IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.004, LIIR166
DOI: https://doi.org/10.52731/liir.v004.166

Evaluating the Performance of Machine Learning Classifiers on Predicting Hypothyroidism for Public Healthcare Good

Senanu Okuboyejo [*], Jawad Haqbeen [†], Takayuki Ito [†]

Abstract

Hypothyroidism is an endocrine disorder in which the thyroid gland cannot secrete enough hormones. If left undetected and treated, it poses grave consequences on the patient's health and quality of life. Early detection is vital for treatment, enhancing the quality of a patient's life. Besides many sectors, artificial intelligence (AI) will drive health sector transformation, offering new approaches to optimize health systems' operation and reliability, ensuring not only techno-economic advantages but also improving patients' quality of life (QoL) in a meaningful way. Therefore, it is critical to find innovative approaches using AI. Towards this end, we initiate the study to evaluate the performance of Machine Learning Classifiers in predicting Hypothyroidism for Healthcare Good. This work uses supervised machine learning (ML) algorithms to predict hypothyroidism based on available features and identifies the best-performing classifier. We built and trained seven classifiers using specified ML algorithms. We presented an experimental case study, validating models and measuring performance. A comparative analysis of the classifiers revealed that the tree-based classifiers (Random Forest, Decision Tree, and Gradient Boost) outperformed other models based on the F1-score and AUC values, consistent with existing literature. This work has implications for the development of health informatics systems.

Keywords: Artificial Intelligence, Machine Learning, Hypothyroidism, Classification, Health Informatics, Healthcare

1 Introduction

The increasing evolution of technologies has resulted in cutting-edge advances in recent years, with artificial intelligence (AI) leading the pack [1], [2]. Healthcare institutions are fast becoming connected knowledge-based communities of practice for sharing knowledge, reducing administrative costs, and improving the quality of care [3]. Besides many sectors, AI will drive health sector transformation, offering new approaches to optimize health systems' operation and reliability, ensuring not only techno-economic advantages but also improving patients' quality of life (QoL) in a meaningful way. AI can potentially transform patient healthcare with its extensive power. AI techniques provide effective knowledge management, sharing, decision-making, and support [4]. For example, it supports healthcare practitioners in patient care by providing up-to-date health information synthesized from publications and clinical practices [5], [33].

[*] Metro State University, MN, USA
[†] Kyoto University, Kyoto, Japan

The connectedness of health institutions has also produced a massive volume of healthcare data [5], [6], characterized by volume, velocity, veracity, and variety. To derive value, it is essential to generate knowledge from these data. Data mining is the process of extracting interesting (non-trivial, implicit, previously unknown, and potentially actionable) patterns or knowledge from large datasets [7], [8]. It uses machine learning algorithms to identify patterns in large datasets that exist in the medical domain. Machine Learning (ML) is a subset of AI techniques that enables computer systems to learn from previous experience (i.e., data observations) and improve their behavior for a given task [8]. ML techniques include Support Vector Machines (SVM), Decision Trees (DT), Gaussian Naïve Bayes (GNB), k-Means clustering, Regression, and Artificial Neural Networks.

Classification (also referred to as prediction modeling), one of the data mining tasks, is a form of supervised learning that assigns data instances to predefined labels. In supervised learning, classifiers are trained on labeled datasets. Classification employs ML techniques to learn models and fit them to unseen datasets. It has been used extensively in healthcare for disease diagnoses such as Type-2 Diabetes [9], Delirium [10], [11], stroke [12], thyroid gland tumor [13], COVID-19 [14], heart disease and failure [15]–[19], predicting the success of clinical procedures such as kidney transplant [20], approach for clinical decision support in the patient selection for targeted therapy in advanced non-small cell lung cancer (NSCLC) [21].

Prior research has also explored the applicability of ML-based classification to thyroid disease [22], [23]. Thyroid disease is an endocrine disorder in the human body that can result in hypo- or hyperthyroidism, depending on the secretion level of the thyroid hormone. Hypothyroidism is a disorder in which the thyroid gland underacts. It does not produce enough hormones and can result in different health problems, such as obesity and heart disease. Early detection is vital for treatment purposes. Improving the prediction accuracy of the disorder is crucial for correct medical diagnosis, which improves the quality of life. This work aims to identify the best-performing ML classifier (measured by accuracy) to predict the presence/absence of hypothyroid based on available features.

In achieving this aim, we will build and learn different ML-based classifiers and subject the classifiers to a comparative analysis to determine the best classifier based on our dataset's appropriate performance evaluation metrics. The rest of the paper is structured as follows: Section 2 describes the materials and methods, a summary of our results, findings, and implications is presented in Section 3, and we draw a conclusion in Section 4.

2 Methods

The proposed approach predicts thyroid disease based on the patient's historical and current data. In this section, we describe the data collection process and the proposed features model, and finally, we focus on the machine learning algorithms used to conduct the study and their validation. The design methodology explored by [11] was adapted for this study and modified to suit our dataset and ML algorithms. The ML-based data analysis process begins with data collection and preprocessing, followed by model training, classification (prediction) with model performance evaluation, and output visualization. We discuss the procedures below.

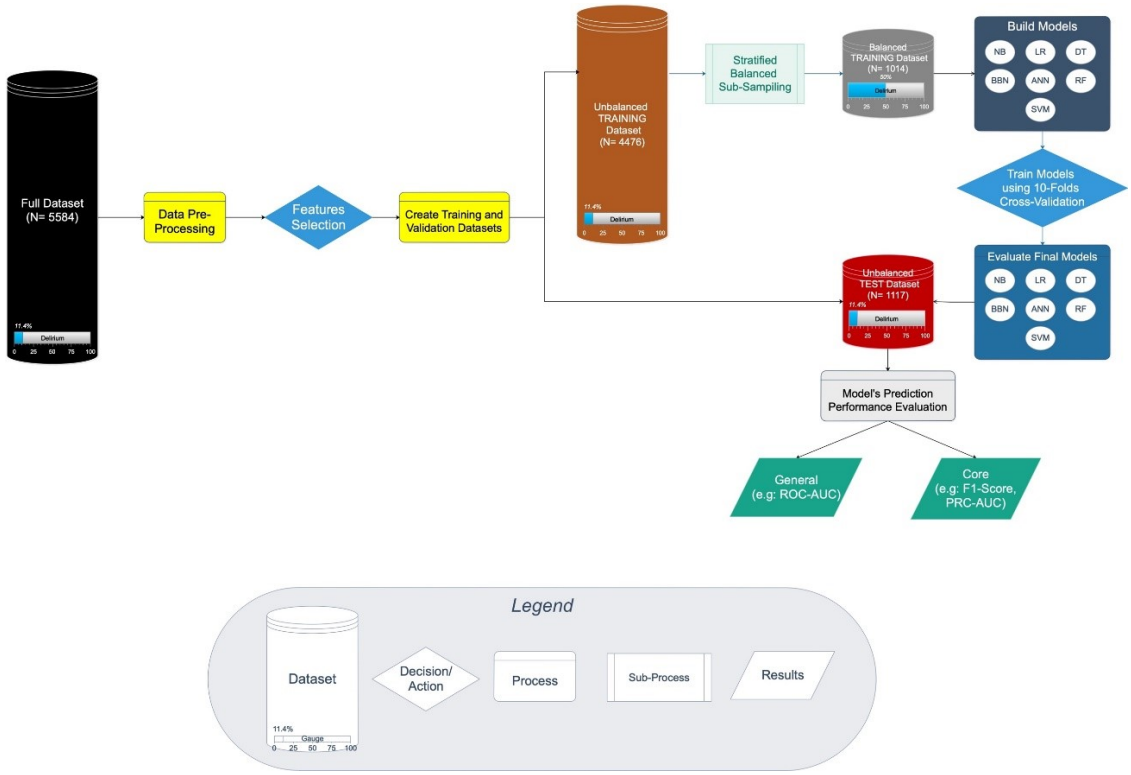

Figure 1: Pipeline of data mining methodology for our study, adopted from [11].

A. Data Acquisition

We imported the hypothyroid disease data from the UCI machine learning repository, available at *(https://doi.org/10.24432/C5D010)*. The dataset was mounted to the processing environment. It comprises 3076 records with 30 attributes (29 predictors; 1 class). A complete description of the attributes is presented in the appendix. The features are categorical and real, with multivariate variables. We used Python running on Google Colab to perform classification on the dataset.

B. Data Preprocessing

The dataset contains large numbers with missing and noisy data. We preprocess the dataset to make accurate predictions. Our preprocessing activities include the following:

- Data Transformation: it changes the data format from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks. In addition, it was also a necessary operation because of the functional constraint of the scikit-learn library [19]. We transformed the categorical features into multiple dummy features to make them machine-readable. All labels were converted into numeric formats using Label encoding, while customary binary encoding was applied to the target attribute "class." The target variable is a binary class that checks the presence or absence of thyroid disease. In the presence of thyroid disease (positive), the value is set to 1, otherwise 0.

- Data cleaning (missing data values): The dataset had 1522 missing values from 6 features. All non-numeric entries, such as special characters, were replaced with "NaN," and all the missing values were replaced with the mean value of the corresponding feature. The missing values in the categorical features were replaced with the most frequent category. The process is called Imputation.

- Feature selection: Among the 30 features of the data set, seven (7) features that did not provide relevant information were dropped. The remaining 23 attributes are considered necessary as they contain vital clinical records. Clinical records are crucial to the diagnosis of thyroid disease.

C. Data Partitioning

The dataset was split into training and validation subsets. The training subset had 80% of the dataset (N=2460), while the remaining 20% was the test subset (N=616). The split was stratified on the outcome variable to consider the high imbalance. The outcome variable had a dominant class value of "negative," which can affect the prediction performance of the models.

D. Training Thyroid Disease Classification Models

The dataset was split into training and validation subsets. The training subset had 80% of the dataset (N=2460), while the remaining 20% was the test subset (N=616). The split was stratified on the outcome variable to consider the high imbalance. The outcome variable had a dominant class value of "negative," which can affect the prediction performance of the models. The prediction modeling for our dataset was implemented as a binary classification problem, where the prediction output represents the hypothyroid outcome per patient. The prediction models were developed using Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting (GB), Gaussian Naïve Bayes (GNB), Decision Tree, Support Vector Machine, and Neural network's multi-layer perceptron (MLP) algorithms. These models have a history of application to disease prediction and classification. **KNN** is a supervised ML algorithm used extensively for classification and pattern recognition [24]. It attempts to classify data points based on their closeness or nearness to each other. The closer the data points are, the higher the likelihood of having the same classification. The similarity likelihood is derived using the Euclidean distance. The "k" is randomly set, and the algorithm identifies the closest data points. The classification for a data instance is the class with the most significant distances. For our KNN, our "k" was set to 5.

GNB is a supervised learning classification technique built on the Bayes theorem. The classifier assumes that class features are independent of each other. As its name implies, a **decision tree (DT**) is a tree-like graph comprised of nodes (leaf, inner, and root). The leaf nodes are the class nodes. The inner nodes (non-leaf nodes) are test attributes, and the branches depict test results. DTs run on a greedy algorithm. It is a binary classification algorithm based on information available for the classification (entropy). It has been used extensively in the medical domain for diagnosis [25] and clinical decision support, with high predictive performance. The Gini index was used for our study to select the best split. **RF and GB** are also tree-based algorithms. They are ensemble methods that combine multiple decision boundaries to get optimal performance. They use the output of a collection of trees to arrive at a prediction decision. They aggregate numerous decision trees to have stronger classifiers with better performances. In RF, the trees are learned

independently (in parallel), while GB learns its trees sequentially. RF outperformed other classifiers in predicting disease risks of a population sample in [26].

SVM-based classifiers create large hyperplanes in high dimensional space, which maximizes the separation between data points [27]. This hyperplane separation leads the classifiers to be discriminative in nature. The SVM provides better accuracy but is expensive in terms of computational time. These classifiers are designed for binary classification and have also been widely used in predictive modeling [28], such as biomedical image classification [29], [30], and disease risk prediction [26]. Our SVM classifier was trained on different values to optimize its parameters.

MLP is a classifier consisting of several perceptron and multiple layers. The classifier receives a signal (number of attributes) at the input layer (non-linear activation), which is transmitted in a forward propagation to the output layer. It then employs a supervised learning technique for its backpropagation. The prediction occurs at the output layer, which may be either binary or multiclass prediction.

Table 1 : ML-ALGORITHM PARAMETERS AND VALUES

ML-Algorithm	**Hyperparameter**	**Values**	**Tuning Range**
Random Forest	Bootstrap	True	[True, False]
	Number of Estimators	1400	[200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]
	Max. Depth	70	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None]
	Min. Samples Split	10	[2, 5, 10],
	Max. features	None	['auto,' None]
	Min. Sample Leaf	1	[1, 2, 4],
Nearest Neighbour	n_neighbour	5	
	Metric	Minkowski	
	p	2	
	Weights	uniform	
Gradient Boosting	Learning rate	0.05	[0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]
	Max_depth	80	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
	Max_features	None	['auto,' None]
	N_estimators	231	[10, 231, 452, 673, 894, 1115, 1336, 1557, 1778, 2000]
	Subsample	0.68	
Gaussian Naïve Bayes	Var_smoothing	1.00E-09	
	Priors	None	
Decision Tree	Criterion	Gini	
	Splitter	Best	
	Max_depth	None	
	Min_sample split	2	
	Min_sample leaf	1	

Support Vector Machine	C	105	[0.1,1, 10, 100], [90,95,100, 105,107, 110]
	Gamma	0.0005	[1,0.1,0.01,0.001], [0.005,0.01,0.015,0.02]
	Kernal	rbf	'rbf', 'poly', 'sigmoid'
Neural Network (MLP)	Solver	Adam	['sgd', 'lbfgs', 'adam']
	Learning rate	adaptive	['constant,' 'adaptive']
	Hidden layer size	110	[(50,50,50), (50,100,50), (100,150,100)]
	Alpha	0.0001	[0.0001, 0.05]
	Activation	tanh	['tanh,' 'relu']

We investigated the seven different ML-based classification models. All the classification models were trained using a 5-fold stratified cross-validation training approach. The stratification ensured that the outcome class ratio in each fold was constant, ruling out any sampling bias that may affect the classification results. We mainly used the sci-kit-learn library [31] to train classification models with different parameter settings. The random state parameter was set to 10 to ensure the same result is obtained when the train test split function divides the matrices into random train and test subsets. Hyperparameters were initially tuned using Grid search to optimize the parameter values. The best parameters were then used as input values for the Randomized Search parameters for the SVM and MLP classifiers. For the other classifiers, only the randomized search was used. The different parameters that were tested during the random search for the different algorithms are provided in Table 2 below. We explored multiple ML algorithms to predict a patient's presence or absence of hypothyroid. We compared the results from the algorithms based on different performance evaluation metrics to determine the best classifier of an instance of our dataset.

3 Results

In Table 2 below, we highlight the values of the confusion matrix for each classifier. Table 3 below presents the prediction performance of the seven ML classifiers. The results of each classifier were examined using the "Area Under Curve" (AUC) and F1- scores. The AUC score was used as the primary performance evaluation metric to select the best model due to the high imbalanced nature of the dataset. For binary classification in healthcare, specificity and sensitivity are essential measures represented by the precision and recall of the classifiers. Sensitivity measures true positives, while specificity measures correct negative findings[32]. For our purpose, the prediction model must be precise in predicting a positive case of hypothyroid. Therefore, to improve the selection of the best model, we also considered the precision and recall of each classifier. In cases where the AUC score was the same for different models, we considered the F1 score, precision, and recall. The F1-scores (weighted average) and AUC values are reported in Table 3 below.

Table 2 : TABLE SHOWING CONFUSION MATRICES FOR ALL CLASSIFIERS ON THE VALIDATION DATASET

CLASSIFIER		ACTUAL VALUES	
		POSITIVE	NEGATIVE
KNN	PREDICTED	557	1
		34	24
GNB		142	416
		4	54
DT		557	1

		3	55
RF		557	1
		2	56
GB		557	1
		3	55
SVM		558	0
		35	23
MLP		557	1
		9	49

Table 3 : PERFORMANCE EVALUATION RESULTS FOR THE CLASSIFIERS

Classifier	**Precision**	**Recall**	**F1**	**AUC**
KNN	0.94	0.94	0.93	0.706
GNB	0.89	0.32	0.38	0.593
Decision Tree	0.99	0.99	0.99	0.97
Random Forest	0.99	0.99	0.99	0.97
Gradient Boosting	0.99	0.99	0.99	0.97
SVM	0.97	0.97	0.97	0.85
MLP	0.97	0.98	0.97	0.9

4 Discussion

In this study, we aimed to select the best classifier for predicting hypothyroidism. We evaluated several machine learning classifiers and found that the Decision tree, Random forest, and Gradient boosting classifiers had excellent performance, having the same values for all performance metrics. The confusion matrix for RF and GB explained the valid and predicted classifications, with minimal errors in the prediction (FP=1, FN=2) and (FP=1, FN=3) respectively, indicating low variance (no overfitting) and low bias (no underfitting). We can conclude that the classifiers have performed well (they accurately predicted 98% of the data class) and are good predictive models based on these metrics, especially for our highly imbalanced dataset. The models were more overfit with small sample sizes, with the validation loss decreasing significantly with more training samples. This suggests that the models may be more prone to overfitting when trained on smaller datasets. Therefore, it is important to use a larger dataset to train these models to avoid overfitting.

While our study demonstrated the good performance of our approach, it is not without limitations. The model's performance on the validation set also improved with more samples, resulting in a good model. However, with the MLP classifier, it performed better on the validation, steadily and significantly rising with an increase in the sample size. The AUC score would have continued to improve with more samples. The AUC showed a better performance of the model on the validation set, steadily and significantly improving with an increase in sample size. Gaussian Naïve Bayes algorithm had the worst performance, recording the lowest values for all its metrics. Thus, we would conduct a study to build more ML-based classifiers for hypothyroidism prediction. There are many directions for future work. For example, extending this approach to different healthcare domains, diverse and more samples backgrounds could yield valuable insights into

the generalizability of our findings. More generally, extending the study to other desirable public health domains is another important direction for future work.

5 Conclusion

Our research has shed light on the efficacy of building various ML-based classifiers for hypothyroidism prediction. We trained and validated these models on our dataset. We adopted the pipeline of Data Mining method, however the values for the variables, sample size was different, as well as the Models were different in our study. Our results revealed the tree-based classifiers as the best classifiers for hypothyroidism, which were consistent with findings in the literature. In the future, the classifiers can be applied to more recent datasets with increased sample sizes to validate their performance further. Unsupervised learning techniques can also be introduced into the dataset to generate hidden patterns beyond binary classification.

Acknowledgement

This research was supported partially by the JST CREST fund (Grant Number: JPMJCR20D1, Japan) and JSPS KAKENHI (Grant Number: 23K17164, Japan).

References

[1] J. Balakrishnan, Y. K. Dwivedi, L. Hughes, and F. Boy, "Enablers and Inhibitors of AI-Powered Voice Assistants: A Dual-Factor Approach by Integrating the Status Quo Bias and Technology Acceptance Model," Inf. Syst. Front. 2021, vol. 1, pp. 1–22, Oct. 2021, doi: 10.1007/S10796-021-10203-Y.

[2] O. A. Nasseef, A. M. Baabdullah, A. A. Alalwan, B. Lal, and Y. K. Dwivedi, "Artificial intelligence-based public healthcare systems: G2G knowledge-based exchange to enhance the decision-making process," Gov. Inf. Q., p. 101618, Aug. 2021, doi: 10.1016/J.GIQ.2021.101618.

[3] R. Bose, "Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support," Expert Syst. Appl., vol. 24, no. 1, pp. 59–71, Jan. 2003, doi: 10.1016/S0957-4174(02)00083-0.

[4] Y. K. Dwivedi et al., "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," Int. J. Inf. Manage., vol. 57, p. 101994, Apr. 2021, doi: 10.1016/J.IJINFOMGT.2019.08.002.

[5] S. S. R. Abidi and S. R. Abidi, "Intelligent health data analytics: A convergence of artificial intelligence and big data," Healthc. Manag. Forum, vol. 32, no. 4, pp. 178–182, Jul. 2019, doi: 10.1177/0840470419846134.

[6] A. Kalantari, A. Kamsin, S. Shamshirband, A. Gani, H. Alinejad-Rokny, and A. T. Chronopoulos, "Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions," Neurocomputing, vol. 276, pp. 2–22, Feb. 2018, doi: 10.1016/J.NEUCOM.2017.01.126.

[7] Aalto-yliopisto, IEEE Computer Society, and Institute of Electrical and Electronics Engineers, ICDE 2016 Workshops : 2016 IEEE 32nd International Conference on Data Engineering Workshops : May 16-20, 2016, Helsinki, Finland. 2016.

[8] G. Nguyen et al., "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," Artif. Intell. Rev., vol. 52, no. 1, pp. 77–124, Jun. 2019, doi: 10.1007/s10462-018-09679-z.

[9] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-Mclallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," Big Data, vol. 3, no. 4, pp. 277–287, Dec. 2015, doi: 10.1089/big.2015.0020.

[10] F. Lucini, K. Fiest, H. Stelfox, and Lee Joon, "Delirium Prediction in the intensive care unit: a temporal approach," in 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society, 2020, pp. 5527–5530.

[11] H. N. Mufti, G. M. Hirsch, S. R. Abidi, and S. S. R. Abidi, "Exploiting machine learning algorithms and methods for the prediction of agitated delirium after cardiac surgery: Models development and validation study," JMIR Med. Informatics, vol. 7, no. 4, Oct. 2019, doi: 10.2196/14993.

[12] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," Neural Comput. Appl., vol. 32, no. 3, pp. 817–828, Feb. 2020, doi: 10.1007/s00521-019-04041-y.

[13] M. Böhland et al., "Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis," PLoS One, vol. 16, no. 9 September, Sep. 2021, doi: 10.1371/journal.pone.0257635.

[14] A. S. Albahri et al., "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," Journal of Medical Systems, vol. 44, no. 7. Springer, Jul. 01, 2020, doi: 10.1007/s10916-020-01582-x.

[15] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes," J. Clin. Epidemiol., vol. 66, no. 4, pp. 398–407, Apr. 2013, doi: 10.1016/J.JCLINEPI.2012.11.008.

[16] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," Int. J. Intell. Eng. Syst., vol. 12, no. 1, pp. 242–252, 2019, doi: 10.22266/ijies2019.0228.24.

[17] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," J. Big Data, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0244-x.

[18] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining

tools and machine learning techniques," Health Technol. (Berl)., vol. 10, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1/Published.

[19] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[20] S. A. A. Naqvi, K. Tennankore, A. Vinson, P. C. Roy, and S. S. R. Abidi, "Predicting kidney graft survival using machine learning methods: Prediction model development and feature significance analysis study," J. Med. Internet Res., vol. 23, no. 8, Aug. 2021, doi: 10.2196/26843.

[21] N. Kureshi, S. S. R. Abidi, and C. Blouin, "A predictive model for personalized therapeutic interventions in non-small cell lung cancer," IEEE J. Biomed. Heal. Informatics, vol. 20, no. 1, pp. 424–431, Jan. 2016, doi: 10.1109/JBHI.2014.2377517.

[22] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," Cancers (Basel)., vol. 14, no. 16, Aug. 2022, doi: 10.3390/cancers14163914.

[23] R. Jha, V. Bhattacharjee, and A. Mustafi, "Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society," Wirel. Pers. Commun., vol. 122, no. 2, pp. 1921–1938, Jan. 2022, doi: 10.1007/s11277-021-08974-3.

[24] H. Abbad Ur Rehman, C. Y. Lin, Z. Mushtaq, and S. F. Su, "Performance Analysis of Machine Learning Algorithms for Thyroid Disease," Arab. J. Sci. Eng., vol. 46, no. 10, pp. 9437–9449, Oct. 2021, doi: 10.1007/s13369-020-05206-x.

[25] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid Disease Prediction Using Machine Learning Approaches," Natl. Acad. Sci. Lett., vol. 44, no. 3, pp. 233–238, Jun. 2021, doi: 10.1007/s40009-020-00979-z.

[26] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Med. Inform. Decis. Mak., vol. 11, no. 1, pp. 1–13, Jul. 2011, doi: 10.1186/1472-6947-11-51/FIGURES/10.

[27] P. Kaur, R. Kumar, and M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," Multimed. Tools Appl., vol. 78, no. 14, pp. 19905–19916, Jul. 2019, doi: 10.1007/S11042-019-7327-8/FIGURES/3.

[28] A. Mathur and G. M. Foody, "Multiclass and binary SVM classification: Implications for training and classification users," IEEE Geosci. Remote Sens. Lett., vol. 5, no. 2, pp. 241–245, Apr. 2008, doi: 10.1109/LGRS.2008.915597.

[29] M. M. Rahman, S. K. Antani, and G. R. Thoma, "A learning-based similarity fusion and filtering approach for biomedical image retrieval using SVM classification and relevance feedback," IEEE Trans. Inf. Technol. Biomed., vol. 15, no. 4, pp. 640–646, Jul. 2011, doi: 10.1109/TITB.2011.2151258.

[30] Z. Camlica, H. R. Tizhoosh, and F. Khalvati, "Medical image classification via SVM

using LBP features from saliency-based folded data," Proc. - 2015 IEEE 14th Int. Conf. Mach. Learn. Appl. ICMLA 2015, pp. 128–132, Mar. 2016, doi: 10.1109/ICMLA.2015.131.

[31] F. Pedregosa FABIANPEDREGOSA et al., "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011.

[32] J. Shreffler and M. R. Huecker, "Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios," StatPearls, Mar. 2022.

[33] Mate A, Madaan L, Taneja A, Madhiwalla N, Verma S, Singh G, Hegde A, Varakantham P, Tambe M (2022) Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. Proceedings of the AAAI Conference on Artificial Intelligence 36(1111):12017–12025.

Appendix

Table 4 : DESCRIPTION OF THE DATASET FEATURES

Attribute	**Description**	**Datatype**
1. age	age of the patient	integer
2. sex	sex patient identifies	string
3. on_thyroxine	- whether the patient is on thyroxine	boolean
4. query on thyroxine	*Whether the patient is on thyroxine	boolean
5. on antithyroid meds	whether the patient is on antithyroid meds	boolean
6. sick	whether the patient is sick	boolean
7. pregnant	whether the patient is pregnant	boolean
8. thyroid_surgery	whether the patient has undergone thyroid surgery	boolean
9. I131_treatment	whether the patient is undergoing I131 treatment	boolean
10. query_hypothyroid	whether the patient believes they have hypothyroid	boolean
11. query_hyperthyroid	whether the patient believes they have hyperthyroid	boolean
12. lithium	whether the patient * lithium	boolean
13. goiter	whether the patient has goiter	boolean
14. tumor	whether the patient has a tumor	boolean
15. hypopituitary	whether the patient * hyper pituitary gland	float
16. psych	whether patient * psych	boolean
17. TSH_measured	whether TSH was measured in the blood	boolean
18. TSH	TSH level in blood from lab work	float
19. T3_measured	whether T3 was measured in the blood	boolean
20. T3	T3 level in blood from lab work	float
21. TT4_measured	whether TT4 was measured in the blood	boolean
22. TT4	TT4 level in blood from lab work	float
23. T4U_measured	whether T4U was measured in the blood	boolean
24. T4U	T4U level in blood from lab work	float
25. FTI_measured	whether FTI was measured in the blood	boolean
26. FTI	FTI level in blood from lab work	float
27. TBG_measured	whether TBG was measured in the blood	boolean
28. TBG	TBG level in blood from lab work	float
29. referral_source		string
30. target	hypothyroidism medical diagnosis	bool-eanean