Ensemble Machine Learning Algor ithms for Predicting Thyroid Disorders in Diabetic Patients: A Comparative Analysis

Salma A. Mahmood *[1], Saad S. Hamadi[2]
[1]Department of Intelligent Medical Systems, University of Basrah, College of Computer Sciences and Information Technology, Basrah, Iraq
[2]Department of Internal Medicine, University of Basrah, College of Medicine, Basrah, Iraq
E-mail:salma.mahmood@uobasrah.edu.iq, and saad.shaheen@uobasrah.edu.iq
*Corresponding author

Keywords: machine learning, classification, ensembled machine learning, diabetes mellitus, thyroid disorders

Received: February 22, 2025

Thyroid diseases represent a significant health challenge due to their high prevalence and complex interactions with other diseases, negatively impacting quality of life and increasing the cost and complexity of treatment. Machine learning techniques have proven effective in medical applications, particularly in enhancing diagnostic accuracy and predictive performance. This study aims to develop an early prediction application for thyroid disorders in diabetic patients by comparing individual machine learning models—Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), and Logistic Regression (LR)—with ensemble learning models, including Random Forest (RF), Voting, Bagging, AdaBoost, Gradient Boosting, and Stacking. The results demonstrated that the Support Vector Machine (SVM) model outperformed other base models, achieving an accuracy of 97%, a precision of 93%, and sensitivity and specificity of 85% each. The K-Nearest Neighbors (KNN) model achieved an accuracy of 97%, a precision of 91%, and sensitivity and specificity of 83% each. Similarly, the Logistic Regression (LR) model achieved an accuracy of 97%, a precision of 90%, and sensitivity and specificity of 84% each. Among the ensemble methods, the Gradient Boosting method achieved the highest performance, with an accuracy of 97%, a precision of 92%, and sensitivity and specificity of 88% each. The Voting model achieved an accuracy of 97%, a precision of 92%, and sensitivity and specificity of 87% each. The Random Forest (RF) model achieved an accuracy of 97%, a precision of 89%, and sensitivity and specificity of 88% each. The significance of this study lies in early prediction of thyroid disorders in diabetic patients. We recommend the use of ensemble learning models due to their effectiveness in early diagnosis and prediction of pathological conditions as part of a computer-based medical diagnostic system.

Povzetek: Narejena je primerjava osnovne in sestavljene metode strojnega učenja za napoved bolezni ščitnice pri sladkornih bolnikih, pri čemer Gradient Boosting, Voting in Stacking dosegajo največjo točnost in AUC.

1 Introduction

Diabetes and thyroid disorders are among the most prevalent endocrine disorders, often exhibiting complex interactions that lead to pathological overlaps of varying severity and clinical impact [1]. Dysfunction in one condition can adversely affect the other, exacerbating symptoms and complications associated with both disorders [2]. The close relationship between Diabetes Mellitus (DM) and Thyroid Disorders is of significant clinical importance, particularly given epidemiological evidence underscoring their reciprocal effects on cardiovascular and metabolic complications [3].

The problem addressed in this study emphasizes the critical need for early detection of thyroid disorders in diabetic patients, particularly through the analysis of diabetes-related data. Given that diabetes is a chronic and serious condition often associated with thyroid dysfunction as one of its most significant complications, early prediction of thyroid disorders can provide timely alerts and enable personalized treatment plans. These plans have the potential to reduce patient suffering and minimize the waste of resources and time for healthcare institutions.

The thyroid gland is a small, butterfly-shaped gland located at the base of the neck. It is an essential component of the endocrine system. A complex system of glands coordinates many vital bodily functions, including producing hormones that regulate metabolism and coordinating other complex physiological activities [4].

Thyroid disorders are medical conditions that can significantly affect the metabolism and overall health of individuals, which include conditions such as hypothyroidism, hyperthyroidism, and thyroid cancer. Hypothyroidism slows down bodily functions, causing symptoms such as chronic fatigue, weight gain, depression, and sensitivity to cold. In contrast, hyperthyroidism causes the metabolism to speed up,

leading to symptoms such as unexplained weight loss, anxiety, irritability, and heart palpitations [5]. In contrast, thyroid cancer has been associated with increased morbidity and mortality, albeit to varying degrees. The majority of deaths associated with thyroid disorders are attributed to thyroid cancer [6]. Thyroid diseases affect millions of people around the world and, if left undiagnosed and untreated, can have a significant negative impact on their quality of life.

Diagnosing thyroid diseases is complex and time-consuming, requiring specialized knowledge and considerable clinical experience [7]. Diagnosis relies mainly on clinical examination and multiple blood tests due to the wide range of symptoms associated with these diseases, which include weight loss difficulties, obesity, constipation, muscle pain, hypersensitivity to colds, fatigue, and exhaustion. However, the high similarity between these symptoms and those associated with other medical conditions makes it difficult to distinguish between thyroid disease and other disorders, further complicating the diagnostic process and increasing the cost of treatment. Therefore, early diagnosis and proper treatment are key to managing these disorders and improving patients' quality of life.

On the other side, diabetes mellitus (DM), recognized as a chronic metabolic disorder, is characterized by high blood glucose levels due to a lack of insulin secretion, action, or both [8]. Insulin is a hormone synthesized and secreted from pancreatic beta cells that stimulates glucose uptake into cells for energy production while preventing the production of additional excess glucose from the liver via feedback mechanisms [9]. There are four primary types of Diabetes mellitus (DM): Type 1 diabetes (T1D), an autoimmune disease resulting from the destruction of pancreatic β-cells; 5–10% of cases are typically diagnosed in early life; Type 2 diabetes (T2D), caused by insulin resistance and progressive β-cell dysfunction, 90–95% of cases, severely attached with aging and obesity; gestational diabetes mellitus (GDM), it is hyperglycemia detected firstly through pregnancy, which may resolve after postpartum but may increase the risk of future T2D; and rare of DM forms, monogenic Diabetes and pancreatic variants-related disease [10]. According to the International Diabetes Federation (2021), Diabetes is one of the most common and prevalent diseases globally, affecting more than 537 million adults, and this number is expected to increase to 643 million by 2030 and 783 million by 2045 [11]. Diabetes is a chronic disease that poses a significant health threat, with the number of deaths related to its medical causes increasing annually, making it one of the significant health challenges in both emerging and developed countries [12]. Therefore, an intelligent and accurate diagnostic approach and effective therapeutic management are essential to minimize the severity of symptoms, prevent potential complications, and maintain overall health and quality of life for patients [13].

Artificial Intelligence (AI) programs mimic human intelligence in performing functions like problem-solving, learning from experience, recognizing patterns, and decision-making. Machine Learning Models (MLM) is one of the promising branches of AI, which depends on the development of algorithms that can automatically extract patterns from primary data and analyze large amounts of patient data, including medical records, laboratory results, and imaging studies, to identify patterns and correlations that may indicate medical disease symptoms. Such enhancements enhance their performance over time and enable them to provide accurate predictions and decisions to enhance the diagnostic process without direct human intervention [9] [14].

On the other hand, developing Ensemble Machine Learning Models (EMLMs) or meta-learners is a promising strategy for increasing the stability and accuracy of base MLMs [15]. EMLMs are based on the principle of "union is strength" where a group of base MLM or weak learners are combined to build a strong and efficient learner [16].

The state-of-the-art models heavily rely on a set of Heterogeneous weak learners and meta-learners. Weak learners are trained on the dataset's attributes to produce initial results, while the meta-learner combines these results to generate an accurate final prediction. Since weak learners perform differently, combining their predictions improves the overall classification performance, making ensemble models more effective than individual classifiers. These techniques can achieve higher diagnostic accuracy and stability, leading to improved patient outcomes and more efficient diagnostic processes. It also contributes to reducing misdiagnosis, improving the reliability of disease diagnosis, and enhancing the overall quality of patient care [9]. Finally, integrating machine learning and combined technologies with precision medicine to deliver personalized treatment through real-time Clinical Decision Support Systems (CDSS) represents the future of AI-powered healthcare.

Despite significant advancements in the application of machine learning for diagnosing thyroid disorders, there remains a notable gap in research specifically addressing the prediction of thyroid disorders in diabetic patients. This study aims to bridge this gap by utilizing a balanced diabetes dataset and employing the Random Forest (RF) approach to identify the most relevant features for accurate predictions. Furthermore, a comprehensive comparative analysis is conducted between base machine learning models (MLMs) and ensemble machine learning models (EMLMs) to evaluate their predictive performance. A focus on binary classification for thyroid disease detection, where Thyroid Patient = 1 and No Thyroid = 0. Following the comprehensive preprocessing of the diabetes dataset, which included cleaning, normalization, and discretization of features to ensure data quality and suitability for analysis, we can formulate a set of research questions that define the contributions of this study.

- What is the effectiveness of using undersampling techniques to achieve a balanced class distribution in the dataset?
- How can Random Forest (RF) be utilized to identify the most impactful features for improving prediction accuracy?

- What is the performance of base machine learning models, such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), and Decision Tree (DT), in detecting thyroid disorders among diabetic patients?
- How can base models be integrated into ensemble architectures using methods such as soft voting, bagging (bootstrap aggregating), adaptive boosting (AdaBoost), gradient boosting, and stacking to enhance prediction accuracy?
- What are the outcomes of the comprehensive evaluation and comparison of models to identify the best-performing approach for thyroid disorder detection?

The remainder of this research is structured as follows: Section 2 presents the literature review, providing an overview of relevant studies and identifying gaps in the existing research. Section 3 details the methodology employed in this study, including data preprocessing, model selection, and evaluation techniques. Section 4 presents the results obtained from the experiments, while Section 5 discusses the findings, their implications, and comparisons with prior work. Finally, Section 6 concludes the study and provides recommendations for future research directions..

2 Literature review

This section reviews various research studies that have utilized different methods to predict and detect thyroid diseases. Providing a comparative overview of multiple studies, focusing on their datasets, algorithms, evaluation metrics, and results.

Salman and Sonuc [18] (2021) their study aimed to categorize thyroid disease into three categories: hyperthyroidism, hypothyroidism, and normal. They utilized Support vector machines, random forest, decision tree, naïve bayes, logistic regression, k-nearest neighbors, multi-layer perceptron (MLP), and linear discriminant analysis to classify thyroid disease. These classifiers analyze 1250 local Iraqi datasets collected from hospitals and labs. RF recorded a maximum accuracy of 98.92%.

In [19], Chaganti et al. (2022) focused on balancing the UCI dataset for efficient findings and feature selection methods. They used backward, forward, and Bi-Directional elimination techniques before applying MLF methods with Extra Trees (ET). They used various machine learning and deep learning classifiers, SVM, RF, LR, Gradient Boosting Machine (GBM), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), CNN-LSTM, and AdaBoost to perform thyroid disease classifying into four main classes: hypothyroid, concurrent non-thyroidal illness, increased binding protein, and compensated hypothyroid. Their outcomes showed that using the RF combined with FS and 10-fold cross-validation attained a superior accuracy of 0.99%, with the lowest computational complexity.

Akhtar et al. [20] (2022) Three types of feature selection approaches, i.e., recursive feature elimination (RFE), selection from a model (SFM), and choosing k-best (SKB) used to select attributes from the "thyroid 0387" dataset (founded in open-source KEEL repository). The authors developed a compelling ensemble of ensembles to improve thyroid disease diagnosis. The homogeneous ensemble activated bagging and Boosting-based classifiers, and the voting ensemble used both soft and hard voting to classify the results. The evaluation metrics, recall, sensitivity, accuracy, and Cohen kappa, were used to check the model performance. The RFE with logistic regression estimator achieved the highest accuracy of 99.27% with 97% precision and Fl-score, 98% recall with low computational cost.

Similarly, The authors of [16] (2022) studied how ensemble techniques can utilized for thyroid prediction. They proposed using ensembled methods, i.e., Boosting, Bagging, and Majority Voting, to enhance the accuracy of the proposed models. They used a local dataset collected from Endocrinology clinics in Kashmir (India), which consists of 1257 records with 11 features. They apply two types of neural networks: Neural network methods, such as Multi-Layer Perceptron (MLP) with Back-Propagation. The experimental outcomes explore that the Bagging Classifier RFC outperformed with an accuracy of 0.99%.

Alshayeji et al. [21] (2023) developed a thyroid disease prediction system that can be integrated with real-time computer-aided diagnosis (CAD) systems to facilitate diagnosis and support early treatment. The study initiated existing research gaps, including the lack of detailed feature analysis and the need to improve prediction accuracy and enhance reliability. In their study, they used several machine learning algorithms such as decision trees (DT), support vector machines (SVM), logistic regression (LR), nearest neighbor algorithm (KNN), and neural networks (NN). The 5-fold cross-validation technique was applied to avoid overfitting. The hyperparameters were optimized using Bayesian optimization. Ensemble learning methods such as bagging and boosting were applied to a public dataset containing 29 thyroid-related traits from the University of California to ensure the model's reliability and improve the accuracy of prediction decisions. The proposed results achieved a high accuracy of 99.5%, with a sensitivity of 99.39% and specificity of 99.59% when using the boosting method.

In the research in [22] (2024), the Authors proposed a practical state-of-the-art artificial intelligence approach for the early diagnosis of thyroid disease. Their study contained several tasks: the nominal continuous synthetic minority oversampling technique (SMOTE-NC) for data balancing, a fine-tuned light gradient booster machine (LGBM) technique to diagnose thyroid illness and handle class imbalance problems, applied advanced machine learning and deep learning methods for comparison to evaluate performance, and Shapley Additive exPlanations (SHAP) to enhance the transparency and interpretability of decision-making processes. These tasks were applied to an open-access thyroid disease dataset based on 3,772 observations. Their results show that the proposed SNL (SMOTE-NC-LGBM)

approach outperformed the state-of-the-art approach with high-accuracy performance scores of 0.96.

In their research [23] (2024), the authors used a local dataset of 4000 records with 15 parameters of thyroid disorder. They classified the data into Hyperthyroidism, Hypothyroidism, and Euthyroid. They preprocessed and cleaned the data, selected the most relevant features, and balanced the dataset. They conducted preprocessing, cleaning, feature selection to retain the most relevant features, and balancing the dataset. In the first strategy, Various forms of the Decision tree, ID3, C4.5, CART, and Random Forest, were applied. The second strategy, ensembled methods, involved the Random Forest algorithm, which combines their outcomes via voting. The outcomes of their method achieved accuracy (68% - 74%).

The study in [24] (2024) Utilized a highly imbalanced UCI thyroid disease dataset, which contains 9,172 samples and 30 attributes, to develop an effective thyroid disease prediction system-based machine learning approach. Their study focused on balancing their dataset using the down-sampling technique. The RF-based self-stacking classifier is present. Their approach diagnoses four primary classes, i.e., hypothyroidism, concurrent non-thyroidal illness, increased binding protein, and compensated hypothyroidism. Achieving 99.5% accuracy and 100% macro precision, recall, and F1-score. To confirm their superior approach, an exhaustive comparative analysis encompasses

In the study [17] conducted in 2024, the authors focused on developing a machine learning (ML) system to predict thyroid disorders in diabetic patients. They employed several ML classifiers, including Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Naive Bayes (NB) to carry out various experimental tasks. The research utilized three local datasets: a balanced dataset with 13438 samples generated through Random UnderSampling (RUS), a subset of Type 2 diabetes (T2D) patients consisting of 11648 samples, and a subset of Type 1 diabetes (T1D) patients that included 1,790 samples. The results showed that the Random Forest (RF) classifier produced the best predictions with the highest accuracy of 0.85 and F1-score of 0.83 in the T2D-targeted dataset among all classifiers tested. Additionally, it demonstrated robust performance on the balanced dataset created using RUS.

Gupta et al. [25] (2024) proposed using various machine learning models for the detection of thyroid disease, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), AdaBoost, Gradient Boosting Machine (GBM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). An imbalanced dataset from the Kaggle repository, consisting of 9,172 samples and 31 features, was utilized for implementation. Several strategies are implementing to enhance model performance. A Conditional Tabular Generative Adversarial Network (CTGAN) was employed to balance the dataset and mitigate model overfitting. Fine-tuned parameters using a Differential Evolution (DE) optimizer. Principal Component Analysis (PCA) utilized for feature reduction. Unlike previous studies, this research classified the data into ten distinct categories: K: Concurrent non-thyroid illness, G: Compensated hypothyroidism, I: Increased binding protein, F: Primary hypothyroidism, R: Discordant assay results, N: Over-replaced, A: Hyperthyroidism, L: Consistent with replacement therapy, M: Under-replaced and -: No condition. The results indicated an impressive accuracy score of 0.998% when using AdaBoost in conjunction with DE optimization.

The author of [26] (2025) focused on enhancing their deep learning, Convolution Neural Network CNN, by integrating practical tools of feature selection, i.e., Random Forests with Principal Component Analysis (PCA) and L1 regularization, making a Hybrid Feature Selection and Deep Learning Framework (HFSDLF) and utilizing the UCI Thyroid Dataset 10,000 images (5,000 benign, 5,000 malignant). The Random Forest classifier achieved the highest accuracy of 96.30 %, outperforming other models such as Decision Trees and Logistic Regression, with notable improvements in sensitivity and specificity. They underscore the novelty of their proposal, and on that, their proposal outperformed other counterpart works.

In the study [27] (2025), the authors addressed and evaluated seventeen machine learning models and an Ensemble ML classifier including KNN, Decision Tree, Random Forest, Bernoulli Naïve Bayes, Logistic Regression, Gradient Boosting, LGBM Classifier with DART, LGBM Classifier with GBDT, XGBoost, Linear Support Vector Machine, CatBoost, AdaBoost, Nearest Centroid Classifier, Voting Classifier, Bagging Classifier, and LightGBM, using a voting strategy. UCI dataset Machine Learning Repository comprises 9172 observations and 31 attributes. Techniques, such as random oversampling for class balancing and feature selection for feature reduction, were used. Their proposal model outperformed by achieving 100% sensitivity and 99.72% accuracy using the XGBoost algorithm and SelectKBest features selection.

The study of Sayyid [28] focus on 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). Using local dataset contianing 44,534 instances, they employed several preprocessing steps two balancing approachs, manual balancing and RandomUnderSampler. Among the evauated 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. indicating its robustness in handling imbalanced datasets.

Table 1: Summarizes the related works.

Ref	Year	Dataset	Sample Size	Target Distribution	Model	Class	Best model	Results accuracy
[18]	2021	Iraqi Dataset	1250	N/A	SVM, RF, DT, NB, LR, KNN, MLP, and LDA	3	RF	98.92%
[19]	2022	UCI	1774	balanced	RF, GBM, LR, SVM, AdaBoost, LSTM, CNN, and CNN-LSTM	4	RF -MLFS	99%
[20]	2022	UCI Thyroid 0387	7200	N/A	RF, BME, AdaBoost, and XGBoost	2	LR-RFE	99.27%
[16]	2022	Local (India)	1257	imbalanced	Boosting, Bagging, and Majority Voting	3	Bagging+ RFC	0.99%
[21]	2023	UCI	3711	balanced	DT, SVM, LR, KNN, NN + Bagging and Boosting	2	boosting method	99.5%, 99.39%
[22]	2024	an open-access dataset [29]	3,772	imbalance	light gradient booster machine (LGBM)	2	SMOTE-NC-LGBM	0.96%
[23]	2024	Local dataset	4000	imbalanced	ID3, C4.5, CART, J4.8, RF and Voting	3	RF + voting	72%
[24]	2024	UCI	9,172 samples and 30 features	imbalanced	RF-based self-stacking	2	RF-based self-stacking classifier	99.5%,
[17]	2024	3 Iraqi local datasets	13438 11648 1790	imbalanced	RF, DT, KNN, SVM, LR, and NB	2	RF	0.83%
[25]	2024	UCI	9,172 samples and 30 features	imbalanced	RF, SVM, LR, AdaBoost, GBM, CNN, RNN, and LSTM+DE optimizer	10	optimized AdaBoost model	0.998%
[26]	2025	UCI	10,000 images	Balanced	RF+PCA+L1	2	RF+PCA+L1	96.30
[27]	2025	UCI	9,172 samples and 30 features	imbalanced	KNN, DT, RF, Bernoulli NB, LR, Gradient Boosting, LGBM + DART, LGBM+GBDT, XGBoost, Linear SVM, CatBoost, AdaBoost, NC, Voting, Bagging, and LightGBM	2	XGBoost + SelectKBest	99.72%
[28]	2025	Iraqi local dataset	44,534 samples with 12 features	imbalanced	RF, DT, KNN, LR, NB, and SVM.	2	RF	95%

Most of the previous works reviewed in this research focus on the detection of thyroid disorders in general, without specifically targeting diabetic patients. Additionally, many of these studies suffer from imbalanced datasets, which can lead to biased results and increase the risk of overfitting. Moreover, the reliance on limited-sized datasets in numerous studies may negatively impact the generalizability of the developed models. On the other hand, ensemble methods, particularly Random Forests (RF), have demonstrated notable superiority in most studies that employed them, highlighting their effectiveness in addressing these challenges.

In this research, we aimed to overcome these issues by using a relatively balanced dataset with a large enough size, which helps in developing more generalizable models suitable for real-world medical applications.

3 Research Methodology

The primary challenge associated with thyroid disorders lies in the multiplicity of their clinical manifestations and the complex interplay of genetic, environmental, and lifestyle factors that contribute to their onset and progression. These disorders are also strongly linked to several comorbid conditions, including diabetes, cardiovascular diseases, and psychiatric disorders. Such interrelationships further complicate the diagnostic process, necessitating a multidisciplinary diagnostic and therapeutic approach to effectively address these complex health issues. This study aims to predict thyroid dysfunction in patients with Diabetes Mellitus (DM) through a comparative analysis of various machine learning methods. The goal is to enhance predictive performance and derive an optimal generalizable model suitable for real-world clinical scenarios.

Figure (1) illustrates the proposed methodological framework, which begins with the development of individual classifier models, including Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), and Logistic Regression (LR). These models are then integrated into a hierarchical ensemble architecture employing techniques such as Random Forest (RF), soft voting, bagging (bootstrap aggregating), adaptive boosting (AdaBoost), gradient boosting, and stacking. This integration aims to synergize the strengths of each algorithm while mitigating individual biases and variances. The research methodological comprising the following stages:

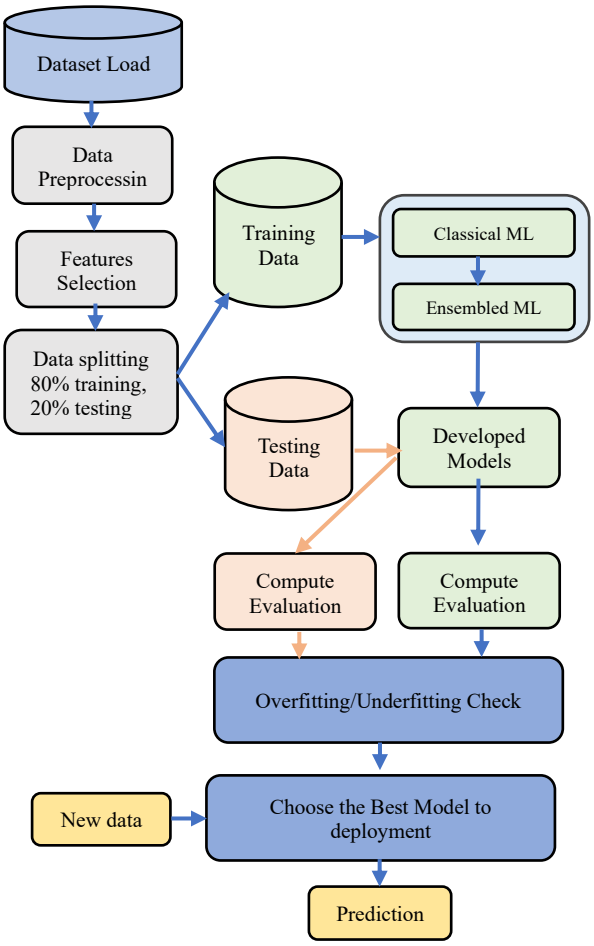

Figure 1: Shows the proposed methodological framework.

3.1 Dataset description

The data were collected from the Faihaa Specialist Center for Diabetes, Endocrinology and Metabolism (FDEMC) in Basra, Iraq, to attain a broad and diverse representation of the city's residents. This accredited and reputable source enhances the accuracy and reliability of the data. In the initial phase, the dataset comprised 44,539 samples. However, it lacked a balanced distribution of categories, with only 15.17% of cases representing patients with thyroid disorders (6,755 cases), compared to 84.83% of cases without thyroid disorders (37,784 cases).

3.2 Dataset preprocessing

This stage involves data cleaning, handling missing or outlier values, standardizing formats, and encoding categorical variables using appropriate techniques and balanced classes. Random UnderSampler was applied to reduce the data size to 13,438 samples, with a 50% equal distribution between patients with and without thyroid disorders. Although this method balances the categories and thus avoids bias towards the larger category, it may result in losing some valuable information by eliminating samples from the larger category. Table (2) displays the used dataset's exploratory data analysis (EDA).

3.3 Features selection

The initial dataset contained 13,438 records and 19 attributes. To improve the training model's efficiency and reduce the number of attributes in the training process., the Random Forest (RF) algorithm was utilized to assess the importance of each attribute to the decision class. Based on the results of this process, the number of input features was reduced to only 12, representing the features that have the most impact on the training results. as the figure (2) shows.

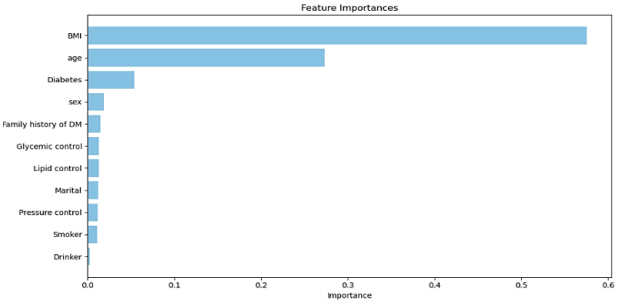

Figure 2: Shows the feature selection.

Table 2: Displays the used dataset's exploratory data analysis (EDA).

	mean	std	25%	50%	75%	min	max
age	55	15	45	56	66	5	100
sex	0	0	0	0	1	0	1
Smoker	0	0	0	0	0	0	2
Drinker	0	0	0	0	0	0	2
Marital	1	0	1	1	1	0	3
Family history of DM	1	0	0	1	1	0	1
BMI	34	8	28	35	39	17	49
Glycemic control	0	0	0	0	0	0	1
Lipid control	0	0	0	0	1	0	1
Pressure control	1	0	0	1	1	0	1
Diabetes	2	0	2	2	2	1	2
Thyroid	1	1	0	1	1	0	1

3.4 Data splitting

The dataset is divided randomly into two subsets: a training set (80%) for model development and a testing set (20%) for performance evaluation.

3.5 Classical machine learning models [29]

Five ML algorithms, including Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), are used, with hyperparameter tuning and Kfold cross-validation with k=5 to balance model complexity and accuracy.

Decision Tree (DT) is one of the most prominent classification algorithms that divides data into subsets using simple rules. It is important in classification because it allows decisions based on specific data attributes, leading to accurate and efficient classification. Decision trees are also valued for their ability to provide a clear visual explanation of the classification process, making it easier to understand and interpret the results. They are widely used in various applications because of their speed, simplicity, and ability to handle complex data.

K-Nearest Neighbors (K-NN) is a simple and efficient classification method that categorizes samples based on their proximity to neighboring points in a multidimensional space. It requires no prior training and effectively handles nonlinear data. K-NN is widely used in classification tasks because of its simplicity, versatility, and pattern recognition accuracy.

Gaussian Naïve Bayes (GNB) algorithm is one of the most frequently used classification algorithms, and it ranks very high in efficiency. This algorithm is based on Bayes' theorem, which assumes that the data belonging to a particular class is normally distributed and that the features are independent. Since this algorithm can be applied in a wide range of fields, its best-known feature is its capability to handle high-dimensional data effectively. Because of its high dimensionality, Gaussian Naive Bayes (GNB) is one of the most frequently used algorithms in classification tasks, as it is simple and cheap to compute and can determine class labels with high accuracy.

The Support Vector Machine (SVM) is one of the strongest classification algorithms because it searches for the best separation between classes in a space of more than two dimensions. This algorithm is highly effective with nonlinear data due to complex kernels, which makes it useful for challenging classification tasks. SVM is common in many applications because of its ability to achieve high precision and distinct separation among classes.

Logistic Regression (LR) is an effective classification algorithm that uses a logistic curve to calculate the probability of an object belonging to a specific class. This algorithm is particularly well-suited for solving linear binary classification problems, contributing to its popularity. Due to its ease of use, speed, and straightforward interpretation, logistic regression is primarily employed in various classification tasks. To implementation, we use the defaul parameter for all base classifiers.

3.6 Ensemble machine learning classifiers [15]

RF, voting, bagging, adaptive boosting (AdaBoost), gradient boosting, and stacking techniques were used in this study to enhance predictive robustness and reduce variance.

Random Forest (RF) is one of the most effective classification algorithms. It relies on constructing a collection of independent decision trees that work together to provide a final classification based on majority voting. This technique is highly effective at handling large and complex datasets, and it resists overfitting due to the use of random sampling and random feature selection during tree construction. Random Forest is widely used in classification tasks due to its accuracy, flexibility, and ease of interpretability.

Voting algorithm is a classification technique that gathers multiple classifiers and uses either the average or majority vote. It is often used because it increases accuracy and stability in complex or imbalanced cases, reducing the impact of individual model errors.

Bagging is an ensemble learning method that trains multiple models of the same learning algorithm on random data samples. After training all the models, the results are combined for better classification accuracy. By "sampling with replacement," this method effectively reduces variance and improves model performance by minimizing the impact of multiple model errors. Bagging is used in classification tasks to improve accuracy and reduce overfitting in high-variance models.

AdaBoost algorithm is an ensemble approach that attempts to improve the performance of models with weak predictive power by assigning more weight to examples with difficult classifications. This method is helpful for bias reduction and improving accuracy by directly combining several weaker models into one strong model. AdaBoost is often employed in classifications where overwhelming progress is needed on complex or unbalanced data sets.

Gradient Boosting algorithm is an effective ensemble learning technique that builds strong models by correcting previous models' errors. This algorithm is valuable for bias reduction and accuracy as predictions progressively improve. Gradient Boosting is widely used in classification to achieve high performance when dealing with complex or unbalanced datasets.

Stacking Algorithm is an ensemble learning technique that improves classification accuracy by combining the outcomes of multiple base models with a complementary model known as a metamodel. This approach produces more accurate and reliable results by leveraging the strengths of different models. Classification uses a stack to achieve superior performance by combining predictions from multiple models. In this research, we use the defaul parameter for all ensembeld classifiers.

3.7 Evaluation metrics

Validation and evaluation are conducted using diverse commonly used metrics: accuracy, precision, recall/sensitivity, F1-measure, specificity, and ROC-AUC [29].

Table 3: Summarizes the evaluation metrics used in this study

Evaluation metric	Equation
Accuracy	It is the ratio of correctly predicted instances to the total number of instances, and overall correctness is measured. $\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$
Precision	is the ratio of true positive predictions to the total predicted positives, indicating the model's ability to avoid false positives. $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
Recall/ Sensitivity	Recall is the ratio of true positive predictions to the total actual positives, measuring the model's ability to identify all relevant instances. $\text{Recall}/= \frac{\text{TP}}{\text{TP} + \text{FN}}$
F1_ Measure	F1_Measure is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. $\text{F1\_Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
Specificity	Specificity or True Negative Rate measures a model's ability to correctly identify negative cases. It is defined as: $\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$
ROC_AUC	The Receiver Operating Characteristic (ROC) curve is a widely used graphical tool for evaluating the performance of classification models across different threshold settings. It illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR), providing insights into the model's discriminatory ability. The ROC curve is beneficial in assessing binary classifiers and is frequently applied in medical diagnostics, machine learning, and statistical decision theory. [30]

Where:

- TP: True Positive is the number of correctly categorized records.
- TN: True Negative is the number of categorized documents that correctly rejected.
- FP: False Positive is the quantity of misclassified records.
- FN: False Negative is the percentage of categorized records that incorrectly rejected.

3.8 Check overfitting and underfitting

Training Accuracy and Testing Accuracy can detect the presence of Overfitting or Underfitting. In the case of overfitting, the training accuracy is very high, in contrast, the testing accuracy is relatively low, showing that the model overfits the training data and neglects to generalize to new data. In the case of underfitting, the testing accuracy is very high, in opposite of training accuracy, which indicating that the model cannot adequately learn the patterns in the data. In the study, we compute training accuracy and testing accuracy for all used models, then finding the accuracy difference between them, we use the 0.2 and -0.2 as a threshold as the following python code.

```
if accuracy_difference > 0.2:
    fitting_status = "Overfitting"
elif accuracy_difference < -0.2:
    fitting_status = "Underfitting"
else:
    fitting_status = "Balanced"
```

The accuracy differences were recoreded in table (4).

4 Results

We conducted an intensive experiment to assess the performance of the different learned models and examine misclassification errors. To perform all experiments in this study, we use a PC with the following specifications: Intel R © Core (TM) i7-8700 CPU with 32 GB RAM, 3.20 GHz frequency. All methods are implemented herein using Python 3.10.0 programming with Anaconda (spider). Python-based ML libraries such as NLTK, pandas, and sci-kit-learn are utilized to investigate the performance metrics by the proposed methods.

The results of the base models demonstrated that the outperformed Support Vector Machine (SVM) model achieved an accuracy of 97%, a precision of 93%, and 85% for both sensitivity and specificity. The K-Nearest Neighbors (KNN) model achieved an accuracy of 97%, a precision of 91%, and 83% for both sensitivity and specificity. The Logistic Regression (LR) model achieved an accuracy of 97%, a precision of 90%, and 84% for both sensitivity and specificity. Among the ensemble methods, the Gradient Boosting method outperformed other models, achieving an accuracy of 97%, a precision of 92%, and 88% for both sensitivity and specificity. The Voting model achieved an accuracy of 97%, a precision of 92%, and 87% for both sensitivity and specificity. The Random Forest model achieved an accuracy of 97%, a precision of 89%, and 88% for both sensitivity and specificity.

These models exhibited improved classification abilities by capturing intricate patterns, which boosts predictive accuracy. Table (4) shows the findings of the used classifiers. Figure (3) shows the visualization comprehensively performace of all used modesl.

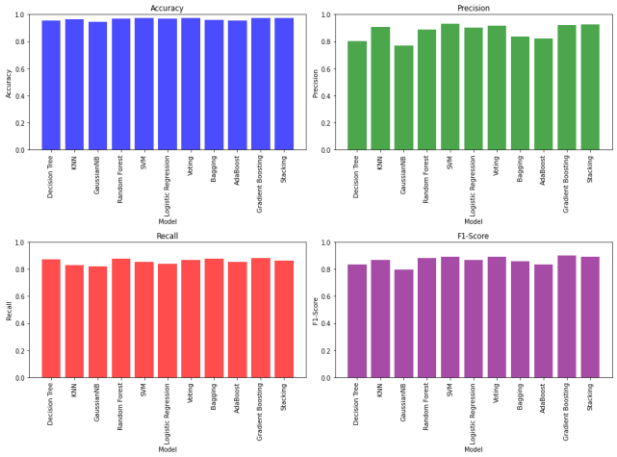

Figure 3: The visualization comprehensively summarizes the model's efficacy.

Table 4: Shows the findings of the used classifiers.

Model	Accuracy	Precision	Recall / Sensitivity	Specificity	F1-Score	Training Accuracy	Testing Accuracy	Accuracy Difference
DT	0.95	0.80	0.87	0.87	0.83	1.00	0.95	0.04
KNN	**0.97**	**0.91**	**0.83**	**0.83**	0.87	0.97	0.97	0.00
Gaussian NB	0.94	0.77	0.82	0.82	0.80	0.94	0.94	0.00
SVM	**0.97**	**0.93**	**0.85**	**0.85**	0.89	0.97	0.97	0.00
LR	**0.97**	**0.90**	**0.84**	**0.84**	0.87	0.96	0.97	0.00
RF	**0.97**	**0.89**	**0.88**	**0.88**	0.88	1.00	0.97	0.03
Voting	**0.97**	**0.92**	**0.87**	**0.87**	0.89	0.98	0.97	0.01
Bagging	0.96	0.84	0.88	0.88	0.86	0.99	0.96	0.03
AdaBoost	0.96	0.82	0.85	0.85	0.84	1.00	0.96	0.04
Gradient Boosting	**0.97**	**0.92**	**0.88**	**0.88**	0.90	0.97	0.97	0.00
Stacking	**0.97**	0.92	**0.86**	**0.86**	0.89	0.98	0.97	0.01

Figure 4: presents Receiver Operating Characteristic (ROC) curves, a graphical representation of classification model performance across varying decision thresholds.

Sensitivity and Specificity are fundamental components of the Receiver Operating Characteristic (ROC) Curve, as they are used to compute the Area Under the Curve (AUC), which serves as a key metric for evaluating a model's ability to distinguish between different classes. Sensitivity (True Positive Rate - TPR) measures the model's ability to correctly identify positive cases, while Specificity (True Negative Rate - TNR) assesses its capability to correctly identify negative cases.

Table 5: Shows the models performance using ROC AUC

Model	Recall (Sensitivity)	Specificity	AUC
Gradient Boosting	0.88	0.88	0.99
Voting	0.87	0.87	0.99
Stacking	0.86	0.86	0.99
Logistic Regression	0.84	0.84	0.99
SVM	0.85	0.85	0.98
Random Forest	0.88	0.88	0.98
Bagging	0.88	0.88	0.97
GaussianNB	0.82	0.82	0.97
KNN	0.83	0.83	0.95
Decision Tree	0.87	0.87	0.92

The results indicate that models such as ensemble techniques like Voting, Stacking, and Gradient Boosting, along with Logistic Regression, exhibit the highest performance, achieving an AUC of 0.99. with balance ratio of sensitivity and speicifity as showen in table (5). This signifies their superior ability to differentiate between classes, which is particularly critical in medical applications such as disease diagnosis and risk assessment, where minimizing errors is paramount. Additionally, Support Vector Machine (SVM) and Random Forest demonstrate strong performance, with an AUC of 0.98, making them viable choices when balancing accuracy and computational efficiency. Conversely, models like Decision Tree, K-Nearest Neighbors (KNN), and AdaBoost exhibit comparatively lower performance, with AUC values ranging between 0.92 and 0.95, rendering them less reliable in high-stakes medical environments that require precise decision-making. Gaussian Naïve Bayes (GaussianNB), despite achieving an AUC of 0.97, may be less suitable for scenarios involving complex and heterogeneous data. Based on these findings, Stacking, Voting, and Gradient Boosting are recommended for medical applications demanding high accuracy, whereas models such as Decision Tree, KNN, and AdaBoost should be avoided when maximizing precision and reliability is a priority.

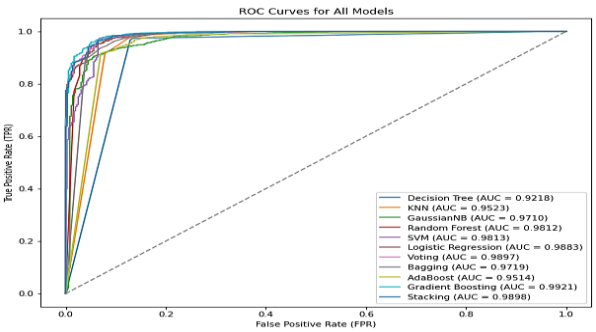

Figure 4: shows the ROC-AUC for all Models.

5 Discussions

The current study presents a comparative analysis of different ML models to find the best models with superior performance to apply in early predition of thyroid disorder in diabetes patient. This section discussion the variations in performance across models and real-world applicability, Comparison with related works and the Limitations.

5.1 The supuerior performance across models and real-world applicability

The findings demonstrated the high efficiency of ensembled ML methods in accurately detecting positive cases, which contributes to reducing treatment costs, shortening the required intervention time, and alleviating patient suffering. Additionally, these methods were distinguished by their ability to reduce false positives, which could lead to subjecting healthy individuals to unnecessary treatment plans. The results also confirmed the stability of these models and the absence of overfitting or underfitting issues. Furthermore, the study showed that implementing feature importance detection algorithms significantly improved the accuracy of the models by reducing the number of features used in the training process, thereby enhancing the efficiency of the resulting models.

5.2 Comparision with privous works

To compare the results of our current study with previous research that utilized ensemble methods, we note that studies [18], [19], [23], and [24] align with our findings in demonstrating the effectiveness of the Random Forest (RF) method as a robust ensemble technique, achieving accuracies exceeding 98% in those studies.

However, it is worth noting that the dataset sizes used in those studies were relatively limited compared to the larger dataset employed in our current research, which achieved an accuracy of 97% using the same method. In contrast, the RF method in our study outperformed the results of study [17], which achieved an accuracy of 83%, and

studies [26], which achieved accuracies of 96.3% and 95%, respectively. Additionally, while source [23] reported the superiority of the Voting method with an accuracy of 72%, our current study achieved a significantly higher accuracy of 97% for the same method.

5.3 Limitations

Despite these promising fidings, several limitations can be summarized as follows:
First, the study relied on a single dataset, which may limit the generalizability of the findings to other populations or healthcare settings. Second, the research used RandomUnderSampler to address the issue of class imbalance, while other techniques such as SMOTE or hybrid approaches could be explored to achieve potentially better results. Third, the Random Forest (RF) model was used to extract important features influencing classification. However, other methods, such as XGBoost, may yield better results due to its high sensitivity in understanding complex interactions among medical features. Finally, in real-world healthcare environments, these applications can serve as valuable decision-support tools, providing actionable insights for clinicians, particularly novice physicians. This contributes to the development of their diagnostic skills and assists them in early diagnosis and the formulation of personalized treatment plans for patients.

6 Conclusion

This research focuses on developing an early detection system for predicting thyroid disorders in diabetic patients with high accuracy. A comparative study was conducted among various machine learning classifiers, baseline models, and ensembled models to identify the most effective approach to be integrated into an Automated Computer-Aided Diagnosis (ACAD) to enhance diagnostic accuracy and improve the quality of medical care. The experimental findings demonstrated a clear superiority of ensemble models such as Gradient Boosting, RF, and Voting, as these models achieved the highest values for the Area Under the Curve (AUC ≈ 0.99%) with consistently high predictive accuracy (Accuracy ≈ 0.97%). These findings confirm the effectiveness of ensembled models in enhancing predictive stability and reducing variance. Such superiority reflects the ability of ensemble models to capture complex patterns in multidimensional data, which is particularly crucial in medical contexts that demand high reliability in predicting comorbidities such as diabetes and thyroid dysfunctions. in minimizing costly diagnostic errors, whether false positives or failures in detecting positive cases (false negatives). This investigation revealed the crucial steps in improving a machine learning model's performance, including data preprocessing, feature selection, hyperparameter tuning, and cross-validation processes. Nonetheless, a primary concern about this investigation is that it has based its conclusions on a single, somewhat limited, local diabetes dataset. Publishing the developed model to guarantee reliability requires the model to be tested against diversified datasets.
The future plan is interesting in improving the data quality by adding more disease relationships and diagnostic tests for diabetes patients to evaluate their effects on prediction accuracies. Additionally, future work will aim to increase the focus on the development of early detection of many disease manifestations in diabetes and thyroid sufferers, such as cardiovascular diseases, so that this patient group can benefit from more effective preventive and therapeutic care.

References

[1] S. Ghimire, P. Sangroula, I. K.C., R. K. Deo, S. Ghimire, and K. Dhonju, "Spectrum of Thyroid Disorders in Patients with Type-2 Diabetes Mellitus," J Nepal Health Res Counc, vol. 20, no. 4, pp. 922–927, Jul. 2023, https://doi.org/10.33314/jnhrc.v20i4.4314.

[2] M. Hage, M. S. Zantout, and S. T. Azar, "Thyroid Disorders and Diabetes Mellitus," Journal of Thyroid Reseach-SAGE, vol. 2011, p. 7, 2011, https://doi.org/10.4061/2011/439463.

[3] H. Y. Abdulrazaq, I. A. Zaboon, and M. A. Maatook, "Prevalence of thyroid disorders among diabetes mellitus patients in al-Basra southern of Iraq," *ATMPH*, vol. 24, no. 04, 2021, https://doi.org/10.36295/asro.2021.24462.

[4] P. Sharma, S. Shrestha, and P. Kumar, "A review on association between diabetes and thyroid disease," *SUJHS*, vol. 5, no. 2, pp. 50–55, Jan. 2020, https://doi.org/10.18231/j.sujhs.2019.013.

[5] D. Das, "Essentiality, relevance, and efficacy of adjuvant/combinational therapy in the management of thyroid dysfunctions," 2022.

[6] Q. Wang, Z. Zeng, J. Nan, Y. Zheng, and H. Liu, "Cause of Death Among Patients with Thyroid Cancer: A Population-Based Study," *Front. Oncol.*, vol. 12, p. 852347, Mar. 2022, https://doi.org/10.3389/fonc.2022.852347.

[7] M. H. Alshayeji, "Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers," *MAKE*, vol. 5, no. 3, pp. 1195–1213, Sep. 2023, https://doi.org/10.3390/make5030061.

[8] A. A. Alsolami *et al.*, "Association between type 2 diabetes mellitus and hypothyroidism: a case– control study," *IJGM*, vol. Volume 11, pp. 457–461, Dec. 2018, https://doi.org/10.2147/ijgm.s179205.

[9] A. M. Rahmani *et al.*, "Machine Learning (ML) in Medicine: Review, Applications, and Challenges," *Mathematics*, vol. 9, no. 22, p. 2970, Nov. 2021, https://doi.org/10.3390/math9222970.

[10] American Diabetes Association Professional Practice Committee, "Glycemic Targets: *Standards of Medical Care in Diabetes—2022*," *Diabetes Care*, vol. 45, no. Supplement_1, pp. S83–S96, Jan. 2022, https://doi.org/10.2337/dc22-s006.

[11] American Diabetes Association, "Diagnosis and Classification of Diabetes Mellitus," *Diabetes Care*, vol. 32, no. Supplement_1, pp. S62–S67, Jan. 2009, https://doi.org/10.2337/dc09-s062.

[12] Md. J. Hossain, Md. Al-Mamun, and Md. R. Islam, "Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused," Health science reports, vol. 7, no. 3, Mar. 2024, https://doi.org/10.1002/hsr2.2004.

[13] A. Raza, F. Eid, E. C. Montero, I. D. Noya, and I. Ashraf, "Enhanced interpretable thyroid disease diagnosis by leveraging synthetic oversampling and machine learning models," *BMC Med Inform Decis Mak*, vol. 24, no. 1, p. 364, Nov. 2024, https://doi.org/10.1186/s12911-024-02780-0.

[14] E. S. Tumpa and K. Dey, "A Review on Applications of Machine Learning in Healthcare," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India: IEEE, Apr. 2022, pp. 1388–1392. https://doi.org/10.1109/icoei53556.2022.9776844.

[15] G. Kunapuli, *Ensemble Methods for Machine Learning*.

[16] Saleem, Mir. "Improving Classification Accuracy of Machine Learning Algorithms for Thyroid Prediction Using Ensemble Models." International Conference on Innovations in Applied Science and Engineering, 2022, http://www.academia.edu/download/107926504/ICIASE_2022_paper_7889_1_.pdf. Accessed 12 May 2025.

[17] H. O. Sayyid, S. A. Mahmood, and S. S. Hamadi, "A Comparative Analysis of Machine Learning Models for Predicting Thyroid Disorders in Type 1 and Type 2 Diabetic Patients," *J. Basrah Res. (Sci.)*, vol. 50, no. 2, pp. 193–203, Dec. 2024, https://doi.org/10.56714/bjrs.50.2.16.

[18] K. Salman and E. Sonuç, "Thyroid Disease Classification Using Machine Learning Algorithms," *J. Phys.: Conf. Ser.*, vol. 1963, no. 1, p. 012140, Jul. 2021, https://doi.org/10.1088/1742-6596/1963/1/012140.

[19] 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, https://doi.org/10.3390/cancers14163914.

[20] T. Akhtar *et al.*, "Ensemble-based Effective Diagnosis of Thyroid Disorder with Various Feature Selection Techniques.," *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pp. 14–19, 2022, https://doi.org/10.1109/smarttech54121.2022.00019.

[21] M. H. Alshayeji, "Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers," *MAKE*, vol. 5, no. 3, pp. 1195–1213, Sep. 2023, https://doi.org/10.3390/make5030061.

[22] A. Raza, F. Eid, E. C. Montero, I. D. Noya, and I. Ashraf, "Enhanced interpretable thyroid disease diagnosis by leveraging synthetic oversampling and machine learning models," *BMC Med Inform Decis Mak*, vol. 24, no. 1, p. 364, Nov. 2024, https://doi.org/10.1186/s12911-024-02780-0.

[23] M. S. Mir, S. A. Fayaz, M. Zaman, and S. Agrawal, "An Application of Traditional and Ensemble Machine Learning Approaches to Redefine Thyroid Disorder Diagnosis," *MMEP*, vol. 11, no. 9, pp. 2437–2446, Sep. 2024, https://doi.org/10.18280/mmep.110916.

[24] S. Ji, "SSC: The novel self-stack ensemble model for thyroid disease prediction," *PLoS ONE*, vol. 19, no. 1, p. e0295501, Jan. 2024, https://doi.org/10.1371/journal.pone.0295501.

[25] P. Gupta *et al.*, "Detecting Thyroid Disease Using Optimized Machine Learning Model Based on Differential Evolution," *Int J Comput Intell Syst*, vol. 17, no. 1, p. 3, Jan. 2024, https://doi.org/10.1007/s44196-023-00388-2.

[26] P. Sanju, N. S. S. Ahmed, P. Ramachandran, P. M. Sajid, and R. Jayanthi, "Enhancing thyroid disease prediction and comorbidity management through advanced machine learning frameworks," *Clinical eHealth*, vol. 8, pp. 7–16, Dec. 2025, https://doi.org/10.1016/j.ceh.2025.01.002.

[27] R. Islam, A. Sultana, and Md. N. Tuhin, "A comparative analysis of machine learning algorithms with tree-structured parzen estimator for liver disease prediction," *Healthcare Analytics*, vol. 6, p. 100358, Dec. 2024, https://doi.org/10.1016/j.health.2024.100358.

[28] H. O. Sayyid, S. A. Mahmood, and S. S. Hamadi, "Comparison of Machine Learning Algorithms for Predicting Thyroid Disorders in Diabetic Patients," Informatica vol. 49, no. 12, pp. 105–114, Feb. 2025, https://doi.org/10.31449/inf.v49i12.6927.

[29] B. AYINDE, "Thyroid Sickness Determination." 2022. [Online]. Available: https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination

[30] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*. canada: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2019.

[31] T. Fawcett, "Introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006, https://doi.org/10.1016/j.patrec.2005.10.010.