Comparison of Machine Learning Algorithms for Predicting Thyroid Disorders in Diabetic Patients

Hiba O. Sayyid[*1], Salma A. Mahmood[2], and 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
E-mail: Itpg.hiba.oudah@uobasrah.edu.iq, Salma.mahmood@uobasrah.edu.iq, and Saad.shaheen@uobasrah.edu.iq
*Corresponding author

Keywords: machine learning, classification, decision tree, random forest, support vector machine, naïve bayes, logistic regression, K-nearest neighbor, diabetes, thyroid disorders

Received: August 17, 2024

Machine Learning (ML), a subfield of Artificial Intelligence (AI), has been used successfully in the healthcare domain for disease diagnosis. Thyroid disorders and diabetes are two of the most prevalent and interconnected chronic diseases, as both play critical roles in regulating various physiological processes in the body. This study aims to predict thyroid disorders in diabetes patients using six machine learning algorithms: Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM). A locally sourced dataset comprising 44,539 instances of diabetic patients was utilized, undergoing preprocessing steps including data cleaning, encoding, and balancing. Two balancing techniques were employed: manual balancing and RandomUnderSampler. The dataset was partitioned into training and testing sets using a Stratified K-Fold cross-validation approach with 10 folds to ensure robust evaluation. Each algorithm's performance was assessed using metrics such as accuracy and F1-score. Among the models, the RF algorithm outperformed the others, achieving the highest accuracy of 95% on the manually balanced dataset and 84% when the RandomUnderSampler technique was employed. Additionally, the F1-scores for RF were 95% and 82%, respectively, indicating its robustness in handling imbalanced datasets. This study highlights the importance of selecting appropriate preprocessing techniques and machine learning methods for healthcare datasets. The findings can assist healthcare providers in making early diagnoses and interventions for thyroid disorders in diabetic patients, potentially improving their quality of life and overall healthcare outcomes.

Povzetek: Opisana je uporaba strojnega učenja za napovedovanje motenj ščitnice pri bolnikih s sladkorno boleznijo. Algoritem naključnih gozdov doseže najvišjo točnost in oceno F1 na uravnoteženem naboru podatkov.

1 Introduction

Diabetes and thyroid disorders are among the most prevalent chronic diseases affecting the endocrine and metabolic systems [1]. These two diseases are often coexisted and strongly linked together, as many studies have shown that there is a higher prevalence of thyroid disorders in diabetic patients and vice versa [2].

Diabetes is a chronic condition that is caused by elevated levels of blood sugar (glucose) [3]. This occurs when the body either cannot use the insulin it produces effectively or cannot produce enough insulin. Insulin is a hormone that allows the body cells to absorb and use glucose for energy and helps regulate blood sugar [4]. As a result, diabetes affects various body functions. There are four types of this disease

- Type 1 diabetes is an autoimmune disease which is usually diagnosed in children and young adults [5], it occurs when the insulin-producing cells of the
- pancreas is attacked by the immune system which leads to little or no insulin [6].
- Type 2 diabetes is the most common type of diabetes that often occurs in older adults when the body doesn't produce enough insulin or becomes resistant to insulin [6].
- Gestational diabetes this type develops as a complication in women during pregnancy and usually goes away after the baby is born [7].
- There are fewer common types of diabetes that are caused by genetic conditions and diseases such as secondary diabetes and monogenic diabetes.

Thyroid disorder is a disease that affects the function of the thyroid gland in producing the appropriate amounts of thyroid hormones T3 (tri-iodothyronine), and T4(tetra-iodothyronine), as these hormones play an important role in controlling many vital activities of the body such as heart rate, energy level, metabolism, bone health, and many other functions. The most common thyroid disorders are Hyperthyroidism and Hypothyroidism [8]. In hyperthyroidism, the thyroid gland overproduces thyroid hormones. While in hypothyroidism the thyroid gland does not produce enough thyroid hormones [8].

Studies show that there is a higher prevalence of thyroid disorders among patients with type 1 or type 2 diabetes in comparison to non-diabetic patients, which reveals their close relationships, it also shows that autoimmunity is a key to understanding the link between type 1 diabetes and autoimmune thyroid disorders [9].

The presence of insulin resistance or diabetes increases an individual's risk of developing thyroid disorders while having thyroid disorder can increase the risk of developing diabetes and metabolic syndrome [10]. It is very important to diagnose thyroid disorder in diabetic patients and a routine screening should also be recommended. It is necessary that the clinician identify the high-risk diabetic groups and manage the thyroid abnormalities if present as soon as possible to reduce the risk of further complications [10].

The primary aim of this study is to assess the effectiveness of six machine learning methods (Decision tree, random forest, Support Vector Machine, Naïve Bayes, k nearest neighbor, logistic regression) in predicting the presence or absence of thyroid disorders in diabetes patients. By comparing the results of each method, we aim to identify the most accurate model to enhance early detection and intervention. Machine learning methods used in this study differ in their nature and work but they are all used for predicting new states.

Logistic regression (LR): is a classification machine learning algorithm that is used for predictive analysis based on the concept of probability [11]. LR classifies the data using the logistic sigmoid function. LR predicts one of two possible outcomes of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It does not give the value of either True or False, Yes or No, 0 or 1, etc. instead, it gives a probabilistic value between 0 and 1[12]. To classify instances into the two classes. a common approach is to use a threshold value (e.g., 0.5), If the predicted probability is above the threshold, the instance is assigned to one class, and if it is below the threshold, then it is assigned to the other class. LR is widely used for many tasks such as fraud detection, disease diagnosis, and prediction, Tumor Malignant or Benign, mail spam or not spam, etc. [11].

Naïve Bayes (NB): is a simple machine learning classification algorithm based on Bayes' Theorem [13]. It is called naïve because of the assumption of conditional independence among the features which means that the presence or absence of one feature in a class is independent of the presence or absence of the other features. It is used for a large amount of data. Bayes' theorem, Rule, or law is used to describe the probability of a hypothesis with existing knowledge. Bayes' theorem formula is [11]:

$$\mathbf{P}(\mathbf{A}|\mathbf{B}) = (\mathbf{P}(\mathbf{B}|\mathbf{A}) * \mathbf{P}(\mathbf{A})) / \mathbf{P}(\mathbf{B}) \quad (1)$$

NB is computationally efficient, easy to create, and can handle large datasets [12]. It is very effective in text classification tasks, such as spam filtering. Despite that it's a simple algorithm with the independence assumption, it can often outperform complex algorithms.

Decision Tree (DT): is one of the supervised Machine learning algorithms that is used for both classification and regression problems [14]. DT is a visual representation of the decision-making process, it's a tree-like graph that partitions the data based on the input features, the tree starts with a root which has the highest gain then nodes and branches. Where each node represents a test that follows the if-then statement and leads to a different branch, each branch leads to one outcome (decision) [15]. It is a widely used algorithm for predicting diseases.

K-Nearest Neighbors (KNN) is one of the simplest lazy learning machine learning algorithms that make predictions based on the entire data [12],[16]. The algorithm is used for solving classification and regression tasks. KNN assumes that similar data points are located near each other, the similarity is called distance. It uses distant measures like Euclidean to measure similarity. Although KNN is a very simple and easy-to-implement algorithm, its results can be very competitive [16].

Support Vector Machine (SVM) is one of the most popular machine learning algorithms. It is used primarily for classification tasks but can also be used for regression [12]. The main goal of SVM is to separate the data with a hyperplane into different classes so that we can easily put the new data point into one of the classes [11]. SVM can be effective in high-dimensional spaces and is widely used in image classification, Face detection, text categorization, and handwriting recognition.

Random Forest (RF) is a machine learning algorithm that belongs to the group of decision-tree-based methods [13]. It can be used for classification tasks and regression. Random forest is a collection of decision trees built during the training process and then the prediction of these trees is combined during the testing process. RF approach gives a better accurate result in comparison to a single decision tree with the ability to limit overfitting [16].

The organization of this paper is as follows, ….

2 Literature review

In the field of machine learning-based prediction of diabetes and thyroid disorders, numerous studies have explored various algorithms and methodologies. This section provides a structured comparison of these studies in terms of the algorithms used, evaluation metrics, and the reported results. By highlighting the strengths and limitations of previous works, we emphasize the novelty and contributions of the present study.

2.1 Diabetes prediction studies

Hassan et al. [17] applied SVM, K-Nearest Neighbors (KNN), and Decision Tree to classify diabetes patients. The study showed that SVM outperformed the other algorithms with the highest accuracy of 90.23%.

Samin Poudel [18] tested 20 machine learning algorithms for diagnosing diabetes based on the Pima Indian Diabetes Dataset. Naive Bayes emerged as the best-performing algorithm with an accuracy of 77%, an F1-score of 0.83, a precision of 0.80, and a recall of 0.86.

Dudkina T et al. [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 results showed that splitting the data by 50% for training and 50% for testing was the best option with 0.71 accuracy.

2.2 Thyroid disease prediction studies

Yadav D et al. [20] used Random Forest, Decision Tree, and Classification and Regression Tree (CART) to predict thyroid disease. The results showed that Random Forest achieved an accuracy of 99%, followed by Decision Tree 98% and CART 93%. Their ensemble approach combining these classifiers achieved a perfect accuracy of 100%.

Priyanka Duggal and Shipra Shukla [21] used feature selection and classification techniques like Naive Bayes, SVM, and Random Forest to diagnose thyroid disorders. The study reported that SVM achieved the highest accuracy with 92.92%.

Chaubey G. et al. [22] tested Logistic Regression, Decision Trees, and KNN for thyroid disease prediction. KNN achieved the highest accuracy at 96.88%.

Chaganti et al. [23] 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.

Table 1:Summary table

Study	Methodology	Algorithms Used	Key Evaluation Metrics	Results
Hassan et al (2020).	Classifying diabetes patients	SVM, KNN, DT	Accuracy	SVM: 90.23%
Samin Poudel (2021)	Diagnosing diabetes	20 ML approaches	Accuracy, Precision, Recall, F1-score	Naive Bayes: Accuracy 77%, F1-score 83%, Precision 80%
Dudkina T et al. (2021)	Classification and detection of diabetes disease	DT based model	Accuracy	DT:71%
Yadav et al. (2020)	Predicting thyroid disease	Random Forest, Decision Tree, CART	Accuracy	RF: 99%
Priyanka Duggal & Shipra Shukla (2020)	Diagnosing thyroid disorders	Naive Bayes, SVM, Random Forest	Accuracy	SVM: 92.92%
Chaubey G. et al. (2012)	Thyroid disease prediction	Logistic Regression, Decision Trees, KNN	Accuracy	KNN: 96.88%
Chaganti et al. (2022)	predicting thyroid disorders	RF, SVM, AdaBoost (ADA), LR, and Gradient boosting machine (GBM), as well as three deep learning models	Accuracy	RF: 99%
Current study	Predicting thyroid disorders in diabetic patients	RF, DT, SVM, KNN, NB, and LR	Accuracy, F1-Score, Precision, Recall, and Specificity	RF with Accuracy: 88%, F1-Score: 85%

From the table above, we can see that various studies have employed different algorithms to predict diabetes and thyroid disorders with varying results. For instance, SVM and Decision Tree techniques are commonly used in diabetes prediction, with SVM often yielding higher accuracy compared to other algorithms. On the other hand, for thyroid disease prediction, Random Forest and KNN have been reported to achieve remarkable accuracy, with Random Forest reaching up to 100% accuracy when combined with ensemble methods.

While these studies have contributed significantly to the field, there remains a gap in comprehensive and reliable approaches for predicting thyroid disorders specifically in the diabetic population. They often focus on either one disorder or use fewer evaluation metrics. Some studies rely primarily on accuracy, which may not reflect the model's true performance, especially when class imbalance exists. The F1-score and AUC metrics are more informative but have not been consistently used across studies.

The current study addresses these gaps by utilizing a comprehensive preprocessing pipeline that includes feature selection technique, and effective class imbalance handling using methods like RandomUnderSampler. Additionally, this study adopts a range of evaluation metrics (accuracy, F1-score, precision, recall, and specificity) to offer a well-rounded analysis of model performance. Furthermore, we compare multiple machine learning models RF, SVM, KNN, DT, NB, and LR using cross-validation, which not only strengthens the model evaluation but also ensures more robust generalization to unseen data.

By offering a balanced prediction model with high accuracy (88%) and F1-score (0.85), the current study surpasses previous works in terms of both the depth of analysis and the performance metrics, which positions it

as a significant advancement in predicting thyroid disorders in diabetic patients.

3 Proposed methodology

The main objective of this study is to predicate the relationship between diabetes mellitus and Thyroid disorders. Six different prediction methods were used for this purpose as aforementioned above. The proposal methodology is shown in the following Figure 1.

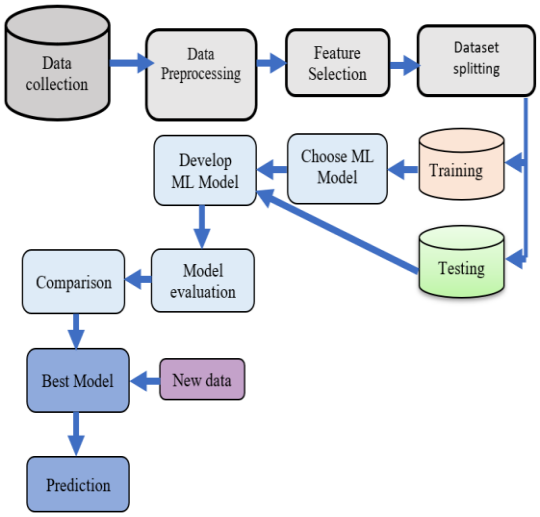

Figure 1: Flowchart of the thyroid disorder prediction system.

The above flowchart is illustrative of the following process:

3.1 Data collection

For this study, we used a medical dataset related to diabetes patients which was obtained from Faiha Specialized Diabetes Endocrine and Metabolism Center (FDEMC) in Basra, Iraq.

3.2 Data preprocessing

Data preprocessing was a critical step in preparing the dataset for effective machine learning model training and evaluation. This section elaborates on the detailed procedures used, including handling missing values, feature engineering, and encoding categorical variables ensuring replicability and transparency.

3.2.1 Handling missing values

Given the sensitivity of clinical data and the potential risks of introducing bias through imputation, instances with missing values were excluded from the dataset. This approach ensured the integrity and reliability of the analysis by working exclusively with complete data. While this reduced the dataset size, it maintained the accuracy required for clinical applications and minimized the risk of introducing errors associated with imputation.

After removing incomplete records, the dataset was carefully inspected to confirm that it remained representative of the original population in terms of key demographic and clinical features, ensuring that the removal process did not introduce unintended biases.

3.2.2 Data cleaning

Data cleaning is a critical process that significantly impacts the quality and reliability of predictive models. A clean dataset ensures accurate and robust machine learning models with improved performance and trustworthy predictions. In this study, thorough data cleaning was performed to address various issues and errors present in the dataset. The data cleaning process involved

- Identifying and rectifying incorrect data entries which included instances where ambiguous letters, words, and symbols were used such as ‘, \\, \L, \N,], B, E, EX, L, M, MN, N, N’, N N, NNNN, N\, N\], N\N, N], N] \, H, U, ص‌ى ,ة ,. Such values represent noises and inaccuracies in the dataset. Furthermore, inconsistencies in data entry were addressed, including the use of 'لا' in Arabic instead of 'No', as well as the recording of 'N' instead of 'No'. Additionally, discrepancies in capitalization were noted, such as 'female' being recorded instead of 'Female'. By rectifying these mistakes, the dataset was standardized, eliminating potential sources of error in the analysis.
- Handling the age field by determining ages in ranges (15-100 years), in line with the policy of the diabetes center catering to adults only.
- Similarly, filtering out heights and weights that fell outside the normal ranges. These actions were essential to preserve the integrity of the dataset and enhance the accuracy of our analyses.

3.2.3 Feature engineering

To enhance the performance of machine learning models, new features were derived from existing ones through feature engineering. For example, the Age feature was computed from the patients' dates of birth, and the Body Mass Index (BMI) was calculated using height and weight measurements. These newly created features provided additional insights into patient characteristics, which contributed to improving the predictive power of the models.

3.2.4 Encoding categorical variables

Since machine learning algorithms generally require numerical input, data encoding is essential to convert categorical variables into a suitable format. This study

used label encoding to transform variables such as sex, family history of DM, glycemic control, lipid control, pressure control, thyroid, marital status, smoker, and drinker into numerical representations compatible with machine learning models. After these steps the dataset consists of 44539 instant and 12 variables, Table 2illustrates each variable along with its corresponding encoded values.

Table 2: Description of the used data

Feature	Description	Value After Encoding
Thyroid	If the patient is diagnosed with a thyroid disorder.	0 means No 1 means Yes
DM	If the patient has type1 or type2 Diabetes Mellitus	1 for type1 2 for type2
Age	The patient's age in years	Range (15-100)
Sex	The patient's gender:	0 for male 1 for female
Family history of DM	If the patient has a family member with diabetes	0 means No 1 means Yes
BMI	Body Mass Index: the patient's weight divided by the square of height	Range (10.8-75.3)
Lipid control	The patient's lipid levels in the bloodstream are managed	0 means No 1 means Yes
Pressure control	The patient's blood pressure levels are managed to stay in a specific target range	0 means No 1 means Yes
Glycemic control	The patient's blood sugar levels are managed in a specific target range	0 means No 1 means Yes
Smoker	If the patient is a current smoker, non-smoker, or former smoker.	0 means No 1 means Yes 2 means X-smoker
Drinker	If the patient is a current drinker, non-drinker, or former drinker	0 means No 1 means Yes 2 means X-drinker
Marital	If the patient is married, single, divorced, or widowed.	0 means Single 1 means Married 2 means Divorced 3 means Widow

3.3 Addressing class imbalance

Class imbalance is a prevalent challenge in machine learning, especially in healthcare datasets where minority classes often represent critical conditions. In this study, the dataset was imbalanced, with only 15.17% of instances representing patients with thyroid disorders (6,755 instances), compared to 84.83% without thyroid disorders (37,784 instances). To address this imbalance, two techniques were employed.

3.3.1 Experiment 1: RandomUnderSampler (RUS)

In the first experiment, the RandomUnderSampler (RUS) technique was used to address the class imbalance. This method randomly reduces the size of the majority class to match that of the minority class, creating a balanced dataset. After applying RandomUnderSampler, the dataset was reduced to 13,438 instances, with an equal distribution of 50% representing patients with thyroid disorders and 50% without. While this approach ensures that the models are not biased toward the majority class, it can result in the loss of valuable information by discarding majority-class instances. Nonetheless, it was chosen for its simplicity and effectiveness in achieving balance without introducing synthetic data.

3.3.2 Experiment 2: manual balancing

In the second experiment, the dataset was manually balanced under the expert supervision of a physician to ensure the process was clinically valid and aligned with medical standards. The dataset was reduced to 2,166 instances, with an equal number of examples from both classes. Unlike RUS, manual balancing involved the careful selection of instances, allowing for greater control over the data distribution while preserving its clinical relevance. This approach mitigated the potential bias introduced by random sampling, ensuring that the balanced dataset reflected real-world clinical scenarios.

Although techniques such as RandomOverSampler (ROS), Synthetic Minority Over-Sampling Technique (SMOTE), and ensemble methods like Balanced Random Forest (BRF) are widely used for handling imbalanced data, they were not employed in this study. The primary concern was that synthetic data might fail to capture the true clinical variability of the minority class, potentially introducing artificial patterns that could distort model predictions and reduce generalizability. Additionally, these methods increase computational complexity and training time, making them less suitable for the objectives of this study. Instead, simpler and more controlled balancing methods were chosen to maintain a representative and manageable dataset.

3.4 Model selection and training

3.4.1 Model selection

In this study, we employed six machine learning algorithms to predict thyroid disorders in diabetic patients: Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM). These models were selected for their diverse characteristics and strengths in classification tasks, particularly in medical datasets. Allowing us to compare their performance in addressing the two different datasets. The rationale for selecting these models is summarized below:

- **Random Forest (RF)** was chosen for its ensemble nature, which combines multiple decision trees to reduce overfitting and improve generalization. RF is particularly effective in handling high-dimensional datasets with complex interactions. Additionally, RF provides feature importance rankings, offering insights into which factors contribute most to predictions.
- **Decision Tree (DT)** was selected for its simplicity, interpretability, and ability to model nonlinear relationships. Furthermore, DTs offer visual representations of decision rules, making them especially useful for understanding model behavior.

- **K-Nearest Neighbors (KNN)** was included due to its ability to perform well in non-linear decision boundaries by evaluating the proximity between instances. It is an intuitive algorithm that can be effective when there are clear clusters in the data.
- **Logistic Regression (LR)** was chosen for its simplicity, interpretability, and strong performance in binary classification tasks. As a linear model, LR serves as a robust baseline, helping to benchmark the performance of more complex approaches.
- **Naïve Bayes (NB)** was selected for its simplicity and efficiency in handling large datasets with categorical features. Its probabilistic nature makes it well-suited for classification tasks with independent features. particularly the Gaussian variant,
- **Support Vector Machine (SVM)** was chosen for its ability to find complex decision boundaries in high-dimensional spaces. It is particularly effective in separating classes with a clear margin.

3.4.2 Training

Initially, a Random Forest classifier was employed to determine the most influential features by ranking them based on their importance scores shown in Figure 2 and Figure 3. These top-ranked features were subsequently utilized for training the models.

All models were trained using Stratified K-Fold cross-validation with 10 folds, ensuring that the distribution of thyroid and non-thyroid patients was maintained in each fold. This method provides a robust evaluation of the models' performance by assessing them across multiple data splits, which helps mitigate the risk of overfitting or underfitting.

To enhance the feature selection process, we used a sequential feature selection approach, where we started by training each model with a single feature and incrementally added more features. This allowed us to identify the most relevant features for each model and ensured that only the most informative variables were used, optimizing the model's performance.

We assessed both training and testing accuracies to evaluate how well each model generalized to unseen data. By comparing these accuracies, we were able to detect potential overfitting (where the model performs well on training data but poorly on testing data) or underfitting (where the model performs poorly on both training and testing data). This evaluation ensured that the models maintained a balance between accuracy and generalization.

3.4.3 Hyperparameter tuning

Hyperparameter tuning was performed to optimize model performance. For RF and DT, fixed parameters such as max_depth=10 and n_estimators=100, were selected after experimenting with various combinations of parameter values. These experiments involved testing different depths for the trees and numbers of estimators to evaluate their impact on the model's performance. While the hyperparameter tuning for KNN involved testing different numbers of neighbors (1–10) and subsets of top features ranked by Random Forest importance, using 10-fold Stratified Cross-Validation to evaluate each combination. The optimal configuration was selected based on the highest cross-validation accuracy and minimal train-test accuracy differences, ensuring good generalization and minimizing the risk of overfitting or underfitting during cross-validation.

The combination of multiple models, cross-validation, sequential feature selection, and hyperparameter tuning ensured that we could rigorously evaluate the performance of each algorithm and select the one best suited for predicting thyroid disorders in diabetic patients. This approach provided a comprehensive understanding of the strengths and weaknesses of each model, helping guide the decision-making process for real-world applications.

3.5 Evaluation

The evaluation phase focused on assessing and comparing the performance of the models using different metrics: accuracy, precision, recall, F1 score, sensitivity, specificity, and a confusion matrix to provide a comprehensive view of the model's ability to correctly classify instances. The metrics were calculated based on the model predictions on the test dataset.

Accuracy: means how many times the model made a correct prediction among the total number of instances [16].

$$\boldsymbol{accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FN} + \boldsymbol{FP}} \quad (2)$$

Precision: means the number of positive (correct) predictions made by the model and belongs to the positive class [12].

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \quad (3)$$

Recall (Sensitivity): means the number of actual positive (correct) predictions made by the model out of all positive examples in the dataset [15].

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad (4)$$

F1score: provides a single score that combines both precision and recall in one number to find balance [24]. It is needed when there is uneven class distribution (more negative).

$$\boldsymbol{f1score} = \mathbf{2} * \frac{\boldsymbol{precision} * \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}} \quad (5)$$

Specificity (True Negative Rate): The percentage of actual negatives properly identified by the model [12].

$$\boldsymbol{specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}} \quad (6)$$

Each model's performance was evaluated using these metrics. The fold that yielded the highest accuracy with equal training and testing accuracies was noted, along with the corresponding optimal number of features.

4 Results

In this section, we present the feature importance ranking results and the evaluation results of the six machine learning models used.

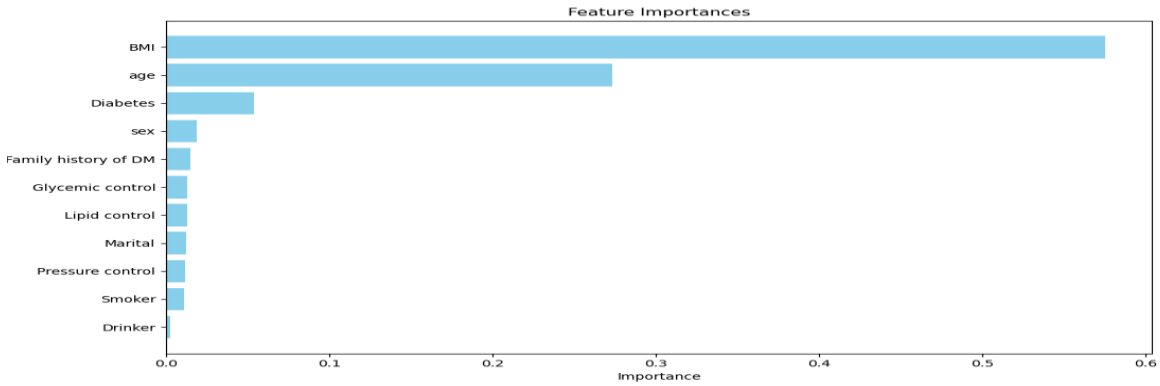

Figure 2: Feature importance ranking for Experiment 1 (on the RandomUnderSampler balanced dataset)

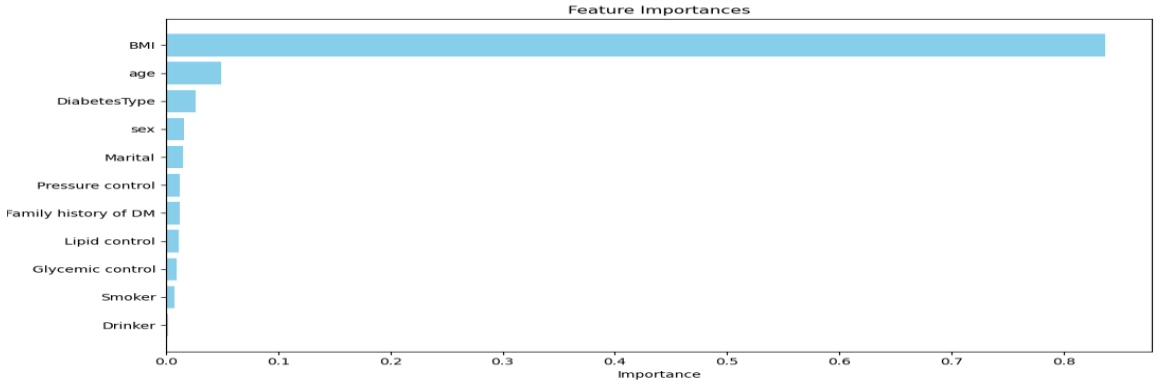

Figure 3:Feature importance ranking for experiment 2 (on the manually balanced dataset)

Figure 2 and Figure 3 display the feature importance ranking derived from a Random Forest model, used to predict thyroid disorders in diabetic patients. The x-axis shows the relative importance of each feature, with higher values indicating greater influence on the model's predictions. BMI and age are identified as the most critical features, with BMI showing the highest impact. Other features, such as diabetes type and sex, also contribute to the model but with comparatively lower importance. This ranking provides valuable insights into the factors most predictive of thyroid disorders in the context of diabetes. The results emphasize the significance of clinical factors like BMI and age in thyroid disorder prediction for diabetes patients.

Table 3: Experiment 1 evaluation metrics comparison table.

Classifier	Accuracy	Precision	F1-Score	Sensitivity (Recall)	Specificity	Confusion Matrix
RF	0.84	0.96	0.82	0.713	0.967	[[649,22], [193,479]]
DT	0.83	0.95	0.81	0.702	0.960	[[644,2], [200,472]]
KNN	0.83	0.92	0.81	0.720	0.934	[[627,4], [188,484]]
SVM	0.79	0.85	0.77	0.708	0.871	[[585,87], [196,476]]
LR	0.78	0.84	0.76	0.701	0.868	[[583,89], [201,471]]
NB	0.78	0.84	0.76	0.701	0.868	[[583,89], [201,471]]

Table 4: Experiment 2 evaluation metrics comparison table.

Classifier	Accuracy	Precision	F1-Score	Sensitivity (Recall)	Specificity	Confusion Matrix
RF	0.95	0.99	0.95	0.917	0.991	[[107,1], [9, 100]]
DT	0.95	0.96	0.95	0.944	0.963	[[105,4], [6, 102]]
KNN	0.94	0.97	0.94	0.917	0.972	[[105,3], [9, 100]]
LR	0.94	1.00	0.94	0.890	1.000	[[108,0], [12, 97]]
SVM	0.93	1.00	0.93	0.861	1.000	[[108,0], [15, 93]]
NB	0.93	1.00	0.93	0.861	1.000	[[108,0], [15, 93]]

The results show that in **Experiment 1** where the RandomUnderSampler technique was employed for data balancing, the Random Forest model demonstrated superior performance across all metrics compared to other models achieving the highest accuracy of 84%, precision of 96%, and F1-score of 82%. Followed by DT and KNN classifiers having the same accuracy of 83%.

However, SVM and LR showed a lower performance, with accuracies of 79% and 78%, respectively.

In **Experiment 2,** which used a manually balanced dataset, all the classifiers performed extremely well across all metrics, with the RF classifier achieving the highest accuracy of 95%, precision of 99%, and an F1-score of 95% indicating the model's high effectiveness in predicting thyroid disorders.

Similarly, the DT and KNN also demonstrated high accuracies of 95%, and 94%, Correspondingly. This great performance is most likely due to the balanced data that ensured a better representation of both classes, leading to more reliable model predictions.

The sensitivity and specificity of these models are significantly higher in Experiment 2 compared to Experiment 1, showcasing the efficacy of the manually balanced dataset in enhancing model performance.

The results showed that while using RandomUnderSampler for data balancing in the first experiment, the models did not reach the same level of effectiveness as in the manually balanced dataset in the second experiment which achieved a consistently high performance across all classifiers. This highlights that choosing thoughtful and effective data-balancing technique can improve the model's overall performance and prediction accuracy.

In summary, for both experiments, the Random Forest model emerged as the best-performing algorithm for predicting thyroid disorders in diabetic patients, followed closely by the Decision Tree and K Nearest Neighbors models. These models demonstrated high accuracy, precision, recall, and F1-score, making them suitable for deployment in clinical settings. Logistic Regression, Naïve Bayes, and SVM, while useful, showed comparatively lower performance and may require further optimization for effective use in this context.

5 Discussion

This study highlights the efficacy of machine learning models, particularly the Random Forest (RF) algorithm, in predicting thyroid disorders among diabetic patients. The findings emphasize the importance of model selection, data preprocessing, and feature analysis in achieving high predictive performance. This section explores comparisons with related works, reasons for Random Forest's superior performance, variations in model effectiveness, and limitations, alongside real-world implications of the findings.

5.1 Comparison with related works

The findings align with recent studies in the literature that emphasize the utility of machine learning for healthcare applications. For instance, studies such as Yadav et al. demonstrated the effectiveness of ensemble-based models like RF in handling structured medical datasets, particularly for classification problems. Compared to other methods, the RF model in this study yielded superior accuracy, recall, and precision, which can be attributed to its ability to handle non-linear relationships and its robust feature selection mechanism.

While [Priyanka Duggal & Shipra Shukla (2020)] also applied Support Vector Machines (SVMs) to medical datasets with 92% accuracy, our results indicate that SVM underperformed relative to RF, potentially due to the high dimensionality of the features or the imbalanced nature of the dataset. This highlights the importance of model selection based on the characteristics of the data.

5.2 Reasons for random forest's performance superiority

The RF model's outperformance can be attributed to several key factors. First, its inherent ability to handle both categorical and numerical data without extensive preprocessing makes it well-suited for medical datasets, which often include diverse feature types. Second, the use of RandomUnderSampler for data balancing helped mitigate the issue of class imbalance, which is a critical challenge in predicting rare conditions such as thyroid disorders in diabetic patients. RF's capacity to combine predictions from multiple decision trees also reduces the risk of overfitting, ensuring more generalized predictions. Furthermore, feature importance analysis revealed that variables such as BMI, age, and diabetes type were among the most predictive, aligning with clinical insights and lending credibility to the model.

5.3 Variations in performance across models

The variations in performance between models can be linked to their differing sensitivities to the dataset characteristics. For example, while K-Nearest Neighbors (KNN) is sensitive to feature scaling and data distribution, its relatively low performance could stem from the high dimensionality of the dataset. Similarly, SVM's reliance

on kernel functions may not have adequately captured the complex interactions within the data. In contrast, Decision Trees (DT) performed reasonably well but lacked the ensemble effect of RF, leading to slightly lower accuracy and recall. These findings suggest that models like RF, which can effectively leverage feature interactions and handle imbalanced data, are better suited for this specific prediction task.

5.4 Limitations and real-world applicability

Despite these promising results, several limitations must be acknowledged. First, the study relied on a single dataset, which may limit the generalizability of the findings to other populations or healthcare settings. Second, while RandomUnderSampler addressed class imbalance, other techniques such as SMOTE or hybrid approaches could be explored for potentially better results. Additionally, the dataset's retrospective nature may introduce biases inherent to the original data collection process.

In real-world healthcare environments, the applicability of this method is promising. The RF model's interpretability, particularly through feature importance scores, provides clinicians with actionable insights, aiding in early diagnosis and tailored treatment planning. However, practical deployment would require rigorous external validation and integration with electronic health records to assess scalability and user-friendliness.

6 Conclusion

Early prediction and diagnosis of diseases remain critical challenges in the medical domain, particularly for interconnected conditions like diabetes and thyroid disorders. While many studies have focused on predicting these diseases individually, limited research exists on predicting thyroid disorders specifically among diabetic patients.

This study aimed to bridge this gap by applying six machine learning algorithms to a local dataset of diabetic patients to predict the likelihood of thyroid disorders. Unlike previous studies that treated these conditions independently, this research explored the relationship between diabetes and thyroid disorders, given their intertwined impact on vital body functions.

Among the tested algorithms, the Random Forest model emerged as the most effective, achieving the highest accuracy, precision, and recall. Its ability to handle imbalanced data and highlight key predictive features, such as BMI, age, and diabetes type, further solidifies its potential as a valuable tool for early diagnosis.

The implications of these findings extend to enhancing healthcare practices by enabling clinicians to identify diabetic patients at risk of thyroid disorders, facilitating timely interventions, and potentially reducing complications. By improving early detection, this approach could significantly enhance the quality of life for individuals affected by both conditions.

In summary, this research contributes to the growing body of evidence supporting machine learning's role in healthcare, particularly for complex, multifactorial diseases. Future work should focus on validating these findings in diverse clinical settings, exploring alternative resampling techniques, and integrating these models into healthcare systems for real-world application.

References

[1] 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.

[2] 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.

[3] N. T. Y. Alibrahim, M. G. Chasib, S. S. Hamadi, and A. A. Mansour, "Predictors of Metformin Side Effects in Patients with Newly Diagnosed Type 2 Diabetes Mellitus," *Ibnosina Journal of Medicine and Biomedical Sciences*, Apr. 2023, doi: https://doi.org/10.1055/s-0043-1761215.

[4] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, Dec. 2022, doi: https://doi.org/10.1049/htl2.12039.

[5] S. Hassan, A.-K. Ali, and R. 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.

[6] 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.

[7] C. McElwain, F. McCarthy, and C. McCarthy, "Gestational Diabetes Mellitus and Maternal Immune Dysregulation: What We Know So Far," *International Journal of Molecular Sciences*, vol. 22, no. 8, p. 4261, Apr. 2021, doi: https://doi.org/10.3390/ijms22084261.

[8] 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. Doi: https://www.researchgate.net/publication/341742234_Thyroid_Disease_Classification_Using_Decision_Tree_and_SVM

[9] M. Nishi, "Diabetes mellitus and thyroid diseases," *Diabetology International*, vol. 9, no. 2,

pp. 108–112, May 2018, doi: https://doi.org/10.1007/s13340-018-0352-4.

[10] 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.

[11] S. Gopal, P. Gaurav, and D. Prateek, Machine learning algorithms using Python programming. New York: Nova Science Publishers, 2021.

[12] A. Panesar, *Machine Learning and AI for Healthcare: big data for improved health outcomes*. Berkeley, CA: Apress, 2021. doi: https://doi.org/10.1007/978-1-4842-6537-6.

[13] F. Pedro and G. Márquez, *Handbook of research on big data clustering and machine learning*. Hershey, PA: Engineering Science Reference (an imprint of IGI Global), 2020.

[14] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–21, Mar. 2021, doi: https://doi.org/10.1007/s42979-021-00592-x.

[15] Yuxi. (Hayden). Liu, *Python Machine Learning by Example Build Intelligent Systems Using Python, TensorFlow 2, Pytorch, and Scikit-Learn, 3rd Edition*. Birmingham: Packt Publishing, Limited, 2020.

[16] S. L. Mirtaheri and R. Shahbazian, *Machine Learning Theory to Applications*. CRC Press, 2022. doi: https://doi.org/10.1201/9781003119258.

[17] A. H. Khassawneh *et al.*, "Prevalence and Predictors of Thyroid Dysfunction Among Type 2 Diabetic Patients: A Case–Control Study," *International Journal of General Medicine*, vol. Volume 13, pp. 803–816, Oct. 2020, doi: https://doi.org/10.2147/ijgm.s273900.

[18] 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.

[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] 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.

[21] 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.

[22] 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.

[23] 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.

[24] G. S. Ohannesian and E. J. Harfash, "Epileptic Seizures Detection from EEG Recordings Based on a Hybrid system of Gaussian Mixture Model and Random Forest Classifier," *Informatica*, vol. 46, no. 6, Sep. 2022, doi: https://doi.org/10.31449/inf.v46i6.4203.