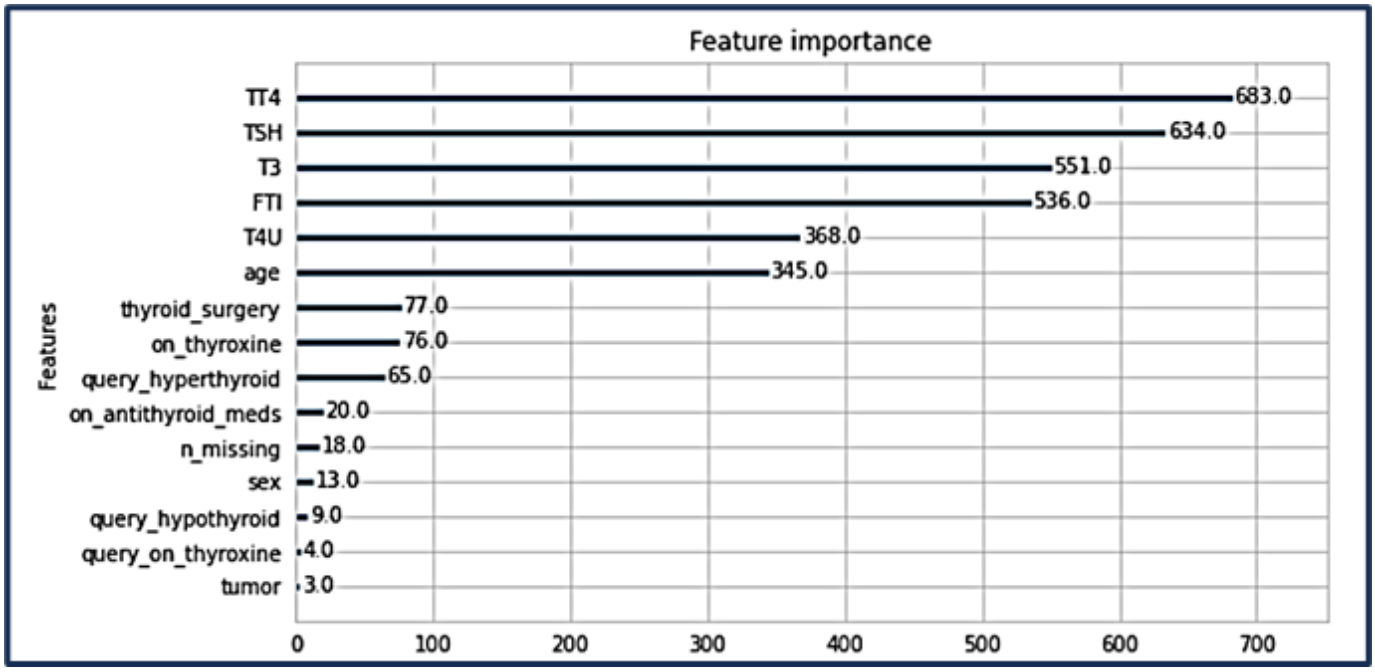

FIGURE 2. Feature Importance of Attributes of the Best Performing Model

As a measure to address the severe imbalance in this dataset, SMOTE was utilized. SMOTE can sufficiently increase the instances of minority samples so that the classification algorithm can increase the learning of minority samples during the training of the data [13]. Machine learning algorithms can be biased to favor the majority class in the presence of an imbalance [14]. SMOTE as an oversampling technique is a common measure utilized in ML to handle imbalanced datasets by creating copies of the minority class instances to balance the dataset which effectively led to a reduction in the bias and in the improvement of the classification accuracy of the model [13, 14, 15, 16].

The feature importance scores of attributes for the best performing model (XGBoost) are seen in Figure 2. The top important features were the hormone test level measurements (TT4, TSH, T3, FTI, T4U) while surprisingly though that the other attributes did not perform well in predicting our target variable. This confirmed our clinical suspicion that the hormone tests are the most helpful in our aim to predict target diagnosis as seen in the correlation heatmap in Figure 3. Only the attributes TT4, TSH, T3 and FTI had a strong positive correlation with the target variable. Feature importance highlights which attributes utilized by the ML model have higher predictive capability as compared to the other attributes. The identification of these features can aid in the model explainability [17]. Feature importance scores also provide insight into the data and the model by identifying most and least relevant in predicting the target variable. It also serves as a basis for dimensionality reduction by removing those attributes with lowest feature importance scores. This act simplifies the model which consequently lead to faster machine execution and also improved diagnostic performance of the model.

As to the metrics and best performing models, our results are comparable with other studies [4, 5, 6, 7, 18] which utilized traditional ML models applied to thyroid disease dataset. These findings suggest the feasibility of applying the machine learning approaches to predict thyroid disorders with acceptable results. The clinical utility of this study is even more highlighted with robust models that can provide faster and with high reliability to assist healthcare professional in predicting thyroid disorders as well as enable clinicians to propose personalized treatment options for our patients [19].

CONCLUSION

Alterations in the levels of thyroid hormones generally lead to two classes of thyroid diseases namely: hyperthyroidism and hypothyroidism characterized by hyperfunction and hypofunction of the thyroid gland, respectively. In this study, several machine learning

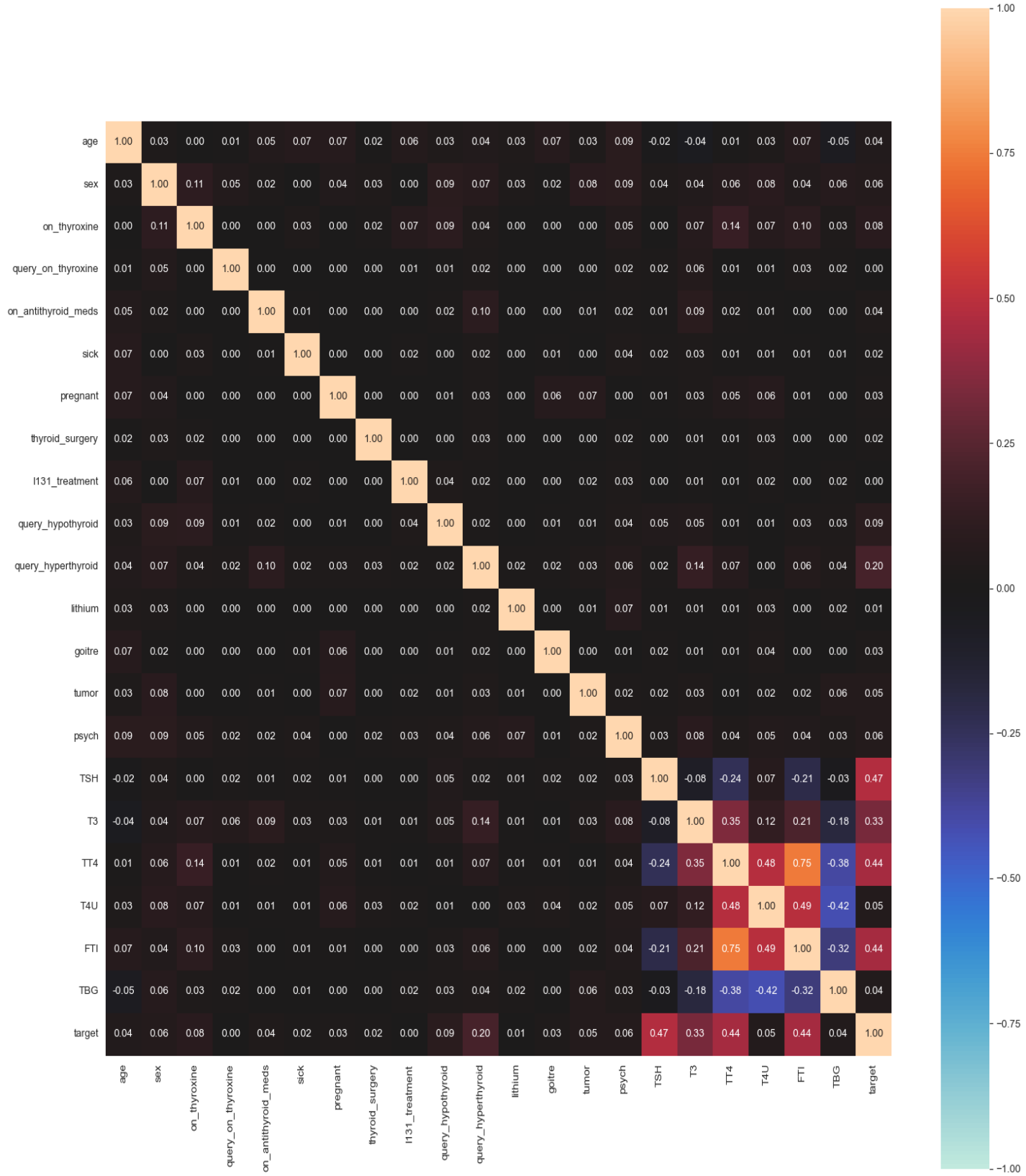

FIGURE 3. Correlation Heatmap of Predictor Variables for Thyroid Disorder Classification

models for classifying thyroid disorders were applied on a publicly available thyroid disease dataset from a machine learning data repository. The best performing model was obtained by with XGBoost with a 99% accuracy and showing very good recall, precision, and F1-scores for each of the three thyroid classes. Generally, all ML models with the exception of Naïve Bayes did well in predicting the negative class generating over 90% in all metrics. For predicting hypothyroidism, XGBoost, decision tree and random forest obtained the most superior performance with metric values ranging from 96-100%. On the other end in predicting hyperthyroidism, all models have lower classification performance as compared to the negative and hypothyroid classes Needless to say XGBoost and random forest did obtain good metric values ranging from 71-89% in predicting hyperthyroid class. Likewise, AdaBoost and decision tree yielded fairly acceptable metric values in predicting hyperthyroid class.

Future enhancement should include explainable artificial intelligence tools for better understanding of the models by the clinicians. Additionally, ML models could also be applied to larger datasets which combines patient symptoms, comorbidities, and radiographic features coming in the quest for excellent diagnostic accuracy. The findings of this study were encouraging and had generated useful insights in the application and development of faster automated models with high reliability which can be of use to clinicians in the assessment of thyroid diseases. The early and prompt clinical assessment coupled with the integration of these ML models in practice can be used to determine prompt and precise diagnosis and to formulate personalized treatment options to ensure the best quality of care to our patients.

REFERENCES

1. Gordon Betts J, Young K.A., Wise J, et al. (2022). The Thyroid Gland. In Anatomy and Physiology, 2nd Ed. OpenStax. https://openstax.org/books/anatomy-and-physiology-2e/pages/17-4-the-thyroid-gland.
2. Kwang-Sig Lee, Hyuntae Park. (2022). Machine learning on thyroid disease: a review. Front. Biosci. (Landmark Ed) 27 (3), 101. https://doi.org/10.31083/j.fbl2703101.
3. Andersen, S., Karmisholt, J., Bruun, N.H. et al. (2022). Interpretation of TSH and T4 for diagnosing minor alterations in thyroid function: a comparative analysis of two separate longitudinal cohorts. Thyroid Res 15, 19. https://doi.org/10.1186/s13044-022-00137-1.
4. Chaganti R, Rustam F, De La Torre Díez I, et al. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. Cancers (Basel). 2022 Aug 13;14(16):3914. doi: 10.3390/cancers14163914. PMID: 36010907; PMCID: PMC9405591.
5. Mollica G, Francesconi D, Costante G, et al. (2022). Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms. IFAC-PapersOnLine, 55 (40) : 67 – 72. https://doi.org/10.1016/j.ifacol.2023.01.050.
6. Alyas T, Hamid M, Alissa K, et al. (2022). Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. BioMed Research International, vol. 2022, Article ID 9809932, 10 pages. https://doi.org/10.1155/2022/9809932.
7. Guleria K, Sharma S, Kumar K, Tiwari S. (2022). Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning. Measurement: Sensors, Volume 24, 2022, 100482. https://doi.org/10.1016/j.measen.2022.100482.
8. Quinlan, Ross. (1987). Thyroid Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C5D010.
9. Debal DA, Sitote TM. (2022). Chronic kidney disease prediction using machine learning techniques. J Big Data 9, 109. https://doi.org/10.1186/s40537-022-00657-5.
10. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 9, 11. https://doi.org/10.1186/s40561-022-00192-z.
11. Hicks SA, Strümke I, Thambawita V, et al. (2022). On evaluation metrics for medical applications of artificial intelligence. Sci Rep 12, 5979 (2022). https://doi.org/10.1038/s41598-022-09954-8.
12. De Diego IM, Redondo AR, Fernández RR, et al. (2022). General Performance Score for classification problems. Appl Intell 52, 12049–12063. https://doi.org/10.1007/s10489-021-03041-7.
13. Wei Du. (2022). Application of Improved SMOTE and XGBoost Algorithm in the Analysis of Psychological Stress Test for College Students. Journal of Electrical and Computer Engineering, vol. 2022, Article ID 2760986, 8 pages. https://doi.org/10.1155/2022/2760986.
14. Nishat MM, Faisal F, Ratul IJ, et al. (2022). A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. Scientific Programming, vol. 2022, Article ID 3649406, 17 pages. https://doi.org/10.1155/2022/3649406.
15. Yakshit, Kaur G, Kaur V, Sharma Y, Bansal B. (2022). Analyzing various Machine Learning Algorithms with SMOTE and ADASYN for Image Classification having Imbalanced Data. 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), Bhopal, India, 2022, pp. 1-7, doi: 10.1109/CCET56606.2022.10080783.

16. Priyadarshinee S, Panda P. (2022). Improving Prediction of Chronic Heart Failure using SMOTE and Machine Learning. 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1-6, doi: 10.1109/ ICCSEA54677.2022.9936470.
17. Collaris D, Weerts H, Miedema D, van Wijk J, Pechenizkiy M. (2022). Characterizing Data Scientists' Mental Models of Local Feature Importance. In Nordic Human-Computer Interaction Conference (NordiCHI '22). Association for Computing Machinery, New York, NY, USA, Article 9, 1–12. https://doi.org/10.1145/3546155.3546670.
18. Pal M, Parija S, Panda G. (2022). Enhanced Prediction of Thyroid Disease Using Machine Learning Method. 2022 IEEE VLSI Device Circuit and System (VLSI DCS), Kolkata, India, 2022, pp. 199-204, doi: 10.1109/ VLSIDCS53788.2022.9811472.
19. Magboo VC, Magboo MS. (2021). Machine Learning Classifiers on Breast Cancer Recurrences. In: Watrobski, J., Salabun, W., Toro, C., Zanni-Merk, C., Howlett, R., Jain, L. (eds.) 25th International Conference on Knowledge-Based and Intelligent Information & Engineering System 2021, Procedia Computer Science, vol 192, pp. 2742–2752. Elsevier, Warsaw, Poland. https://doi.org/10.1016/ j.procs.2021.09.044.