A Comparison of Classification Methods of Hypothyroid Disease Prediction

Kulchaya Pongsawaeng*, Ausron Binmaduereh*, Panuphong Jenrotphondet*, and Orasa Patsadu*

Received: June 19, 2024
Revised: September 18, 2024
Accepted: September 23, 2024

* Corresponding Author: Orasa Patsadu, E-mail: orasa.p@mail.rmutk.ac.th

DOI: 10.14416/j.it/2025.v2.002

Abstract

This paper proposes a comparison of classification methods of hypothyroid disease prediction using data mining techniques. A dataset from the UCI repository with the thyroid disease dataset is used to prepare data with missing value handling, imbalance class handling, and suitable attribute selection. Then, the dataset is used to build the model by comparing the performance of classification methods such as Multilayer Perceptron, Support Vector Machine, and Decision Tree. The result shows that the Decision Tree achieves high performance with an accuracy of 99.61%, which is higher than the Multilayer Perceptron and Support Vector Machine with an accuracy of 96.46 % and 92.93%, respectively. In addition, we compared the result with state-of-the-art, which uses a similar technique to our proposed method. The result shows that our proposed method also outperforms previous research. Therefore, we decided to use Decision tree model for the prototype system development in hypothyroid disease prediction to support physicians' decision-making for diagnosis and treatment. Furthermore, this paper proposes data visualization to help users for primary risk assessment of a chance of hypothyroid disease to acknowledge risk before deciding to meet physicians using demographic information. Therefore, it will reduce the cost of medical and death rates.

Keywords: Hypothyroid Disease, Data Mining, Comparison of Classification Methods, Healthcare, Decision Support System.

1. Introduction

Thyroid disease is abnormal in the thyroid gland. Thyroid patients tend to increase, especially with hypothyroid disease. Most patients are women, who have an opportunity of thyroid disease more than men, especially females older than 40 years. Hypothyroid disease is divided into 2 states such as primary hypothyroidism and secondary hypothyroidism. The symptom indicates hypothyroid disease such as tiredness, pains, and aches, and gain weight. In addition, side effects from using certain drugs affect to hormone production of the thyroid gland such as cardiovascular drugs, mental conditions, and cancer disease. In addition, thyroidectomy affects to stop thyroid hormone production [1], [2]. Several researchers use classification techniques to solve medical [3] - [5]. In particular, research proposes a method for hypothyroid disease prediction using the classification method. Several classification techniques are used to build models for hypothyroid disease prediction, and each research focuses on model building [6] - [10]. According to literature reviews, most of the research also lacks a decision support system for primary risk self-assessment using data visualization. It became the motivation for our research. Therefore, our hypothesis is to propose a high-performance model for hypothyroid disease prediction and show data visualization for decision support in treatment.

This paper proposes a comparison of classification methods of hypothyroid disease prediction. The thyroid disease dataset from UCI is used to prepare data for model building by missing value handling, imbalance class handling, and appropriated attribute selection. Then, data is used to build the model by comparing the performance of classification methods such as Multilayer Perceptron, Support Vector Machine, and Decision Tree. In addition, our research proposes data visualization to help users for primary risk self-assessment to acknowledge primary risk before deciding to meet physicians

* Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Krungthep.

using demographic information to reduce medical costs and death rate.

The rest section of our research is section 2, which is related works. Section 3 is the research methodology. Section 4 is the results and case study. The last section is the conclusion.

2. Related Works

"Hypothyroidism is thyroid hormone production less than what is usual in the body, which causes hormone deficiency" [2], [11]. Several researchers proposed methods for hypothyroid disease detection. Olalekan, et al. [6] proposed a method for hypothyroid disease prediction using Ensemble learning with Bagging. There is comparative Ensemble learning by Bagging with Decision Tree and Bagging with SimpleCart. The result shows that the method for hypothyroid disease prediction using Bagging with Decision Tree achieves an accuracy of 99.60%. The accuracy of hypothyroid disease prediction using Bagging with SimpleCart is 99.55%.

Guleria, et al. [7] proposed a method to predict hypothyroidism using the classification method. Decision Tree, Random Forest, Naïve Bayes Multiclass classifier, and Deep Learning (ANN) are used to build the model. The result shows that Decision Tree and Random Forest achieve hypothyroidism prediction with an accuracy of 99.5758% and 99.3107%, respectively.

Naeem, et al. [8] proposed hypothyroidism disease prediction using the classification method. K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine are used to build a model for hypothyroidism disease prediction. The result shows that Support Vector Machine achieves an accuracy of 84.72% when compared to other classifiers.

Kumar, et al. [9] proposed hypothyroid prediction using the classification method. Decision Tree and Neural Networks are used to build models for hypothyroid prediction. The result shows J48 can predict hypothyroid disease with an accuracy of 99.58%.

Gothane [10] proposed a method to predict hypothyroidism using the classification method. ZeroR classifier is used to build a model for hypothyroidism prediction. The result shows that this model can predict hypothyroidism with an accuracy of 92.2853 %.

Hongboonmee and Trepanichkul [12] proposed a comparative performance of classification to analyze risk factors that affect hypothyroid disease prediction with data mining techniques. There are 3 techniques: Artificial Neural Network, Decision Tree, and Naïve Bayes with a dataset from Phitsanulok Hospital. The result shows that Artificial Neural Network achieves an accuracy of 82.97%.

Parimala and Vadivu [13] presented a method to predict thyroid disease using Support Vector Machine and K-Nearest Neighbor. This model can predict accuracy.

Lim, et al. [14] proposed a method to detect thyroid disease using the classification method. The dataset is used to prepare data using featurewiz to select features. Decision Tree, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Machine, and Ensemble machine learning algorithms (Random Forest and Extreme Gradient Boost) are used to build the model. The proposed model can detect thyroid disease with an accuracy of 99.45%.

Alyas, et al. [15] presented a method for thyroid disease prediction using Decision Tree, Random Forest, K-Nearest Neighbor, and Artificial Neural Networks. The result shows that Random Forest can predict thyroid disease with an accuracy of 94.8%.

Kurnaz, et al. [16] proposed a method for thyroid disease prediction using the classification method. The dataset is prepared using the Non-Sorting Genetic algorithm to select attributes for model building. Decision Tree, K-Nearest Neighbor, and Support Vector Machine are used to build a model for thyroid disease prediction. The result shows that this model can accurately detect thyroid disease.

Chaganti, et al. [17] presented a method for thyroid disease prediction using the classification method. The dataset is used to prepare for attribute selection and build the model using RF, GBM, ADA, LR, and SVM. The result shows that RF archives

thyroid disease prediction with an accuracy of 99% when compared to other methods.

Zhang, et al. [18] proposed a framework for comprehensively understanding thyroid disease using Association rule. The dataset is used to build a model with Apriori and the FP-Tree algorithms. The result shows that this model can create rules to detect thyroid disease.

3. Research Methodology

The model-building process consists of 3 processes to predict hypothyroid disease, as shown in Figure 1.

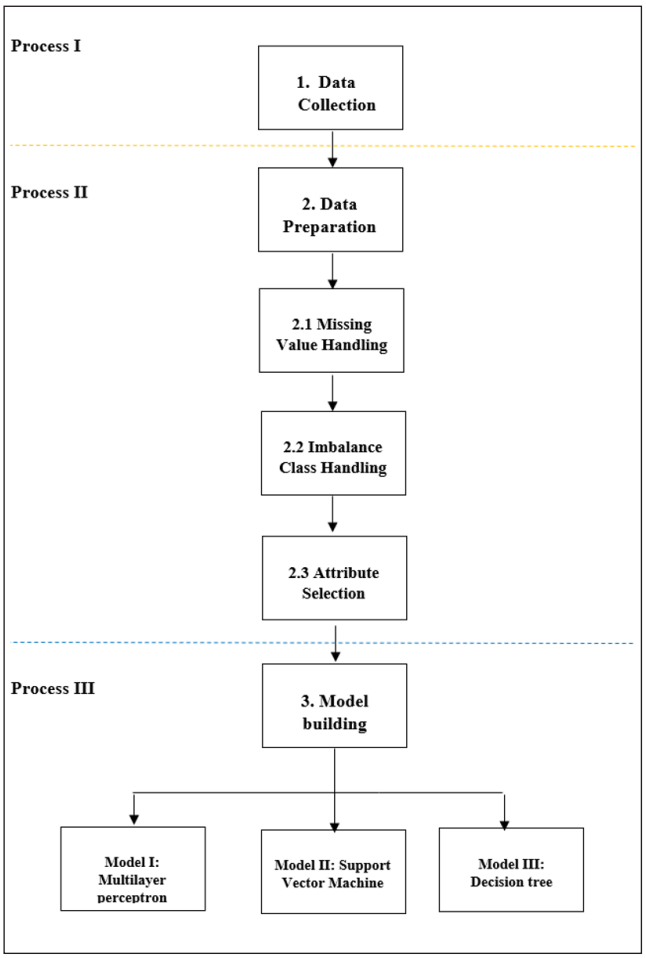

Figure 1. *The process of model building.*

From Figure 1, the model-building process consists of 3 processes: data collection, data preparation, and model-building, as explained in the next step.

3.1 Data Collection

This research uses a thyroid disease dataset, which is open data from the UCI repository. Thyroid disease data set collects from Garavan Institute in Sydney, Australia [19] using data of "allhypo". There are 30 attributes (29 attributes and 1 target). There is a total of 2,800 rows [20].

3.2 Data Preparation

Once we get data, we prepare data, which consists of 3 steps as follows.

1) Missing Value Handling

From the dataset, we found that data has numerous missing values. We handle missing values using the mode for nominal and compute the mean for numeric [21]. In addition, we removed 2 attributes: query hyperthyroid TBG because it does not data 100% and referral source. Therefore, there are 27 attributes from 29 attributes (2800 rows).

2) Imbalance Class Handling

When we handle missing values, we handle imbalanced classes. We found that the target class consists of class 0 = 2580 and class 1 = 220. Therefore, we handle the imbalance class using SMOTE [21], which increases data by random from the original dataset to distribute data of small class sizes and reduce the overfitting of data. There are 2800 rows (equal number of classes) for model building.

3) Attribute Selection

Once, we handle the imbalance class. We have set experiments to select suitable attributes for model building. There are 2 methods for comparing the model performance by considering accuracy and processing time. The first method is all attribute selection. The second method is attribute selection using ClassifierSubsetEval with BestFirst [21]. From our experiment, the result shows that suitable attribute for model building is shown in Table 1.

Table 1. *The comparison results of suitable attribute selection for model building.*

Method	Accuracy	Processing time
All attributes (27 attributes)	99.43 %	7.72 seconds
our experiment result (16 attributes)	99.61 %*	0.02 seconds*

*high accuracy and less processing time

From Table 1, our experiment result of suitable attribute selection for model building consists of 16 attributes such as on thyroxine, query on thyroxine, sick, pregnant, thyroid surgery, I131 treatment, lithium, goiter, tumor, TSH measured, TSH, T3, TT4 measured, T4U, FTI measured, and TBG measured. These attributes have high accuracy and take less time to process when compared to all attribute selections. Therefore, we use these attributes for model building.

3.3 Model Building

After, we prepared the data. We build a model to predict hypothyroid disease by comparison of classification methods such as Multilayer Perceptron, Support Vector Machine, and Decision Tree [21], [22]. These techniques were selected to build the model because they are easy to interpretability, robust, fast to process, suitable to complex problems, and high accuracy. Therefore, we use these techniques for model building. Each classifier sets the parameters as follows.

For Multilayer Perceptron, we have experimented to define the best parameter for high-performance model building. We set learning rate = 0.3, momentum = 0.2, hidden layer node = 9, input node = 16, and output node = 2 (yes, no), as shown in Figure 2.

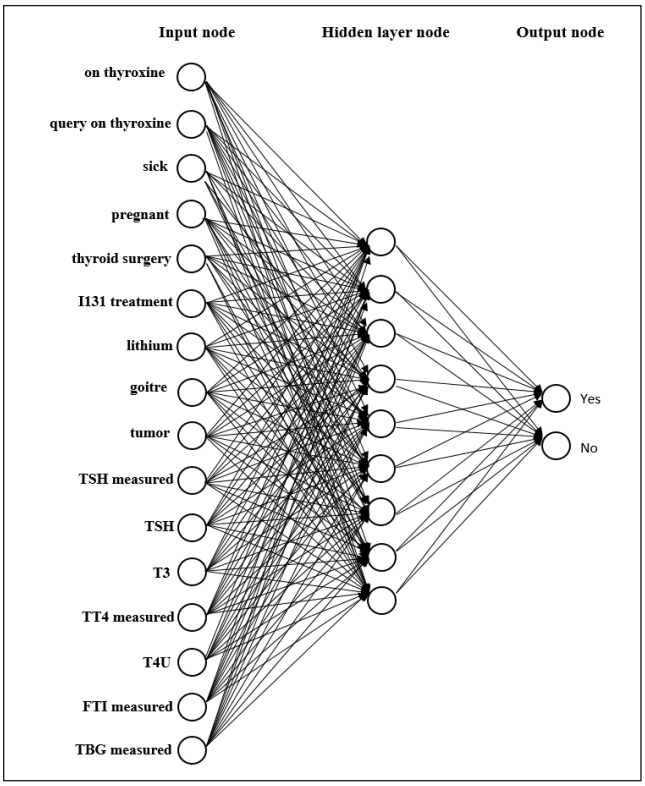

Figure 2. *Model for hypothyroid disease prediction using MLP.*

For the Support Vector Machine, we have experimented in several cases. The result shows that the high-performance model is a kernel function set (polynomial kernel).

For Decision Tree, we have experimented to build the model. The result shows that the suitable parameter is J48, confidence factor = 0.001, and seed = 10.

3.4 Model Performance Measurement

For performance measurement of the model, we use a confusion matrix (Table 2) to compute measurements such as Accuracy, Precision, Recall, and F-measure [21], [22] as shown in Equation 1-4.

Table 2. *Confusion matrix [21].*

Actual result / Prediction result	True	False
True	TP	FP
False	FN	TN

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

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

$$\text{Accuracy} = \frac{TP + TN}{TP+FN+TN+FP} \quad (3)$$

$$\text{F-measure} = \frac{(\text{Precision x Recall}) \text{ x } 2}{(\text{Precision+Recall})} \quad (4)$$

4. Results and Case Study

4.1 Experimental Result and Discussions

After model building for hypothyroid disease prediction using data mining techniques, we compared model performance with 3 techniques: Multilayer Perceptron, Support Vector Machine, and Decision Tree using a dataset from sample data of a total of 2,800 rows (16 attributes), which is described in section 3.2. The dataset is divided into a training dataset and a testing dataset with 5-fold cross-validation. In the model processing, we use an Intel core i5 Central Processing Unit (CPU) @3.10GHz and 4 GB RAM processing platform. The result is shown in Table 3.

Table 3. *The result of model performance measurement with our proposed model.*

	Precision	Recall	F-measure	Accuracy
Multilayer Perceptron	0.963	0.965	0.962	96.46 %
Support Vector Machine	0.929	0.929	0.903	92.93%
Decision Tree	0.996	0.996	0.996	99.61 %*

*high accuracy

From Table 3, Decision Tree model achieves an accuracy of 99.61% and uses a processing time of 0.02 seconds when compared to other models. Multilayer Perceptron and Support Vector Machine have accuracy of 96.46% and 92.93%, respectively. The result of the model performance comparison shows that Decision Tree can effectively predict in several cases such as on antithyroid medication and thyroxine. This model can detect accuracy. Therefore, we selected Decision Tree model to develop a prototype system for hypothyroid disease prediction because it has high accuracy when compared to other models and conforms to our hypothesis.

In addition, we compared our proposed model and state-of-the-art, which uses similar techniques to our proposed method. Guleria, et al. [7] use a public dataset of hypothyroidism. The dataset is extracted feature to select appropriate attributes for model building. Decision Tree, Random Forest, Naïve Bayes Multiclass classifier, and Deep Learning based on ANN are used to build the model. Also, Kumar, et al. [9] use a dataset from the UCI repository. The dataset's features are extracted using the dimensionality reduction technique. Appropriated attributes consist of 12 attributes such as On thyroxine, Pregnant, TSH measured, TSH, Goitre, T3, TT4 measured, TT4, Query hypothyroid, Thyroid surgery, and FT1. Decision Tree and Neural Networks are used to build models. Lim, et al. [14] use a dataset (demographic data, the characteristics, the mediations and treatments, different types of hormoneindexes, referral sources) from Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. The dataset is extracted features using Featurewiz. Decision Tree, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Classifier, and ensemble machine learning algorithms (Random Forest and Extreme Gradient Boost) are used to build the model. The resulting model performance comparison is shown in Table 4.

Table 4. *The result of model performance comparison.*

	Guleria, et al. [7]	Kumar, et al. [9]	Lim, et al. [14]	Our proposed model
Accuracy	99.5758%	99.58%	99.45%	99.61 %*
Precision	NA	NA	NA	0.996
Recall	0.996	NA	NA	0.996
F-measure	NA	NA	0.99	0.996

*high accuracy

From Table 4, the result shows that our proposed model is also high performance when compared to previous research. In conclusion, our proposed model can be applied to predict hypothyroid disease, which is beneficial for physicians to support decisions in diagnosis for further treatment.

4.2 Case Study

In this section, we demonstrate a case study to show the process of our prototype system. We randomly selected the test dataset in 1 case, as shown in Figure 3.

From Figure 3, we test data in our prototype system to predict hypothyroid disease. When the system is processed, the result shows that this case study has a chance of hypothyroid disease, as shown answer "Yes". Therefore, the user can acknowledge for preliminary risk estimation.

To make it easier to understand for users and physicians, we proposed data visualization to present demographic information, as shown in Figure 4.

From Figure 4, data visualization shows demographic information to support the decision of preliminary risk estimation of a chance of hypothyroid disease for the user before deciding to meet a physician to reduce the cost of medical and receive timely treatment. For example, the user is female and age 65 years. She takes thyroxine, is pregnant, and queries hypothyroid. From these information considerations, the user has a preliminary risk of a chance of hypothyroid disease. Therefore, the user should hurry up to meet the physician for further diagnosis.

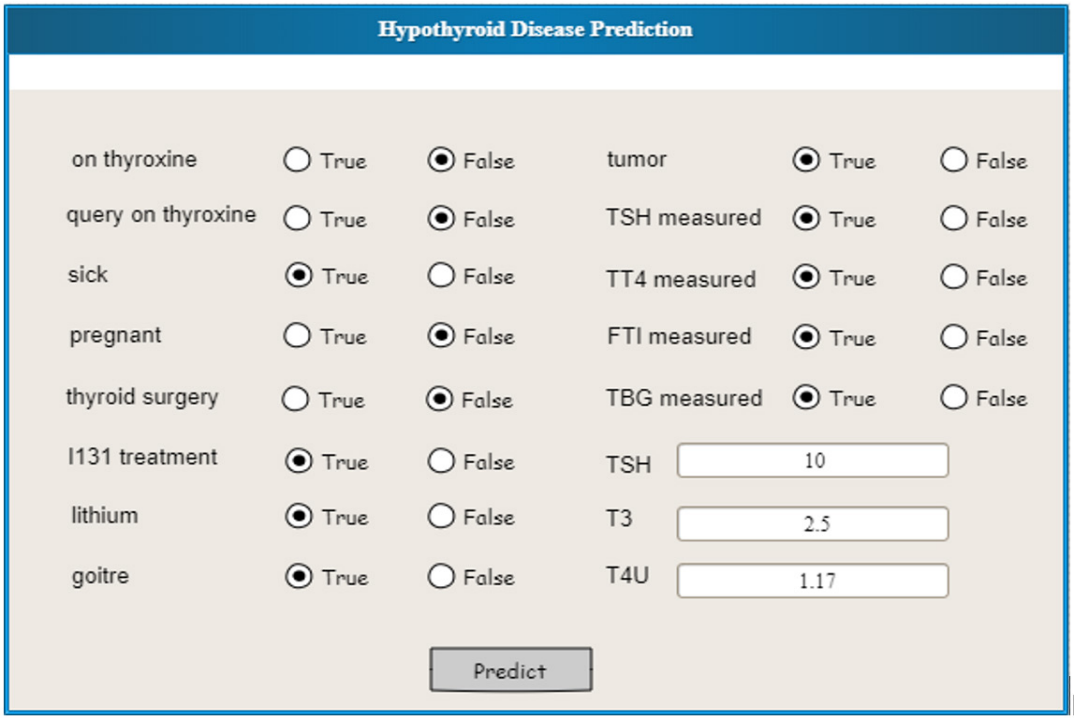

Figure 3. *Case study of hypothyroid disease prediction.*

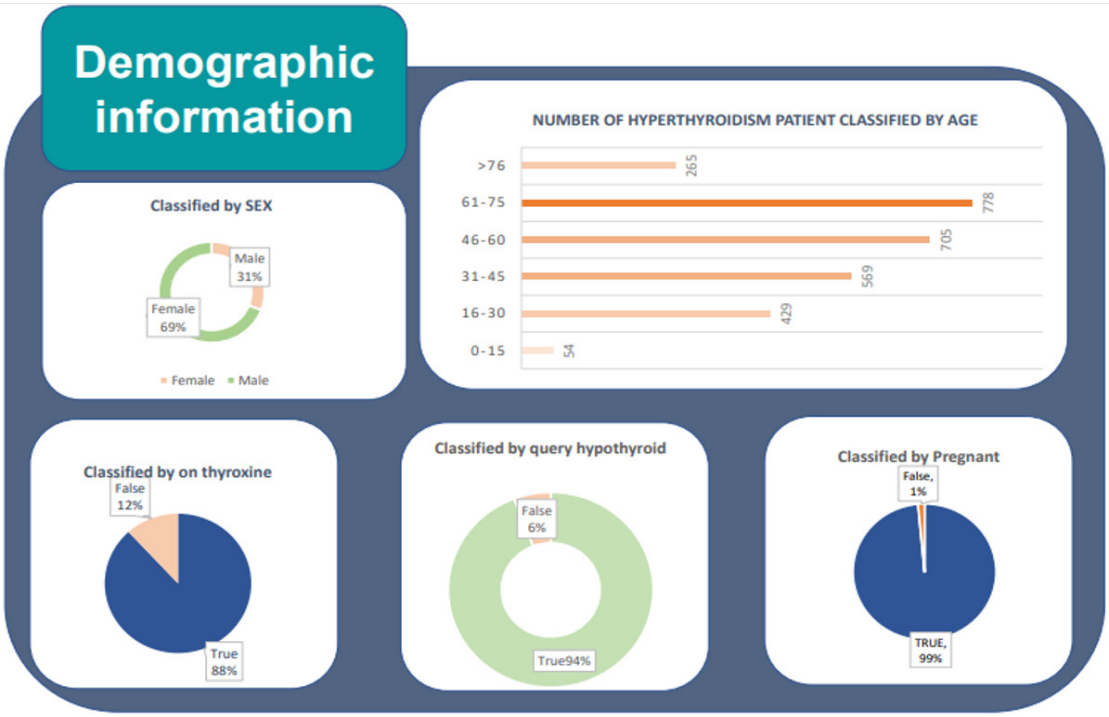

Figure 4. *Data visualization for decision support to detect hypothyroid disease.*

5. Conclusion

In this paper, we proposed a model for hypothyroid disease prediction using Multilayer Perceptron, Support Vector Machine, and Decision Tree based on a thyroid disease dataset. Firstly, the dataset is prepared to handle missing values. Secondly, the dataset deals with the imbalanced class. Thirdly, the dataset is selected as a suitable attribute for model building. Lastly, Multilayer Perceptron, Support Vector Machine, and Decision Tree are used to build models. From our experiment, Decision Tree has a high accuracy of 99.61% and uses a processing time of 0.02 seconds compared to other classifications. Multilayer Perceptron has an accuracy of 96.46%, and Support Vector Machine has an accuracy of 92.93%. In addition, we compared our proposed model to previous research. Guleria, et al. [7], Kumar, et al. [9], and Lim, et al. [14] use a similar technique to our proposed model. The result shows that Decision Tree also has high accuracy. Therefore, we use Decision Tree to develop a prototype system for hypothyroid disease prediction to support the decision of physicians for diagnosis and treatment to reduce the death rate. Moreover, we proposed data visualization for users to estimate the preliminary risk of a chance of hypothyroid disease before deciding to meet a physician to reduce costs and receive timely treatment.

In the future, we plan to develop the system for real use to reduce the cost and risk of a chance of hypothyroid disease in the severe stage.

6. References

[1] National Institute of Diabetes and Digestive and Kidney Diseases, *Hypothyroidism (Underactive Thyroid)*. Available Online at https://www.niddk.nih.gov/health-information/endocrine-diseases/hypothyroidism, accessed on 27 April 2024.

[2] O. M. Ahmed and R. G. Ahmed. "Hypothyroidism." *In Tech Open Access Publisher,* Chapter 1, pp. 1-20, 2015.

[3] O. Patsadu, K. Thepmee, K. Phuengliam, and M. Sirimongkol. "Abnormal Gait Pattern Recognition of Stroke Patient in Initial Stage Using Smartphone and Hybrid Classification Methods." *Information Technology Journal,* Vol. 18, No. 2, pp. 21-33, July-December, 2022.

[4] O. Chunhapran, D. Noolek, P. Labcharoenwongs, and T. Yampaka. "Multi-View Combination using Mutual Information and 3-D Euclidean Distance for Breast Cancer Classification." *Information Technology Journal,* Vol. 18, No. 2, pp. 44-54, July-December, 2022.

[5] N. Nonsiri, R. Manassila, and K. Somkanta. "Data Classifying to Diagnose Diabetes Risk Using Data Mining Techniques." *The Journal of King Mongkut's University of Technology North Bangkok,* Vol. 33, No. 2, pp. 538-547, April-June, 2023.

[6] J. A. Olalekan, F. Ogwueleka, and P. O. Odion. "Effective and Accurate Bootstrap Aggregating (Bagging) Ensemble Algorithm Model for Prediction and Classification of Hypothyroid Disease." *International Journal of Computer Applications,* Vol. 176, No. 39, pp. 40-48, July, 2020.

[7] K. Guleria, S. Sharam, S. Kumar, and S. Tiwari. "Early Prediction of Hypothyriodism and Multiclass Classification Using Predictive Machine Learning and Deep Learning." *Measurement: Sensors,* Vol. 24, pp. 1-7, December, 2022.

[8] A. B. Naeem, B. Senapati, A.S. Chauhan, M. Makhija, A. Singh, M. Gupta, P.K. Tiwari, and W. M. F. Abdel-Rehim. "Hypothyroidism Disease Diagnosis by Using Machine Learning Algorithms." *International Journal of Intelligent System and Applications in Engineering,* Vol. 11, No. 3, pp. 368-373, July, 2023.

[9] R. P. R. Kumar, M. S. Lakshmi, B. S. Ashwak, K. Rajeshwari, and S. Zaid. "Thyroid Disease Classification using Machine Learning Algorithms." *E3S Web of Conferences,* Vol. 391, pp. 1-7, June, 2023.

[10] S. Gothane. "Data Mining Classification on Hypo Thyroids Detection: Association Women Outnumber Men."

International Journal of Recent Technology and Engineering, Vol. 8, No. 6, pp. 601-604, March, 2020.

[11] L. Chake, S. Razvi, I. M. Bensenor, F. Azizi, E. N. Pearce, and R. P. Peeters. "Hypothyroidism." *Nat Rev Dis Primers,* Vol. 8, No. 1, pp. 1-17, June, 2022.

[12] N. Hongboonmee and P. Trepanichkul. "Comparison of Data Classification Efficiency to Analyze Risk Factors that Affect the Occurrence of Hyperthyroid Using Data Mining Techniques." *Journal of Information Science and Technology,* Vol. 9, No. 1, pp. 41-51, January-June, 2019.

[13] S. Parimala and P. S. Vadivu. "Optimizing Thyroid Stage Classification System Using Enhanced Data Mining Algorithms." *Proceedings of 2nd International Conference on Mathematical Techniques and Applications,* India, 2022.

[14] S.T. Lim, K.W. Khaw, X. Chew, and W.C. Yeong. "Prediction of Thyroid Disease Using Machine Learning Approaches and Featurewiz Selection." *Journal of Telecommunication, Electronic and Computer Engineering,* Vol. 15, No. 3, pp. 9-16, July-September, 2023.

[15] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad. "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach." *BioMed Research International,* Vol. 2022, pp. 1-10, June, 2022.

[16] S. Kurnaz, M.S. Mohammed, and S. J. Mohammed. "A High Efficiency Thyroid Disorders Prediction System with Non-Dominated Sorting Genetic Algorithm NSGA-II as a Feature Selection Algorithm." *Proceedings of 2020 International Conference for Emerging Technology,* India, pp. 1-6, 2020.

[17] R. Chaganti, F. Rustam, I.D.L.T. Díez, J.L.V. Mazón, C.L., and I. Ashraf. "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques." *Cancers,* Vol. 14, No. 16, pp. 1-23, August, 2022.

[18] X. Zhang, V.C.S. Lee, and J.C. Lee. "Unveiling Thyroid Disease Associations: An Exceptionality-Based Data Mining Technique." *Endocrines,* Vol. 4, pp. 558-572, July, 2023.

[19] Q. Ross, *Thyroid Disease. UCI Machine Learning Repository.* Available Online at https://doi.org/10.24432/C5D010, accessed on 2 March 2024.

[20] UCI Machine Learning Repository, *Thyroid disease.* Available Online at https://archive.ics.uci.edu/dataset/102/thyroid+disease, accessed on 2 March 2024.

[21] J. Han and M. Kamber. *Data Mining Concepts and Techniques.* Morgan Kaufmann Publishers, Third Edition, 2012.

[22] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining.* Pearson Education, Limited, International Edition, 2014.