Journal of Basrah Researches (Sciences) **50**(2), 193 (2024)
DOI: https://doi.org/10.56714/bjrs.50.2.16

A Comparative Analysis of Machine Learning Models for Predicting Thyroid Disorders in Type 1 and Type 2 Diabetic Patients

Hiba O. Sayyid[1,*], **Salma A. Mahmood**[2], **Saad S. Hamadi**[3]

[1] Department of Computer Science, University of Basrah, College of Computer Sciences and Information Technology, Basrah, Iraq.
[2] Department of Intelligent Medical Systems, University of Basrah, College of Computer Sciences and Information Technology, Basrah, Iraq.
[3] Department of Internal Medicine, University of Basrah, College of Medicine, Basrah, Iraq.

ARTICLE INFO

Received 20 July 2024
Revised 30 August 2024
Accepted 30 September 2024
Published 31 December 2024

Keywords:

Thyroid Disorder Prediction, Type 1 Diabetes, Type 2 Diabetes, Machine Learning, Feature Importance, Random Forest, Imbalanced Data Handling, Random Under-Sampling.

Citation: H. O. Sayyid et al., J. Basrah Res. (Sci.) **50**(2), 193 (2024). DOI:https://doi.org/10.56714/bjrs.50.2.16

ABSTRACT

Machine learning (ML) is increasingly indispensable in modern medicine, particularly for disease prediction and improving patient outcomes. This study applies ML techniques to predict thyroid disorders in diabetic patients, a critical task given the frequent co-occurrence and complex interplay between these conditions. six ML classifiers namely Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB) were evaluated across three experiments on a local dataset: (1) a balanced dataset using Random Under-Sampling (RUS), (2) a subset of Type 2 diabetes (T2D) patients, and (3) a subset of Type 1 diabetes (T1D) patients. Random Forest classifier consistently outperformed other classifiers, achieving the highest accuracy (0.85) and F1-score (0.83) in the T2D-focused dataset and showing robust performance on the balanced dataset using RUS. These results highlight the suitability of Random Forest for deployment in clinical settings and underscore the importance of balancing techniques like RUS in improving predictive accuracy. However, challenges remain in predicting thyroid disorders among T1D patients due to the low prevalence of thyroid disorders in this group. The findings reinforce the potential of ML in advancing diagnostics and personalized care in diabetic populations.

1. Introduction

Machine learning (ML) is increasingly recognized as a powerful tool in healthcare, enabling the analysis of large datasets to predict disease outcomes, identify at-risk patients, inform clinical decision-making, personalize treatment plans, and improve patient outcomes. The prediction of

*Corresponding author email : Itpg.hiba.oudah@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq

comorbid conditions like thyroid disorders in diabetic patients is crucial due to the intricate relationship between diabetes and thyroid function.

Diabetes, a chronic condition characterized by high blood sugar levels, it occurs when the body does not produce enough insulin or is unable to use the insulin it produces effectively [1]. Diabetes can be categorized into primary diabetes and secondary diabetes [2]:

The primary diabetes includes two types, type 1 is an autoimmune response in which the body's immune system attacks and destroys insulin-producing beta cells in the pancreas. As a result, the body produces little or no insulin [3]. Type 2 is the most common, and it occurs when the body becomes insulin-resistant or the pancreas inability to produce enough insulin [4].

Secondary diabetes occurs due to another medical condition or the use of certain medications that affect insulin production or effectiveness. Common causes of secondary diabetes include pancreatic diseases hormonal imbalances, certain medications, and genetic disorders [2].

Thyroid disorders are medical conditions that affect the function of the thyroid gland. The most common thyroid disorders include hypothyroidism and hyperthyroidism [5]. Hypothyroidism is when the thyroid gland does not produce enough thyroid hormones. This can slow down the body's metabolism and lead to various symptoms [6]. Hyperthyroidism is when the thyroid gland overproduces thyroid hormones leading to an elevated metabolism rate [7].

Thyroid disorders, including hypothyroidism and hyperthyroidism, frequently occur in patients with diabetes and affect metabolic regulation [7]. Highlighting the significance of treating thyroid disorders in people with DM to improve their prognosis. Thyroid dysfunction, if left untreated, can severely affect the metabolic control of diabetic patients, leading to major consequences in the outcome of both conditions [8]. Diabetes mellitus and thyroid diseases often coexist, with both Type 1 and Type 2 diabetes showing a higher prevalence of thyroid dysfunction compared to non-diabetic individuals. People with Type 1 diabetes are more likely to develop autoimmune thyroid disorders [9],[10]. Factors like poor glycemic control in diabetic patients can cause impaired T4 to T3 conversion and low T3 levels, affecting thyroid function. Higher insulin levels in type 2 diabetes mellitus associated with insulin resistance could stimulate thyroid tissue development and contribute to thyroid dysfunction. Sex, central obesity, elevated HbA1c levels, nephropathy, and the duration of diabetes, especially beyond five years, are identified as risk factors for thyroid dysfunction in individuals with diabetes [5]. Thyroid dysfunction has been linked to Type 2 Diabetes Mellitus (T2DM) in multiple studies, indicating a bidirectional relationship. High insulin levels in prediabetes and early-stage type 2 diabetes can cause thyroid enlargement, nodule formation, and consequences such as diabetic nephropathy, retinopathy, and peripheral neuropathy. Changes in serum TSH and thyroid hormones are associated with glycemic control and cardiovascular events [11].

Diabetes and thyroid diseases often have similar symptoms, making diagnosis and management difficult. For example, these disorders might result in weariness, weight changes, and mood swings. Both diseases share some common risk factors, such as age, sex, obesity, and family history. These factors can impact the development of both diseases. Overall, the relationship between diabetes and thyroid diseases is multifaceted and complicated, with interactions across hormones, immune responses, and metabolic pathways.

Despite the known association between diabetes and thyroid disorders, there is a lack of predictive models specifically designed to identify thyroid disorders among diabetes patients. The application of machine learning techniques to this problem can provide more accurate predictions, leading to better management strategies, and potentially reducing complications and healthcare costs. This research is motivated by the need to bridge this gap by developing and evaluating the effectiveness of various ML algorithms in predicting the presence of thyroid disorders among diabetic patients to provide a comprehensive tool for clinicians to identify high-risk patients for improved diagnostic processes and personalized appropriate treatment plans.

We conducted three experiments to determine the best-performing model: (1) using a balanced dataset achieved through RandomUnderSampler (RUS), (2) focusing on Type 2 Diabetes (T2D) patients, and (3) focusing on Type 1 Diabetes (T1D) patients. By comparing the performance of six ML classifiers—Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support

Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB)—this study seeks to identify the most reliable model for deployment in clinical settings.

The following section explores studies related to the current research, highlighting their methodologies, findings, and relevance.

2. Literature Review

Posonia et. al. 2020 [12] proposed Decision Tree J48 classification method for diabetes prediction, the method applied to the Pima Indians Diabetes Database that considered 768 pregnant women with diabetes. It consists of eight features, such as the number of times pregnant, plasma glucose concentration 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ml), body mass index kg/(height in m)^2), diabetes pedigree function, age, and class variable as positive or negative values. The primary aim of this study was to classify gestational diabetics or non-gestational diabetics. The dataset is analyzed using the Weka tool. The result of this study shows that Decision Tree J48 calculation gives 91.2% efficiency with less processing time. The authors suggested that this research can be improved in the future by applying the feature selection method before training the model.

Dharmarajan et al. 2020 [13] discussed a study that performed thyroid disease diagnosis using machine learning Classification techniques such as Decision Trees, Support Vector Machine (SVM), and Naïve Bayes. The data was collected as blood samples from 500 thyroid patients. This proposed classification method gave the best accuracy of 99.89% in comparison to their previous work using the Decision tree method with (97.35%) accuracy.

Hassan et al. [14] 2020 classified diabetes mellitus patients using classification techniques such as the Support Vector Machine, Decision Tree, and K-Nearest Neighbors, the dataset that was used in the implementation of these different methods (Pima Indian Diabetic Dataset) which was collected from female patients over 21 years old. The performance of these applied techniques is determined by using performance measures such as precision, accuracy, Sensitivity, and Specificity. The results obtained proved that SVM has the highest accuracy of 90.23% and outperforms the decision tree and KNN. In the future, the authors plan to improve the performance and the accuracy prediction of their Classification techniques and test them with huge datasets.

Duggal et al. 2020 [15] presented many feature selection and classification methods to diagnose thyroid disease. Tree-Based Feature Selection, Univariate Selection, and Recursive Feature Elimination are the methods that were proposed for feature selection. The Dataset contains 7200 instances and 27 attributes and was obtained from The UCI Machine Learning Repository. The purpose of this dataset was for research, development, and experimental uses. by using Three classification methods: Random Forest, Support vector machine, and Naive Bayes. The results show that when using the feature selection method of Recursive Feature Elimination (RFE) the Support Vector Machine method is the most accurate method with an accuracy of 92.92%.

Yadav et al. 2020 [16] used three methods namely decision tree, random forest, and classification and regression tree (CART) to examine the thyroid disease dataset. They used the bagging ensemble technique to enhance the results of these classifiers. The experiment was done on a dataset of thyroid patients that has 3710 instances with 29 features. The prediction accuracy was calculated based on different num-fold and seed values. The results obtained that the accuracy of the decision tree, random forest tree, and extra tree is 98%, 99%, and 93%, respectively. Then, by developing a bagging ensemble method that combines the three basic tree classifiers and applies them to the same dataset, the results of the new method obtain a better accuracy of 100%. The suggested future work on this paper is identifying different factors that affect the thyroid dataset and testing them on different large datasets such as heart disease or diabetes dataset

Chaubey et al. 2021 [17] have presented a study of how to predict thyroid disease and highlighted how to apply machine learning methods as a tool for classification. For this study, a thyroid dataset that contains 215 instances and 5 attributes was taken from UC Irvin's knowledge discovery in the databases archive. By applying three machine learning methods Logistic regression, Decision trees, and KNN the result shows that the KNN classifier is a better method for this dataset to predict thyroid disease with 96.875%. accuracy.

Permana et al. 2021 [18] suggested using (C4.5) decision tree to predict diabetes disease and help doctors with early diagnosis. The research aims to identify the most effective variable of the many variables causing diabetes complications. in this study, the dataset used is an early-stage diabetes dataset derived from secondary data that is available at https://www.kaggle.com/singhakash/early-stage-diabetes-risk prediction-datasets. The dataset contains 520 instances with several variables (age, sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity). with 90.38 % accuracy, the result shows that polydipsia plays a huge role in diabetes since it is One of the most common symptoms of diabetics. This indicates that the proposed method model is very accurate.

Dudkina et al. 2021 [19] presented a study that is dedicated to handling the problem of Classification and detection of diabetes disease. the study focuses on developing a decision tree-based machine learning model to solve this problem. The model was tested by using a dataset that contains 768 instances of diabetes patients and 9 attributes from the Pima Indians Diabetes database to test the model. The conclusion was that they could get better accuracy by allocating more data for training the model. In this case, splitting the data by 50% for training and 50% for testing was the best option with 0.71 accuracy.

Samin Poudel 2021 [20] presented a study to test different 20 ML approaches performance such as K Nearest Neighbours (KNN), Naïve Bayes, Support Vector Machine (SVD), perceptron and robust deep neural networks, XGBoost, etc. The study is done on the Pima Indian Diabetes Dataset which has 768 instances with 8 attributes and one output variable with a 0 or 1 value. For this data set, the results showed that the Naïve Bayes method outperformed the other methods when considering the combined analysis of all evaluation metrics with 0.77 accuracy, 0.83 F-score, 0.80 precision, and 0.68 recall. This shows that using complex and computationally expensive methods does not always improve disease diagnosis accuracy.

Chaganti et al. 2022 [21] presented a method that focuses on the multi-class problems to predict thyroid disorders using five machine learning models including RF, SVM, AdaBoost (ADA), LR, and Gradient boosting machine (GBM), as well as three deep learning models. They created a dataset from the UCI thyroid disease datasets that contained 9173 patient records,31 features, and 6771 normal patient records with no sign of thyroid disease. The dataset was randomly balanced by taking 400 samples from the 6771 records, and at least 200 samples for the other classes. The results showed that when using the random forest classifier with the presented method it can achieve a 0.99 accuracy in predicting ten types of thyroid diseases. One significant limitation of this study is the small dataset size, which may affect the performance of deep learning models training.

Tasin I et al. 2022 [22] employed various machine learning methods, including decision tree, logistic regression, KNN, random forest, SVM, and ensemble techniques to determine which method will provide the best results in predicting diabetes. The study utilized a private dataset of 203 individuals from a local textile industry in Bangladesh, referred to as the RTML dataset. Additionally, the Pima Indian dataset was used for model training and comparison. To manage the imbalance classes SMOTE and ADASYN balancing techniques were used. The performance of the used classifiers was evaluated using metrics such as precision, recall, F1 score, AUC, and classification accuracy. The best performance was achieved using the XGBoost classifier with an ADASYN approach, yielding 81% accuracy, an F1 score of 0.81, and an AUC of 0.84.

Almost all of the existing studies have focused on predicting a single disease either diabetes or thyroid disorders, with limited to no research focusing on predicting thyroid disorders specifically in diabetes patients. This research aims to bridge this gap by developing robust ML models customized to predict thyroid disorders in diabetes patients.

3. Methodology

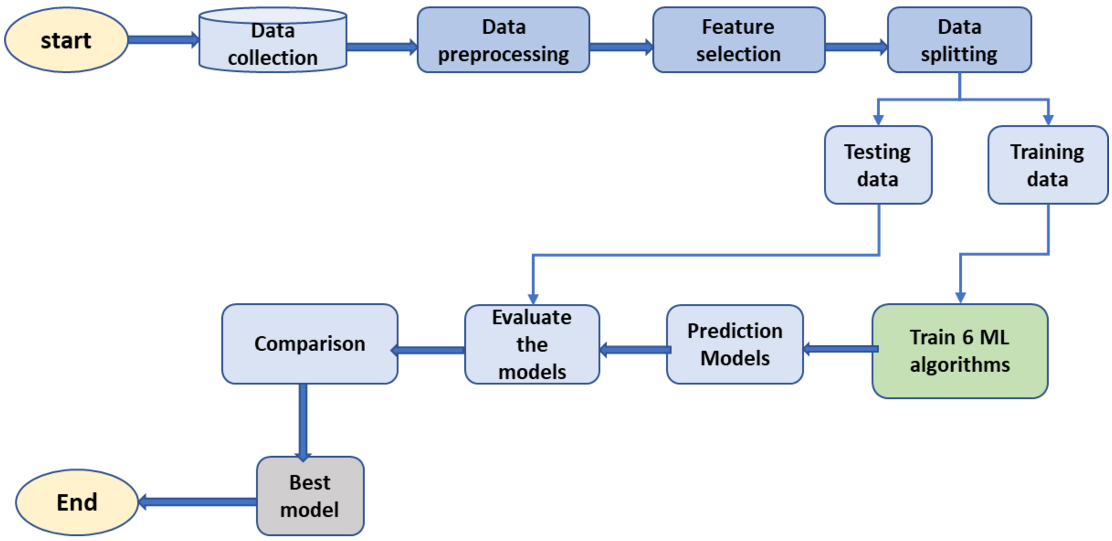

Fig. 1. Thyroid disorders predictions system workflow.

3.1. Data Collection

The dataset used for this study is a local dataset obtained from the Faiha Specialized Diabetes Endocrine and Metabolism Center (FDEMC) in Basra, Iraq. It consisted of records of diabetic patients with both Type 1 and Type 2 diabetes, with features such as age, sex, BMI, diabetes type, glycemic control, lipid control, pressure control, thyroid status indicating the presence or absence of thyroid disorders, and other relevant factors.

3.2. Data Preprocessing

Data preprocessing steps included handling missing values, outliers, and inconsistencies. Categorical variables such as sex, family history of DM, glycemic control, lipid control, pressure control, Thyroid, smoking status, drinking status, and marital status were encoded appropriately.

After the preprocessing steps, the final clean dataset consists of 44539 instances, with 6,755 labeled as having thyroid disorders. The dataset includes 12 features, those are thyroid, diabetes type, age, sex, family history of diabetes, Body Mass Index (BMI), glycemic control, lipid control, pressure control, smoking status, drinking status, and marital status.

3.3. Data Balancing

To tackle the issue of class imbalance, the RandomUnderSampler technique was employed. This method under-samples the majority class (non-thyroid disorder cases) to create a balanced dataset, which allows the ML models to focus equally on predicting both classes. This approach was used in experiment 1 to assess its impact on model performance. For Experiments 2 and 3, we filtered the balanced dataset to focus exclusively on T2D and T1D patients, respectively.

3.4. Feature Importance and Model Training

For feature importance, we used a Random Forest classifier to determine the most important of features (Risk Factors) in predicting thyroid disorders. The Random Forest model was trained on the dataset, and the feature importances were extracted to identify the most relevant features.

Using the features ranked by importance, we trained the ML method, using Stratified 10-Fold cross-validation. The process involved training the model across different folds and feature sets to determine the optimal number of features, avoid Overfitting and Underfitting problems, and to find the best-performing configuration (best model) based on training and testing accuracy. This approach was consistently applied across all six machine learning models: Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Naive Bayes.

3.5. Experimental setup

Three primary experiments were conducted to evaluate the performance of the models:

Experiment 1 (Balanced Data Using RUS): To address class imbalance, the Random Under-Sampling (RUS) technique was applied, resulting in a balanced dataset of 13,438 instances, with equal representation of both classes: thyroid disorder present and absent. The primary objective of this experiment was to assess the performance of the six ML models on a balanced dataset.

Experiment 2 (Type 2 Diabetes): The balanced dataset from Experiment 1 was further filtered to include only patients diagnosed with Type 2 Diabetes (T2D), yielding a balanced dataset of 11,648 instances. This experiment aimed to evaluate the performance of the machine learning models specifically within this subgroup, providing insights into their predictive capabilities for thyroid disorders among T2D patients.

Experiment 3 (Type 1 Diabetes): The balanced dataset was filtered to focus exclusively on patients with Type 1 Diabetes (T1D), resulting in a balanced subset of 1,790 instances. The goal of this experiment was to evaluate the performance of the machine learning models in predicting thyroid disorders within this specific group, where the occurrence of thyroid disorders is relatively less common compared to other diabetic populations.

3.6. Model Evaluation

Model evaluation is a crucial step in assessing the performance of the machine learning models, it is useful to understand how well the models generalize to new unseen data and identify areas for improvement. The models were evaluated based on accuracy, precision, F1-score, sensitivity, and specificity, ensuring a comprehensive comparison of their performance across different scenarios. These metrics equations are:

$$accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{1}$$

$$precision = \frac{TP}{TP+FP} \tag{2}$$

$$f1score = 2 * \frac{precision*recall}{precision+recall} \tag{3}$$

$$sensitivity\ (recall) = \frac{TP}{TP+FN} \tag{4}$$

$$specificity = \frac{TN}{TN+FP} \tag{5}$$

4. Results

4.1. Experiment 1: Balanced Data Using RUS (13,438 instances)

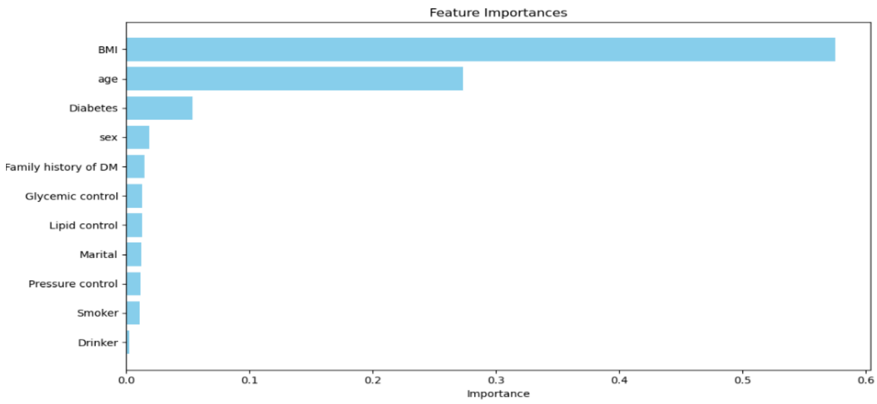

Fig. 2 Experiment 1 features importance ranking.

Table 1. Experiment 1 comparison results.

Classifier	Accuracy	Precision	F1-Score	Sensitivity (Recall)	Specificity
RF	0.84	0.96	0.82	0.713	0.967
DT	0.83	0.95	0.81	0.702	0.960
KNN	0.83	0.92	0.81	0.720	0.934
SVM	0.79	0.85	0.77	0.708	0.871
LR	0.78	0.84	0.76	0.701	0.868
NB	0.78	0.84	0.76	0.701	0.868

The feature importance ranking, shown in Figure 2, highlighted BMI and age as the most important factors in predicting thyroid disorders, followed by diabetes type and sex.

As shown in Table 1 in the first experiment, the Random Forest classifier achieved the highest accuracy of 0.84 and an F1-score of 0.82. The Decision Tree and KNN models also performed well, with accuracies of 0.83 and F1-scores of 0.81. SVM and LR models showed slightly lower performance, with accuracies of 0.79 and 0.78, respectively. Naive Bayes also performed comparably to LR. The results demonstrate that balancing the dataset effectively enhances model performance.

4.2. Experiment 2: (T2D) Patients (11,648 instances)

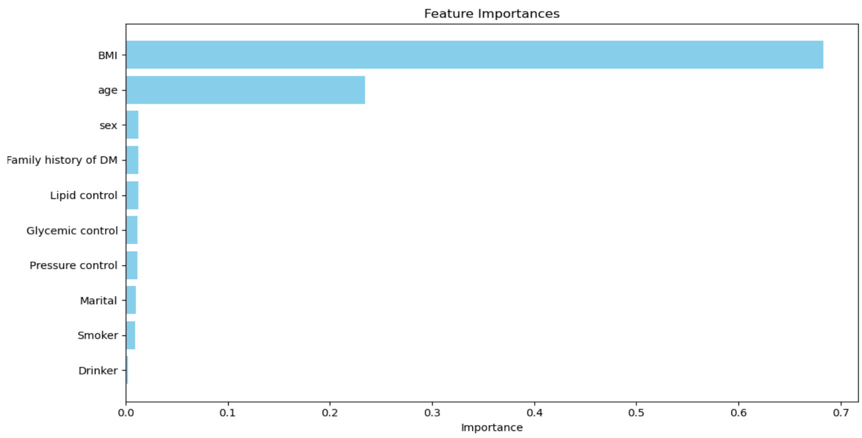

Fig. 3 Experiment 2 features importance ranking.

Table 2. Experiment 2 comparison results.

Classifier	Accuracy	Precision	F1 Score	Sensitivity (Recall)	Specificity
RF	0.85	0.99	0.83	0.719	0.991
DT	0.85	0.99	0.83	0.717	0.993
KNN	0.85	0.99	0.83	0.714	0.993
NB	0.84	0.96	0.82	0.708	0.973
SVM	0.83	0.99	0.80	0.678	0.991
LR	0.80	0.85	0.79	0.739	0.868

In the second experiment, the feature importance ranking in Figure 3 showed that BMI and age were the most influential factors for predicting thyroid disorders in Type 2 diabetes patients.

The models' performance in this Experiment was improved across all models with RF, DT, and KNN classifiers showing the best performance, each achieving an accuracy of 0.85 and F1-scores of 0.83. The Naive Bayes classifier performed slightly lower, with an accuracy of 0.84 and an F1-score of 0.82, slightly lower than the top models but still significantly improved from Experiment 1. This experiment underscored the model's capability to predict thyroid disorders effectively in T2D patients, where the prevalence is higher compared to T1D.

4.3. Experiment 3: (T1D) Patients (1,790 instances)

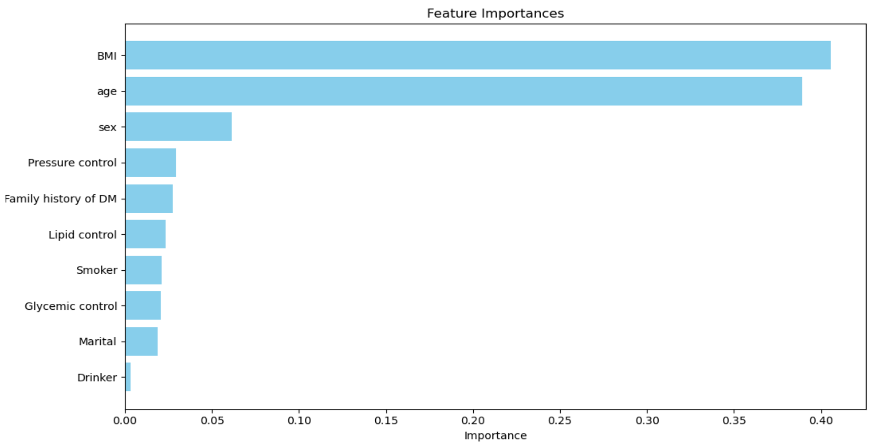

Fig. 4 Experiment 3 features importance ranking.

Table 3. Experiment 3 comparison results.

Classifier	Accuracy	Precision	F1 Score	Sensitivity (Recall)	Specificity
DT	0.66	0.74	0.61	0.511	0.820
RF	0.59	0.63	0.52	0.438	0.744
SVM	0.58	0.65	0.47	0.367	0.798
LR	0.57	0.57	0.59	0.611	0.528
NB	0.57	0.57	0.56	0.556	0.584
KNN	0.57	1.00	0.24	0.135	1.000

In the third experiment, which focused on Type 1 Diabetes patients, The feature importance analysis revealed BMI, age, and sex as the top predictors. The models struggled to achieve high accuracy. The Decision Tree classifier performed the best with an accuracy of 0.66 and an F1-score of 0.61. The Random Forest classifier followed with an accuracy of 0.59 and an F1-score of 0.52. The KNN model, which performed well in the previous experiments, showed poor results in this scenario, with an accuracy of 0.57 and an F1-score of only 0.24. These results highlight the challenges of predicting thyroid disorders in type 1 diabetes patients, where the disorder's prevalence is relatively low.

5. Discussion

The results from the three experiments demonstrate the importance of identifying key features and balancing datasets when predicting thyroid disorders in diabetic patients. Random Forest consistently outperformed other models, particularly in Experiment 1 (balanced data) and Experiment 2 (Type 2 diabetes subset), achieving the highest accuracy and F1 scores. Feature importance analysis showed that BMI and age were the most critical factors for thyroid disorder prediction in diabetic patients, aligning with medical literature that links body mass index and age with thyroid dysfunction. Diabetes type and sex also played important roles.

5.1. Impact of Data Balancing (Experiment 1)

Experiment 1 demonstrated that balancing the dataset using the RandomUnderSampler significantly enhances the predictive capabilities of ML models. The Random Forest classifier, in particular, benefitted from the balanced dataset, achieving the highest accuracy and F1 score among all tested models. This finding underscores the importance of addressing class imbalance, which is common in medical datasets, to improve model performance.

5.2. Type 2 Diabetes Focus (Experiment 2)

In Experiment 2, the focus on Type 2 Diabetes patients revealed that ML models, particularly the Random Forest, Decision Tree, and KNN classifiers, perform exceptionally well when the prevalence of thyroid disorders is higher. The consistently high accuracy and F1 scores across these models suggest that ML can be effectively deployed in clinical settings to predict thyroid disorders in T2D patients, thereby aiding in early diagnosis and improved patient management.

5.3. Challenges with Type 1 Diabetes (Experiment 3)

Experiment 3 highlighted the challenges associated with predicting thyroid disorders in Type 1 Diabetes patients due to the low prevalence of thyroid disorders in this group.

The lower performance across all models, particularly the Random Forest and KNN, indicates that additional factors, possibly related to the unique pathophysiology of T1D, need to be considered to enhance predictive accuracy. This finding suggests the need for further research into the specific characteristics of T1D that may impact thyroid function and how these can be integrated into ML models.

6. Conclusion

This study provides a comprehensive analysis of the application of machine learning methods to predict thyroid disorders in diabetic patients, emphasizing the importance of data balancing and the challenges posed by different diabetes types. Random Forest emerged as the most robust model, consistently delivering high accuracy and F1-scores across the balanced and Type 2 Diabetes-focused datasets, making it the most suitable model for deployment in predictive systems for thyroid disorders in diabetic populations. However, the difficulty in accurately predicting thyroid disorders in Type 1 Diabetes patients suggests that further research is needed. Future work will focus on exploring more advanced techniques for handling imbalanced data, testing the developed best-performing models, such as RF on new external datasets from different populations, and incorporating additional clinical features to enhance prediction performance.

Based on the results, Random Forest is recommended for deployment in clinical decision support systems aimed at managing thyroid disorders in diabetic populations.

The findings of this study highlight the critical role of ML in advancing medical diagnostics and improving patient care through early and accurate disease prediction.

References

[1] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” Procedia Computer Science, vol. 132, pp. 1578–1585, 2018, Doi: https://doi.org/10.1016/j.procs.2018.05.122.

[2] I. D. Penman, S. Ralston, M. W. J. Strachan, and R. P. Hobson, Davidson’s principles and practice of medicine, 24th Edition. [Edinburgh]: Elsevier, 2023.

[3] R. Kumar, P. Saha, S. Sahana, Yogendra Kumar, A. Dubey, and O. Prakash, “A Review On Diabetes Mellitus: Type1 & Type2,” World Journal of Pharmacy and Pharmaceutical Sciences, vol. 9, no. 10, pp. 838–850, Aug. 2020, Doi: https://doi.org/10.20959/wjpps202010-17336.

[4] S. A. Hassan, A.-K. M. Ali, and R. I. Saleem, “Relationship between glycemic control and different insulin regimens in pediatric type 1 diabetes mellitus,” The Medical Journal of Basrah University, 2023, Doi: https://doi.org/10.33762/mjbu.2023.140990.1138.

[5] S. U. Ogbonna and I. U. Ezeani, “Risk Factors of Thyroid Dysfunction in Patients With Type 2 Diabetes Mellitus,” Frontiers in Endocrinology, vol. 10, Jul. 2019, Doi: https://doi.org/10.3389/fendo.2019.00440.

[6] H. M. Almahshi, E. A. Almasri, H. Alquran, W. A. Mustafa, and A. Alkhayyat, “Hypothyroidism Prediction and Detection Using Machine Learning,” 2022 5th International Conference on Engineering Technology and its Applications (IICETA), May 2022, pp. 159–163. Doi: https://doi.org/10.1109/iiceta54559.2022.9888736.

[7] P. Sharma, S. Shrestha, and P. Kumar, “A review on association between diabetes and thyroid disease,” Santosh University Journal of Health Sciences, vol. 5, no. 2, pp. 50–55, Jan. 2020, Doi: http://doi.org/10.18231/j.sujhs.2019.013.

[8] B. Biondi, G. J. Kahaly, and R. P. Robertson, “Thyroid Dysfunction and Diabetes Mellitus: Two Closely Associated Disorders,” Endocrine Reviews, vol. 40, no. 3, pp. 789–824, Jan. 2019, Doi: https://doi.org/10.1210/er.2018-00163.

[9] H. Y. Abdulrazaq, I. A. Zaboon, and M. A. Maatook, “Prevalence of thyroid disorders among diabetes mellitus patients in al-Basra southern of Iraq,” Annals of Tropical Medicine and Public Health, vol. 24, no. 04, 2021. Available: https://www.researchgate.net/profile/Ibraheem Abood/publication/352135204_Abdulrazaq_et_al_2021_Prevalence_of_thyroid_disorder_among_diabetes_mellitus_patients_C_Annals_of_Tropical_Medicine_Public_Health_Prevalence_of_thyroid_disorders_among_diabetes_mellitus_patients_in_al/links/60ba801392851cb13d7984a7/Abdulrazaq-et-al-2021-Prevalence-of-thyroid-disorder-among-diabetes-mellitus-patients-C-Annals-of-Tropical-Medicine-Public-Health-Prevalence-of-thyroid-disorders-among-diabetes-mellitus-patients-in.pdf .

[10] S. Kalra, S. Aggarwal, and D. Khandelwal, “Thyroid Dysfunction and Type 2 Diabetes Mellitus: Screening Strategies and Implications for Management,” Diabetes Therapy, vol. 10, no. 6, pp. 2035–2044, Oct. 2019, Doi: https://doi.org/10.1007/s13300-019-00700-4.

[11] F. Rong et al., “Association between thyroid dysfunction and type 2 diabetes: a meta-analysis of prospective observational studies,” BMC Medicine, vol. 19, no. 1, Oct. 2021, Doi: https://doi.org/10.1186/s12916-021-02121-2.

[12] A. M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," 2020 3rd International Conference on Intelligent

Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 498-502, Doi: https://doi.org/10.1109/ICISS49785.2020.9316001.
[13] K. Dharmarajan, K. Balasree, A.S. Arunachalam, and K. Abirmai, “Thyroid Disease Classification Using Decision Tree and SVM,” Indian Journal of Public Health Research & Development, vol. 11, no. 03, pp. 229, Mar. 2020. Available: https://www.researchgate.net/publication/341742234_Thyroid_Disease_Classification_Using_Decision_Tree_and_SVM .
[14] A. S. Hassan, I. Malaserene, and A. A. Leema, “Diabetes Mellitus Prediction using Classification Techniques,” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 5, pp. 2080–2084, Mar. 2020, Doi: https://doi.org/10.35940/ijitee.e2692.039520.
[15] P. Duggal and S. Shukla, "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 670-675, Doi: https://doi.org/10.1109/Confluence47617.2020.9058102.
[16] C. Yadav and S. Pal, “Prediction of thyroid disease using decision tree ensemble method,” Human-Intelligent Systems Integration, vol. 2, no. 1–4, pp. 89–95, Apr. 2020, Doi: https://doi.org/10.1007/s42454-020-00006-y.
[17] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, “Thyroid Disease Prediction Using Machine Learning Approaches,” National Academy Science Letters, vol. 44, no. 3, pp. 233–238, May 2020, Doi: https://doi.org/10.1007/s40009-020-00979-z.
[18] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, “Classification of diabetes disease using decision tree algorithm (C4.5),” Journal of Physics: Conference Series, vol. 1869, no. 1, p. 012082, Apr. 2021, Doi: https://doi.org/10.1088/1742-6596/1869/1/012082.
[19] Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, “Classification and Prediction of Diabetes Disease using Decision Tree Method,” Symposium on Information Technologies & Applied Sciences, Bratislava, Slovakia , Mar. 2021. Available: https://ceur-ws.org/Vol-2824/paper16.pdf
[20] S. Poudel, “A Study of Disease Diagnosis Using Machine Learning,” Medical Sciences Forum, vol. 10, no. 1, p. 8, Feb. 2022, Doi: https://doi.org/10.3390/iech2022-12311.
[21] 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, vol. 14, no. 16, p. 3914, Aug. 2022, Doi: https://doi.org/10.3390/cancers14163914.
[22] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, “Diabetes prediction using machine learning and explainable AI techniques,” Healthcare Technology Letters, Dec. 2022, Doi: https://doi.org/10.1049/htl2.12039.

Journal of Basrah Researches (Sciences) **50**(2), 193 (2024)
DOI: https://doi.org/10.56714/bjrs.50.2.16

التعلّم الآلي للتنبؤ باضطرابات الغدة الدرقية لدى مرضى السكري: دراسة مقارنة

هبة عوده سيد [1,*]، سلمى عبد الباقي محمود [2]، سعد شاهين حمادي [3]

[1] قسم علوم الحاسوب، جامعة البصرة، كلية علوم الحاسوب وتكنولوجيا المعلومات، البصرة، العراق.

[2]قسم النظم الطبية الذكية، جامعة البصرة، كلية علوم الحاسوب وتكنولوجيا المعلومات، البصرة، العراق.

[3]قسم الطب الباطني، جامعة البصرة، كلية الطب، البصرة، العراق.

معلومات البحث

الاستلام	20 تموز 2024
المراجعة	30 آب 2024
القبول	30 أيلول 2024
النشر	31 كانون الأول 2024

الكلمات المفتاحية

التنبؤ باضطراب الغدة الدرقية، السكري من النوع الأول، السكري من النوع الثاني، التعلم الآلي، أهمية السمات، معالجة عدم توازن البيانات، تقنية التوازن العشوائي.

Citation: H. O. Sayyid et al. , J. Basrah Res. (Sci.) **50**(2), 193 (2024). DOI:https://doi.org/10.56714/bjrs.50.2.16

الملخص

أصبح تعلّم الآلة (ML) أداة لا غنى عنها في الطب الحديث، خاصة في التنبؤ بالأمراض وتحسين نتائج المرضى. تركز هذه الدراسة على تطبيق طرق التعلم الآلي للتنبؤ باضطرابات الغدة الدرقية لدى مرضى السكري، وهي مهمة صعبة نظرًا للعلاقة المعقدة بين هذه الحالات والتكرار المتزايد لظهورها معًا. تم تقييم ستة مصنفات تعلم آلي وهي: الغابة العشوائية(RF) ، شجرة القرار(DT) ، الجيران الأقرب (KNN)، آلة المتجهات الداعمة(SVM) ، الانحدار اللوجستي(LR) ، ونايف بايز (NB)، عبر ثلاثة تجارب على مجموعة بيانات محلية: (1) مجموعة بيانات متوازنة باستخدام أخذ العينات العشوائية تحت العينة(RUS) ، (2)مجموعة فرعية من مرضى السكري من النوع الثاني(T2D) ، و(3) مجموعة فرعية من مرضى السكري من النوع الأول .(T1D) تفوق مصنف الغابة العشوائية بشكل مستمر على المصنفات الأخرى، حيث حقق أعلى دقة (0.85) وأعلى درجة F1 (0.83) في مجموعة بيانات مرضى السكري من النوع الثاني، وأظهر أداءً قوياً على مجموعة البيانات المتوازنة باستخدام RUS .تؤكد هذه النتائج على ملاءمة استخدام الغابة العشوائية في القطاع الطبي ، وتبرز أهمية تقنيات التوازن مثل RUS في تحسين دقة التنبؤ. ومع ذلك، تبقى التحديات في التنبؤ باضطرابات الغدة الدرقية بين مرضى السكري من النوع الأول بسبب الانتشار المنخفض لهذه الاضطرابات في هذه المجموعة. تدعم هذه النتائج إمكانيات التعلم الآلي في تحسين التشخيص والرعاية الشخصية بين مرضى السكري.

[*]**Corresponding author email :** Itpg.hiba.oudah@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq