BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Enhanced interpretable thyroid disease diagnosis by leveraging synthetic oversampling and machine learning models

Ali Raza[1], Fatma Eid[2], Elisabeth Caro Montero[3,4,5,6], Irene Delgado Noya[3,7,8] and Imran Ashraf[9*]

Abstract

Thyroid illness encompasses a range of disorders affecting the thyroid gland, leading to either hyperthyroidism or hypothyroidism, which can significantly impact metabolism and overall health. Hypothyroidism can cause a slowdown in bodily processes, leading to symptoms such as fatigue, weight gain, depression, and cold sensitivity. Hyperthyroidism can lead to increased metabolism, causing symptoms like rapid weight loss, anxiety, irritability, and heart palpitations. Prompt diagnosis and appropriate treatment are crucial in managing thyroid disorders and improving patients' quality of life. Thyroid illness affects millions worldwide and can significantly impact their quality of life if left untreated. This research aims to propose an effective artificial intelligence-based approach for the early diagnosis of thyroid illness. An open-access thyroid disease dataset based on 3,772 male and female patient observations is used for this research experiment. This study uses the nominal continuous synthetic minority oversampling technique (SMOTE-NC) for data balancing and a fine-tuned light gradient booster machine (LGBM) technique to diagnose thyroid illness and handle class imbalance problems. The proposed SNL (SMOTE-NC-LGBM) approach outperformed the state-of-the-art approach with high-accuracy performance scores of 0.96. We have also applied advanced machine learning and deep learning methods for comparison to evaluate performance. Hyperparameter optimizations are also conducted to enhance thyroid diagnosis performance. In addition, we have applied the explainable Artificial Intelligence (XAI) mechanism based on Shapley Additive exPlanations (SHAP) to enhance the transparency and interpretability of the proposed method by analyzing the decision-making processes. The proposed research revolutionizes the diagnosis of thyroid disorders efficiently and helps specialties overcome thyroid disorders early.

Keywords Machine learning, Deep learning, Thyroid disorders, Medical diagnosis, Explainable artificial intelligence

*Correspondence:
Imran Ashraf
imranashraf@ynu.ac.kr
[1] Department of Software Engineering, University of Lahore, Lahore 54000, Pakistan
[2] Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea
[3] Universidad Europea del Atlantico, Santander 39011, Spain
[4] Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[5] Universidad de La Romana, La Romana, República Dominicana
[6] Universidad Internacional Iberoamericana Arecibo, Puerto Rico 00613, USA
[7] Universidade Internacional do Cuanza, Cuito, Bie, Angola
[8] Fundacion Universitaria Internacional de Colombia, Bogota, Colombia
[9] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Introduction

Thyroid sickness or disorders refer to a group of medical conditions that affect the normal functioning of the thyroid gland, a small butterfly-shaped gland located in the neck [1]. The thyroid gland regulates various bodily functions, including metabolism, energy production, growth, and development. Common thyroid disorders include hypothyroidism, where the thyroid gland produces insufficient amounts of thyroid hormones, and hyperthyroidism, characterized by an overproduction of thyroid hormones [2, 3]. These conditions can cause a wide range of symptoms, such as fatigue, weight changes, mood swings, and disrupted menstrual cycles. Various factors, including autoimmune diseases, iodine deficiency, genetic predispositions, and certain medications, can cause thyroid disorders. Diagnosis typically involves a combination of clinical examination, blood tests to measure thyroid hormone levels, and imaging tests [4]. Treatment options vary depending on the specific disorder but can include hormone replacement therapy, medication to regulate hormone levels, or, in some cases, surgery.

Thyroid disorders, which include a range of conditions such as hypothyroidism, hyperthyroidism, and thyroid cancer, have been associated with mortality rates and deaths, although to varying degrees [5]. Mortality due to thyroid disorders is predominantly attributed to thyroid cancer, which accounts for most thyroid-related deaths [6]. Thyroid cancer mortality rates have shown a relatively stable trend in recent years, with advances in diagnostic techniques and treatment options contributing to improved outcomes. However, certain subtypes of thyroid cancer, particularly anaplastic thyroid carcinoma, continue to exhibit high mortality rates. Furthermore, while hypothyroidism and hyperthyroidism are generally manageable with appropriate medical interventions, untreated or poorly treated cases can lead to severe complications, potentially increasing the risk of mortality [7]. Although the overall mortality burden of thyroid disorders is comparatively lower than that of other significant diseases, ongoing research aims to refine diagnostic and therapeutic strategies [8], ultimately reducing mortality rates for individuals affected by thyroid disorders. Early detection, accurate diagnosis, and appropriate management of thyroid disorders using an intelligent approach are essential to minimize symptoms, prevent complications, and maintain overall health and well-being.

Thyroid disorders pose a significant challenge for an accurate and timely diagnosis, often requiring extensive clinical evaluation and laboratory tests [9]. However, recent advances in artificial intelligence (AI) present promising opportunities to enhance the diagnostic process [10]. AI systems can leverage machine learning algorithms to analyze large amounts of patient data, including medical records [11], laboratory results [12], and imaging studies [13], to identify patterns and correlations that may indicate medical disease symptoms. Similar thyroid dysfunction can be identified based on medical data fed to the machine learning model [14]. By training AI models in large datasets and incorporating diverse patient populations, AI algorithms can learn to recognize subtle indicators of thyroid disorders and provide valuable information to healthcare professionals [15]. Integrating AI into the diagnosis of thyroid disorders can facilitate the diagnostic process, improve accuracy, and ultimately help early detection and appropriate treatment of the disorder.

Existing methods [5, 16–19] for diagnosing thyroid disorders often exhibit moderate performance scores and rely on traditional classification techniques built on imbalanced datasets. These approaches often fail to address data balance issues, leading to biased results and reduced diagnostic accuracy. Furthermore, the lack of explainable AI models in current diagnostic methods limits the interpretability and trustworthiness of the predictions. The advanced machine learning approach addresses these gaps by incorporating robust data balancing techniques and employing explainable AI models, enhancing the accuracy and transparency of thyroid disorder diagnoses.

In this study, we used an advanced machine learning approach to diagnose thyroid conditions. The dataset utilized in this study contains thyroid disease records consisting of 3,772 observations with 30 features for both male and female genders, which are categorized as 'sick' and 'negative' [20]. The dataset is imbalanced, which required the application of the SMOTE-NC method [21]. By generating synthetic samples, SMOTE-NC helped balance the class distribution, resulting in a more evenly distributed dataset for training machine learning models. For each instance in the minority class, SMOTE-NC selected its k nearest neighbor instances from the same class, based on a chosen distance metric. This technique proves particularly beneficial when dealing with datasets that contain continuous and nominal features, as SMOTE alone would not apply directly to the nominal features [22, 23]. The study introduced a unique combination of the SMOTE-NC method with a fine-tuned LGBM technique to diagnose thyroid illnesses and address class imbalance problems. The proposed LSN approach outperformed the state-of-the-art methods, demonstrating high-accuracy performance scores.

The primary research contributions of the proposed study are as follows:

- We proposed an innovative SNL (SMOTE-NC-LGBM) approach that combines SMOTE-NC with a

fine-tuned LGBM technique to diagnose thyroid illness and address the class imbalance problem, which has been lacking in previous studies. The results demonstrate that the proposed approach outperforms state-of-the-art methods, achieving high performance in the diagnosis of thyroid diseases.
- We comprehensively evaluated the proposed approach against four advanced machine learning and two deep learning methods. In addition, we used a hyperparameter optimization approach to enhance the performance of the proposed approach for the diagnosis of thyroid disease.
- We have applied an eXplainable artificial intelligence (XAI) mechanism based on the SHAP chart to enhance the transparency and interpretability of the proposed method. This approach allows us to analyze the decision-making processes for diagnosing thyroid illness, providing a better understanding of how the model reaches its decision.

The rest of the study is organized as follows. "Literature analysis" section reviews the literature analysis for the detection of thyroid disorders. "Proposed methodology" section presents a stepwise analysis of the proposed methodology. The results of the proposed approach are shown in "Results and discussions" section. Finally, "Conclusions and future work" section concludes the study and highlights the main findings and future work.

Literature analysis

Thyroid disease is a prevalent health problem that affects a large population worldwide. Over the years, there has been growing interest in using machine learning techniques to aid in the early diagnosis of [16]. A comprehensive literature analysis examines the existing research on machine learning applications [5]. This analysis also highlights the challenges and limitations posed during thyroid diagnosis in previous studies. In [24], the study aimed to use machine learning to extract radiomic characteristics from two-dimensional ultrasound (2D-US) and contrast-enhanced ultrasound (CEUS) images for the classification and prediction of benign and malignant thyroid nodules. Conducted retrospectively, the research included 313 thyroid nodules (203 malignant and 110 benign) with pathological diagnoses. The diagnostic performance of both junior and senior radiologists was evaluated, with Area Under the Curve (AUC) scores of 0.755, 0.750, and 0.784 for US, CEUS, and combined US and CEUS assessments by junior radiologists, respectively. Senior radiologists achieved AUCs of 0.800, 0.873, and 0.890. The Random Forest (RF) classifier outperformed other classifiers, achieving an AUC of 1 for the training cohort and 0.94 (95% confidence interval 0.88–1) for the test cohort. This research underscores the potential of combining machine learning with radiomics features from US and CEUS images to enhance the accuracy of thyroid nodule classification.

The research article [17] focuses on the application of various machine learning algorithms for the prediction of hypothyroidism and hyperthyroidism. To conduct their study, the authors utilize three datasets: hypothyroid, hyperthyroid, and sick, which collectively contain 3221 entries. The research aims to identify crucial features that can improve the precision of detecting thyroid diseases. To achieve their goals, the paper undergoes pre-processing and feature selection steps and then applies modified and original data to multiple classification models for thyroid prediction. Notably, the results demonstrate that Random Forest (RF) outperforms other models across all sectors of the dataset, while Naive Bayes performs poorly. The researchers used the RF feature importance method to achieve a remarkable precision of 91. 42%. However, diagnosis performance scores are relatively low, indicating a need for further improvement in this area.

In [25], the study evaluated the efficacy of two distinct classifier models in detecting thyroid issues. A dataset from the UCI repository is used to train and evaluate their models. The study mainly focused on examining the accuracy and precision of the Convolutional Neural Network (CNN) and Support Vector Machine (SVM) algorithms to distinguish between hypothyroidism and hyperthyroidism. The outcomes indicate that the CNN classifier performs better than the SVM classifier, achieving an accuracy of 89% and a precision of 87%. However, it is important to highlight that this study considers the overall performance scores of both classifiers.

Chaganti et al., [18] focused on detecting thyroid disease using a machine learning approach, including SVM, the K-Nearest-Neighbors Algorithm (KNN), Decision Tree, Naive Bayes, and Random Forest. The study was based on collected samples of thyroid datasets from GitHub repositories. The results indicated that SVM, KNN, Decision Tree, and Naive Bayes achieved high accuracy levels, up to 90%. However, the existing Random Forest algorithm only managed an accuracy of approximately 70%. To improve accuracy, the authors introduced Principal Component Analysis (PCA), a dimensionality reduction technique that significantly improved accuracy to around 90%. The study also noted that classical feature engineering approaches still yielded relatively low-performance scores.

Kumar et al., [26] focused on their study on predicting the early stages of thyroid development, specifically emphasizing common types of hypothyroidism. To accomplish this objective, the study incorporates feature selection strategies and various categorization methods.

The data set used in the study is obtained from the UCI data repository, which includes 7,200 different categories of multivariate data, and each record consists of 25 distinct characteristics. Disease identification is performed using a deep convolutional neural network (DeepCNN), while the Gray Wolf Optimizer (GWO) is employed for model training, as both have demonstrated a close association leading to enhanced accuracy. Through dataset refinement, the proposed model achieves an impressive 95% accuracy and 92% specificity in classifying thyroid diseases.

In [27], Nayak et al. introduced a unique multi-instance-based learning technique for the cytopathological diagnosis of thyroid conditions. The technique utilizes Multi-Scale Feature Fusion (MSF) and Convolutional Neural Networks (CNN) to process Whole Slide Images (WSIs) containing multiple occurrences per section. The architecture is designed to identify significant sections within the images automatically. The approach achieves improved classification results by incorporating a feature-fusion architecture that combines minimal features through an instance-level awareness model. The proposed model undergoes extensive training and validation using clinical data, demonstrating an impressive accuracy of 93.2%, outperforming all existing methods. Moreover, the model's superiority becomes evident when compared to a modern deep multi-instance technique applied to a publicly available histopathology dataset.

In [25], the study investigated the effectiveness of Support Vector Machine (SVM) classifiers and Logistic Regression models in predicting and classifying thyroid disease. They argue that SVM classifiers outperform logistic regression models in terms of accuracy and precision during performance evaluation. To test their proposed prediction model, the authors used an original dataset obtained from Sawai Man Singh Hospital (SMS) in India. The experimental findings indicate that the SVM classifier achieved a precision of 84% and an overall accuracy of 86%, showing performance scores lower than the baseline.

Alyas et al. [28] investigated the use of various machine learning algorithms for the classification of thyroid diseases. The study performed a comparative analysis of the performance of the decision tree, random forest algorithm, KNN, and artificial neural networks on a dataset obtained from the UCI thyroid disease repository. Furthermore, they perform the classification on both sampled and unsampled datasets to enable a comprehensive comparison. As a result of their analysis, the random forest algorithm achieved the highest accuracy, scoring 94.8%, with a specificity of 91%.

In [19], the study investigated the application of machine learning algorithms to assess the risk of thyroid disease. For this purpose, the authors utilized the Sick-euthyroid dataset [29]. Since the dataset contains imbalanced target variable classes, relying solely on the accuracy score might not accurately reflect the prediction performance. To mitigate this issue, the evaluation metric considers both accuracy and recall ratings. Additionally, the F1 score, which offers a balanced measure of precision and recall for uneven class distributions, serves as a crucial performance metric for the machine learning algorithms employed in the study. The findings suggest that the ANN Classifier achieved the highest performance, surpassing the other nine studied algorithms in accuracy when predicting the risk of thyroid disease with an F1-score of 95%.

In [30], Pal et al., dedicated their efforts to designing an approach for the detection of thyroid disease using machine learning algorithms, including KNN, decision tree (DT), and multilayer perceptron (MLP). The authors obtained a thyroid disease dataset from the UCI repository. To evaluate the models' performance, accuracy and area under the curve were used as metrics. The results of the comparative analysis revealed that the multilayer perceptron (MLP) outperformed the other models in accurately classifying thyroid disease, achieving an impressive accuracy of 95.73% and an Area Under the Curve (AUC) value of 94.23%. The study involved a substantial dataset of 3,163 cases with 24 thyroid characteristics. An analysis of existing works on the diagnosis of thyroid illness is given in Table 1.

Proposed methodology

The proposed research methodology is illustrated stepwise in Fig. 1. To begin the experiments, we utilized an open-access thyroid disease dataset consisting of 3,772 observations with 30 features. Upon analyzing the dataset, we identified an imbalance issue. First, we split the dataset into training and testing sets, using 90% of the data for training purposes. The training data was then input into SMOTE-NC for data balancing. The resulting balanced dataset was used to train various machine learning and deep learning models with hyperparameter tuning. Each model's performance was evaluated using unseen test data, which accounted for 10% of the original dataset. The best-performing AI model was selected for diagnosing thyroid illness. Finally, we applied the XAI mechanism to analyze the decision-making processes of the chosen model, providing a better understanding of how the model reaches its conclusions.

Thyroid sickness data

This study used an open-access dataset on thyroid disease [20] acquired from a famous Kaggle repository. The downloaded data set contains thyroid disease records

Table 1 The literature summary analysis for diagnosis of thyroid illness

Ref.	Year	Dataset	Technique	Preprocessing	Validation	Performance accuracy
[17]	2023	Three datasets with 3221 patients	RF	Yes	train-test	0.91
[25]	2023	UCI Thyroid Disease data	CNN	No	train-test	0.89
[18]	2023	Thyroid dataset from GitHub repository.	SVM	Yes	cross validation	0.90
[26]	2023	UCI thyroid illness data with 7200 patients	CNN	Yes	cross validation	0.95
[27]	2023	Histopathology dataset	CNN	Yes	train-test	0.93
[25]	2023	Thyroid data obtained from Sawai Man Singh (SMS) hospital in India.	SVM	No	train-test	0.86
[28]	2022	UCI thyroid illness data with 7200 patients	RF	Yes	train-test	0.84
[19]	2022	Sick-euthyroid dataset	ANN	Yes	cross validation	0.95
[30]	2022	UCI dataset consisting of 3163 patients	MP	No	train-test	0.95

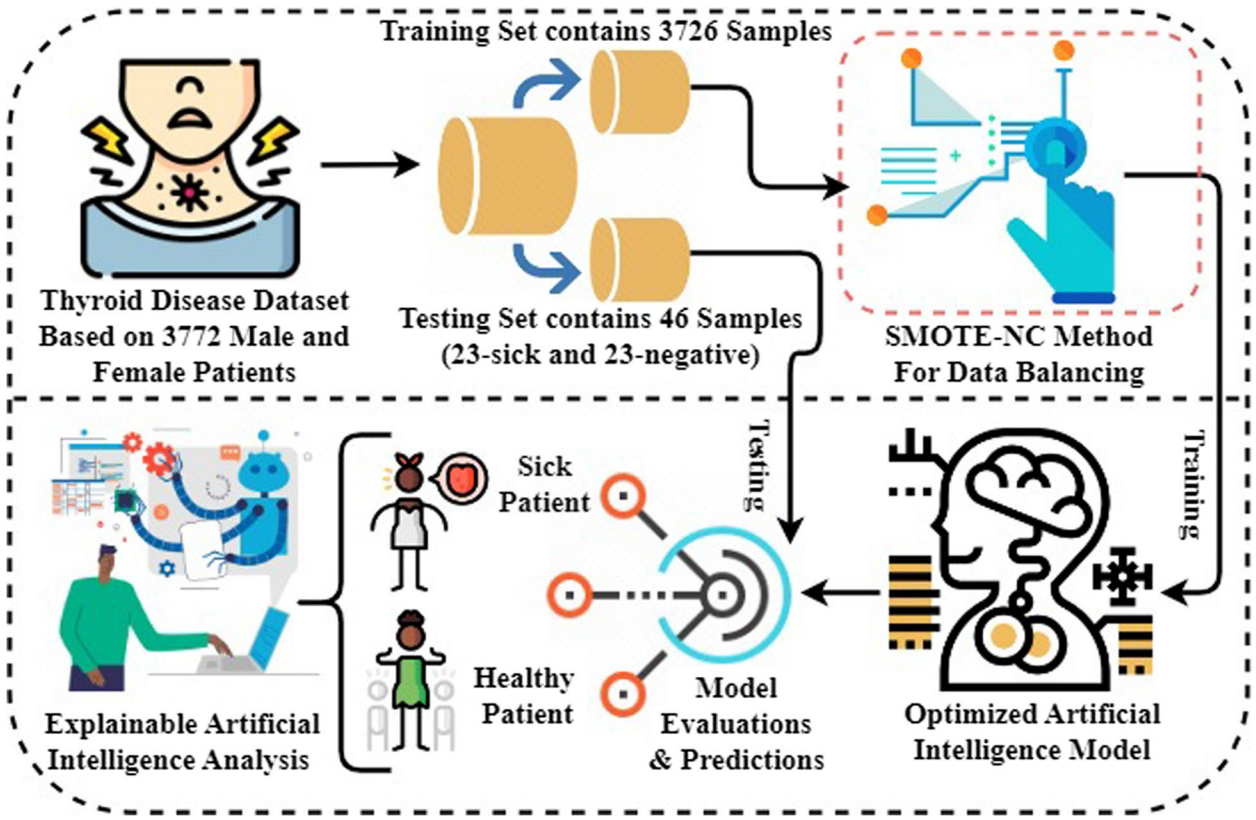


Fig. 1 The proposed methodology for the diagnosis of thyroid illness

collected and supplied by the Garavan Institute and J. Ross Quinlan of the New South Wales Institute, Sydney, Australia, in 1987. The dataset consists of 3,772 observations with 30 features. The target thyroid class is binary, categorized as 'sick' and 'negative.' The observations in the dataset cover both male and female genders. We also analyzed the correlation of dataset features, as shown in Fig. 2. The analysis demonstrates that all the features of the data set exhibit strong correlations among them to diagnose thyroid disease.

Proposed SNL approach

In this section, we analyze the proposed unique SNL research approach. Initially, we observed that the target class of the dataset is unbalanced, as illustrated in Fig. 3a. The analysis reveals that the 'sick' class contains a low

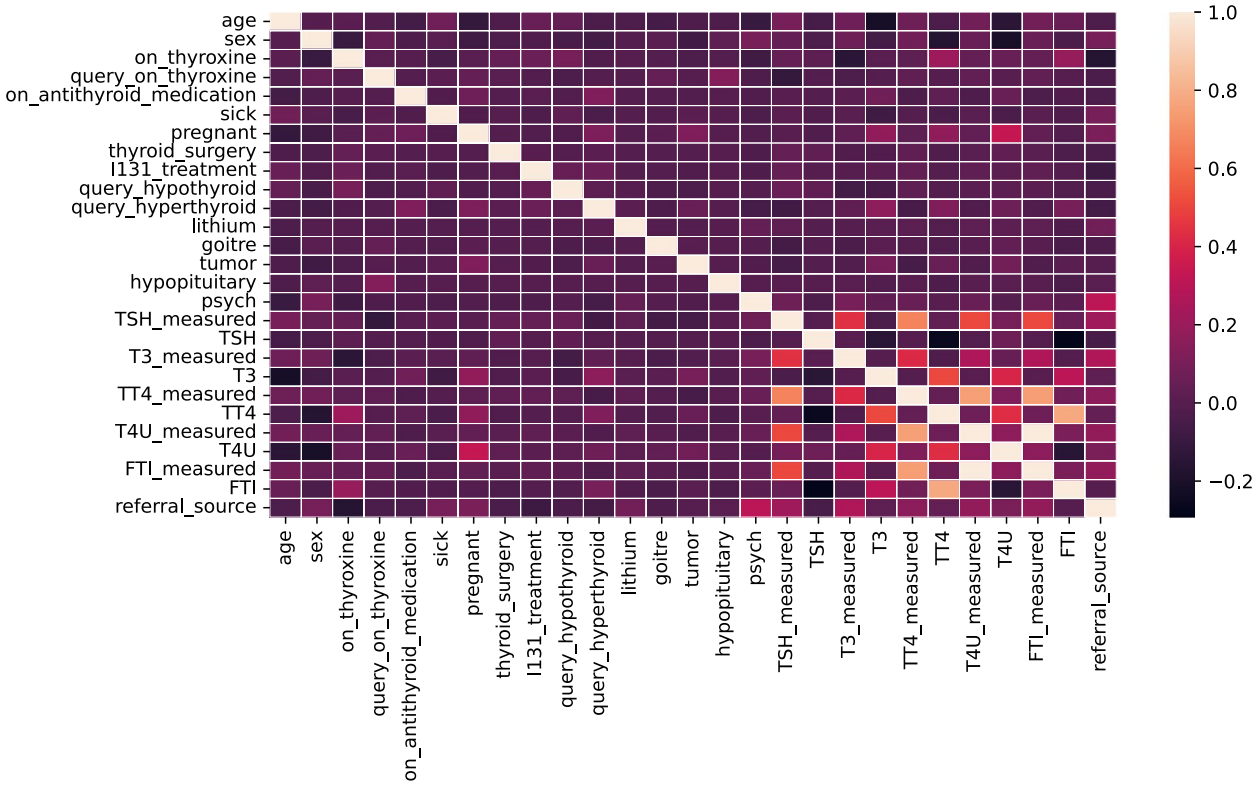
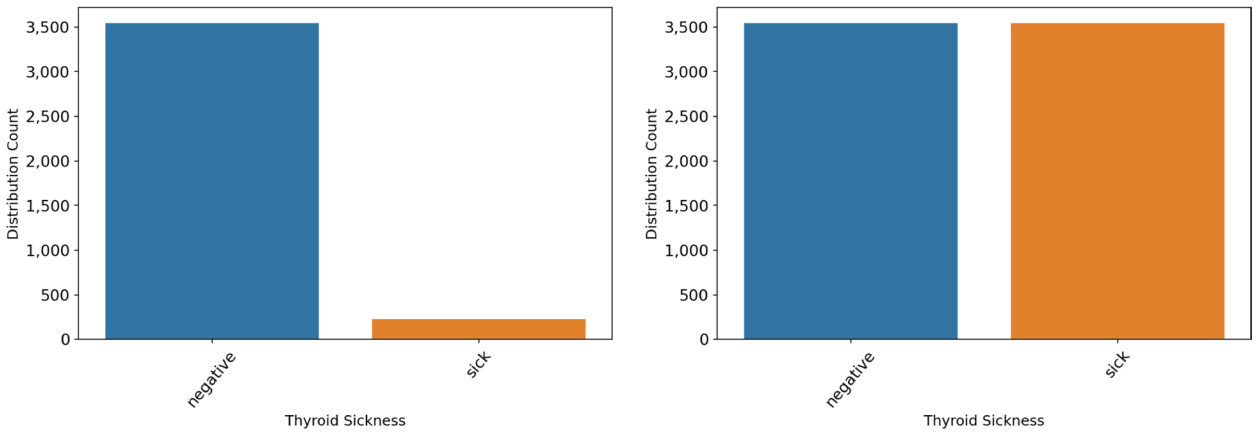


Fig. 2 The correlation analysis of thyroid disease-related features



(a) Without data balancing (b) With data balancing

Fig. 3 The distribution analysis of target label for thyroid illness diagnosis

number of samples, which consequently impacts the performance of thyroid diagnosis scores.

For data balancing, we applied the Synthetic minority oversampling TEchnique nominal continuous(SMOTE-NC) method, and the results of data balancing are shown in Fig. 3b. SMOTE-NC is an advanced variant of the SMOTE algorithm designed to address class imbalance in datasets containing both categorical (nominal) and continuous features. SMOTE-NC effectively generates realistic synthetic data, improving the performance of classifiers on imbalanced datasets with mixed feature types [21]. The use of the SMOTE-NC method allowed

the applied Light Gradient Boosting Machine (LGBM) technique to achieve high-performance scores for thyroid diagnosis, as described in "Performance results with machine learning" section.

SMOTE-NC is chosen to balance the dataset after thoroughly evaluating various techniques due to its superior performance in handling numerical and categorical data [21]. Compared to other methods such as random under-sampling, random over-sampling, and simple SMOTE, SMOTE-NC effectively preserves the underlying data structure while generating synthetic samples [31], thus enhancing the classifier's ability to generalize from imbalanced data. Its widespread application and validation in the literature further reinforce its reliability. Consequently, SMOTE-NC is a robust and versatile technique for addressing class imbalance issues in datasets with mixed feature types.

SMOTE-NC mathematical mechanism

Let N be the number of minority samples, and N_{new} be the desired number of synthetic samples to be generated. You can set N_{new} based on the desired balance level. For example, suppose that you want the minority class to be represented at a percentage p of the majority class after over-sampling. In that case, you can set $N_{\text{new}} = p \times$ (number of majority samples).

To generate synthetic samples using SMOTE-NC:

1. Randomly select a minority sample x_{min} from the dataset.
2. For each selected minority sample x_{min}, identify its k nearest neighbors (k-NN) from both the minority and majority classes. The k-NN can be determined using a distance metric such as the Euclidean distance.
3. For each of the k-NN, calculate the difference vector diff between the feature values of the current neighbor and the selected minority sample x_{min}.
4. Generate synthetic samples x_{syn} for each k-NN as follows:
   - For each feature j:
     - If the feature is nominal:
       * Randomly choose one of the nominal values from the current neighbor and the selected minority sample x_{min}.
       * Assign the chosen nominal value to the synthetic sample x_{syn} for the feature j.
     - If the feature is continuous:
       * Calculate the difference ratio rat for the current neighbor: rat = random_number() (A random number between 0 and 1)
       * Calculate the feature value for the synthetic sample x_{syn} for feature j:

$$x_{\text{syn}}[j] = x_{\text{min}}[j] + \text{rat} \times \text{diff}[j]$$

5. Add the generated synthetic samples to the minority class, thus increasing its size.

Applied artificial intelligence methods

Applied artificial intelligence (AI) methods have emerged as promising tools for diagnosing thyroid diseases [32, 33], revolutionizing the field of medical diagnostics. Thyroid disorders are often challenging to diagnose accurately due to the subtle and diverse nature of symptoms. AI-powered diagnostic systems can detect patterns and associations that might escape the human eye [34–38], enabling early and precise identification of thyroid conditions such as hypothyroidism, hyperthyroidism, and thyroid nodules.

Logistic regression

Logistic Regression (LR) is a widely used method for diagnosing thyroid illness due to its effectiveness in binary classification tasks [39]. The LR algorithm is a type of supervised learning technique that aims to predict the probability of an event occurring; in this case, it is used to determine the presence or absence of thyroid disease. The algorithm operates by modeling the relationship between a set of input features and the binary output variable representing the diagnosis. The working principle of LR involves transforming the linear combination of input features using the logistic function, also known as the sigmoid function. The LR equation for the diagnosis of thyroid illness is given by:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \quad (1)$$

Where:

$P(y = 1|\mathbf{x})$ is the probability of thyroid illness diagnosis

$\mathbf{x}$ is the input feature vector

$\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients to be learned

$x_1, x_2, \ldots, x_n$ are the input features

Linear support vector machines

The Linear Support Vector Machines (LSVM) method can be used for the diagnosis of thyroid illness [40]. In the binary classification setting, the LSVM algorithm aims

to find the optimal hyperplane that separates the two classes in feature space.

Given a training dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^d$ represents the feature vector of the i-th patient and $y_i \in \{-1, 1\}$ is the corresponding class label indicating whether the patient is healthy ($y_i = -1$) or has a thyroid illness ($y_i = 1$).

The LSVM seeks to find the optimal hyperplane represented by the equation:

$$\mathbf{w}^T \cdot \mathbf{x} + b = 0 \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, and $b \in \mathbb{R}$ is the bias term.

The decision function of the LSVM is defined as:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \cdot \mathbf{x} + b) \tag{3}$$

where sign$(\cdot)$ is the sign function that returns $+1$ if the argument is positive, -1 if the argument is negative, and 0 if the argument is zero.

We want to maximize the margin between the two classes to find the optimal hyperplane. The margin is the perpendicular distance from any training sample to the hyperplane. Mathematically, the margin is given by:

$$\text{margin} = \frac{1}{\|\mathbf{w}\|} \tag{4}$$

Subject to the constraint that for all i:

$$y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 \tag{5}$$

The optimization problem can be formulated as:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{6}$$

$$\text{subject to} \quad y_i(\mathbf{w}^T \cdot \mathbf{x}_i + b) \geq 1 \quad \text{for all } i \tag{7}$$

Once the optimal hyperplane is obtained, we can use the decision function $f(\mathbf{x})$ to predict the class of a new patient based on their feature vector $\mathbf{x}$.

Random forest

Random Forest (RF) is an ensemble learning method that combines multiple decision trees to make predictions [18, 41]. For the diagnosis of thyroid illness, we can represent the Random Forest model as follows:

Let X be the feature matrix representing the input data with n samples and m features. Each sample is denoted by x_i for $i = 1, \ldots, n$, and the corresponding target labels are represented by y_i.

RF algorithm generates B decision trees $T_1, T_2, \ldots, T_B$ by using bootstrapped samples from the original data with replacement. Each tree is trained on a random subset of the features. Let T_b represent the b-th tree, where $b = 1, \ldots, B$.

The prediction of the RF model for a new input sample x_{new} is obtained by aggregating the predictions of individual decision trees. For classification tasks, this aggregation is typically done by majority voting. Let y_{new} be the predicted class for the new sample x_{new}.

The predicted class y_{new} is given by:

$$y_{\text{new}} = \text{mode}(T_1(x_{\text{new}}), T_2(x_{\text{new}}), \ldots, T_B(x_{\text{new}})), \tag{8}$$

where mode represents the majority voting function.

For regression tasks, the predictions of individual trees are averaged to obtain the final prediction. Let f_{new} be the predicted value for the new sample x_{new}.

The predicted value f_{new} is given by:

$$f_{\text{new}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x_{\text{new}}). \tag{9}$$

Light gradient boosting machine

Light Gradient Boosting Machine (LGBM) is a popular gradient boosting framework used for both classification and regression tasks [42]. It is an ensemble learning method that combines the predictions of several weak learners (typically decision trees) to build a more accurate and robust model. The general idea of LGBM can be summarized as follows.

Let $\{(x_i, y_i)\}_{i=1}^n$ be the training dataset, where x_i represents the features of the i-th instance, and y_i is the corresponding label.

Prediction of the m-th Tree: The prediction of the m-th tree is given by:

$$\hat{y}^{(m)} = f_m(x) = f_{m-1}(x) + T_m(x; \Theta_m)$$

where $f_m(x)$ is the prediction of the m-th tree, $f_{m-1}(x)$ is the prediction of the $(m-1)$-th ensemble, and $T_m(x; \Theta_m)$ is the prediction of the m-th tree with its parameters Θ_m.

Objective Function (Loss Function): The objective function is defined as:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m)$$

Where $L(y, \hat{y})$ is the loss function that measures the discrepancy between the predicted value $\hat{y}$ and the true label y, and $\Omega(f_m)$ is the regularization term for the m-th tree.

Training Process: The LGBM training process involves finding the parameters Θ_m for each tree that minimizes the objective function $\mathcal{L}$.

Gated recurrent unit

The Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) that can be used to diagnose thyroid disease [43]. It has two main gating mechanisms, the reset gate and the update gate.

The update gate z_t determines how much of the previous hidden state h_{t-1} should be retained, and the reset gate r_t decides how much of the previous hidden state should be ignored. The candidate hidden state $\tilde{h}_t$ is computed as follows:

$$\tilde{h}_t = \tanh(W \cdot (r_t \odot h_{t-1}) + U \cdot x_t) \tag{10}$$

Where:

- x_t is the input at time step t,
- h_{t-1} is the hidden state at the previous time step,
- W and U are weight matrices,
- $\odot$ represents element-wise multiplication.

The update gate z_t and the reset gate r_t are computed as follows:

$$z_t = \sigma(W_z \cdot x_t + U_z \cdot h_{t-1}) \tag{11}$$

$$r_t = \sigma(W_r \cdot x_t + U_r \cdot h_{t-1}) \tag{12}$$

Where:

- σ is the sigmoid activation function,
- W_z, W_r, U_z, and U_r are weight matrices for the update and reset gates.

The final hidden state h_t is computed by combining the previous hidden state with the candidate hidden state:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{13}$$

The output of the GRU at time step t can be used to diagnose thyroid illness based on the specific task and dataset. Table 2 shows the layer-wise architecture of the GRU model.

Long short-term memory

Long short-term memory (LSTM) [44] is a type of recurrent neural network (RNN) that is well suited for handling sequential data such as time series, making it a useful approach to diagnosing diseases based on sequential medical data. The LSTM architecture consists of several equations that govern the network's information flow. These equations involve various matrices and vectors representing input, output, cell states, and activation functions. In this example, we have provided a simplified version of the LSTM equations for illustration. The following are the mathematical equations of the LSTM method for diagnosing thyroid disease.

The LSTM cell has three main gates (input, forget, and output gates) that control the flow of information.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{14}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{15}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{16}$$

$$C'_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{17}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot C'_t \tag{18}$$

$$h_t = o_t \odot \tanh(C_t) \tag{19}$$

The final output is obtained by passing the last hidden state h_T through a fully connected layer:

$$y = \sigma(W_{hy}h_T + b_y) \tag{20}$$

Table 3 shows the layer-wise architecture of the LSTM model.

Explainable artificial intelligence

In this study, the main objective is to diagnose thyroid disease using XAI with a SHAP chart [45, 46]. The XAI method provides insights into how the AI model makes predictions, allowing us to interpret the results more effectively. The SHAP chart is a method used to interpret

Table 2 GRU model layer architecture

Layer	Shape	Param #
gru_2 (GRU)	(None, 64)	12864
dense_4 (Dense)	(None, 32)	2080
dropout_2 (Dropout)	(None, 32)	0
dense_5 (Dense)	(None, 1)	33
Total params	14,977	

Table 3 LSTM model layer architecture

Layer	Shape	Param #
lstm (LSTM)	(None, 64)	16896
dense_6 (Dense)	(None, 32)	2080
dropout_2 (Dropout)	(None, 32)	0
dense_5 (Dense)	(None, 1)	33
Total params	14,977	

the predictions made by machine learning models. It helps us to understand the contribution of each feature to the final prediction.

For the proposed Thyroid Illness diagnosis model, we use a machine learning model f that takes a set of features $x = (x_1, x_2, \ldots, x_n)$ as input and outputs a prediction $f(x)$. The SHAP value for each feature x_i is defined as follows:

$$\text{SHAP}(x_i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [f(x_S \cup \{x_i\}) - f(x_S)] \quad (21)$$

Where

- N is the set of all features in the model ($N = \{x_1, x_2, \ldots, x_n\}$).
- S is a subset of N excluding the feature x_i.
- x_S is the input with features from the subset S fixed at a reference value.
- $f(x_S \cup \{x_i\})$ is the model's prediction when including the feature x_i with input x_S.
- $f(x_S)$ is the prediction of the model with input x_S.

The SHAP values represent the average contribution of each feature to the predictions in all possible combinations of features.

For the specific cases of diagnosis of thyroid disease, we have a set of characteristics related to patient health and test results. Let $x = (x_1, x_2, \ldots, x_n)$ represent the feature vector for a patient.

The machine learning model f takes x as input and predicts the patient's probability of thyroid disease, denoted $P(\text{Thyroid Illness}|x)$.

The SHAP chart allows us to visualize how each feature contributes to the final prediction probability. Positive SHAP values indicate that the feature positively influences the prediction, while negative values indicate a negative influence.

This study used the SHAP chart as an Explainable AI method to diagnose Thyroid Illness. The SHAP values provide valuable insights into the contribution of each feature to the model's predictions, enhancing the interpretability of the AI system.

Results and discussions

In this section, we present the results obtained from this study on the use of AI methods for thyroid diagnosis. The study aimed to develop and evaluate an AI-based model that assists in accurately and efficiently diagnosing thyroid disorders. The data set used for AI model training and performance evaluations contained numerous medical records related to thyroid, including patient demographics, symptoms, and laboratory test results.

Experimental setup

The Python 3.0 programming language was used to conduct the experiments. The study's experiments were performed in an environment with a graphics processing unit (GPU) back-end, 16GB of RAM, and 90GB of disk space. The machine learning libraries employed in this study include Scikit-learn, Keras, TensorFlow, NumPy, Seaborn, and Matplotlib. We used f1 score, recall, precision, and accuracy for performance evaluations as metrics.

Hyperparameter tuning

In this study, we applied the hyperparameter tuning mechanism to enhance the performance of the applied deep learning and machine learning methods for thyroid diagnosis. The hyperparameter tuning mechanism is based on a recursive training and testing process to find the optimal hyperparameters. The k-fold cross-valuation approach is also used for the selection of the optimal performance parameters. We selected a systematic approach to carefully explore and control the hyperparameters, such as learning rate, batch size, and number of hidden layers, before training the model. The rationale for choosing systematic parameter tuning over other approaches is that this approach has low computational complexity. While other approaches, such as grid search, random search, or Bayesian optimization might provide slightly better parameter-fit for models, they cost in terms of computational resources. The objective of this study is to set a baseline performance for thyroid detection, systematic tuning is a more practical and feasible approach. In addition, due to the complex architecture of the proposed approach, selecting grid search or Bayesian optimization requires sophisticated parameter optimization. On the other hand, systemic hyperparameter tuning seems a more realistic approach. The best-fit hyperparameters for this study are analyzed in Table 4.

Performance results with machine learning

In this section, the performance of applied machine learning approaches for diagnosing thyroid disease is analyzed. The precision of performance and the results of the classification report for each machine learning method are outlined in Table 5. The analysis demonstrates that the linear models LR and LSVM achieved poor performance scores, leading to the conclusion that the features of the dataset are not highly linearly separable. In contrast, the tree-based models RF and LGBM exhibited strong performance in this analysis. The proposed LGBM method, specifically, achieved the highest accuracy score of 0.96 for diagnosing thyroid disease.

Table 4 Hyperparameter tuning analysis

Technique	Hyperparameter description
LR	random_state=0, max_iter=300, multi_class='auto', C=1.0
LSVM	random_state=0, max_iter=500, multi_class='auto', C=1.0
RF	n_estimators=200, max_depth=200, random_state=0
LGBM	n_estimators=300, boosting_type='gbdt', num_leaves=31, importance_type='split'
GRU	loss = 'binary_crossentropy', activation='sigmoid', metrics='accuracy', optimizer = 'adam', epochs=20
LSTM	loss = 'binary_crossentropy', activation='sigmoid', metrics='accuracy', optimizer = 'adam', epochs=20

Table 5 Performance analysis of applied machine learning methods for unseen test data

Technique	Accuracy	Target class	Precision	Recall	F1
LR	0.80	Negative	0.75	0.91	0.82
		Sick	0.89	0.70	0.78
		Average	0.82	0.80	0.80
LSVM	0.80	Negative	0.75	0.91	0.82
		Sick	0.89	0.70	0.78
		Average	0.82	0.80	0.80
RF	0.93	Negative	0.88	1.00	0.94
		Sick	1.00	0.87	0.93
		Average	0.94	0.93	0.93
LGBM	**0.96**	**Negative**	**0.92**	**1.00**	**0.96**
		Sick	**1.00**	**0.91**	**0.95**
		Average	**0.96**	**0.96**	**0.96**

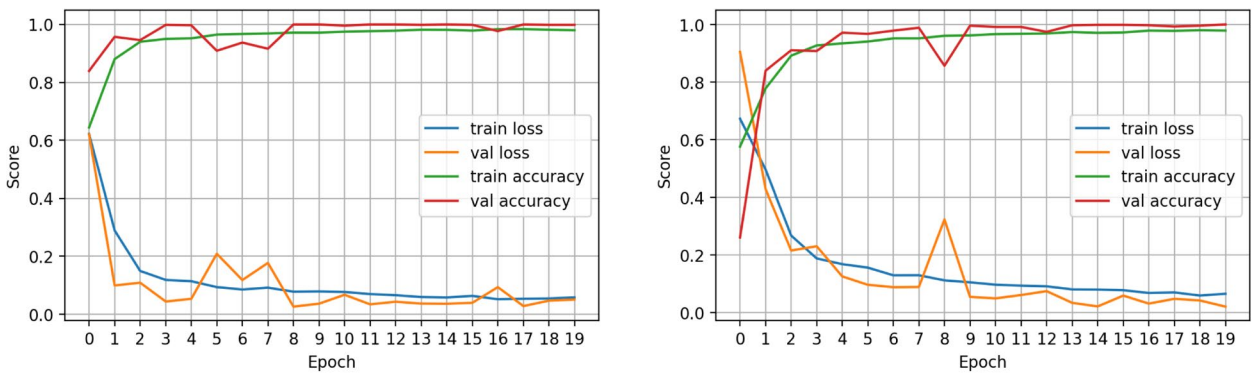
Performance results with deep learning

For a detailed comparison, we utilized an advanced deep learning approach and presented the performance results in this section. Performance evaluation of the GRU and LSTM methods applied during training is shown in Fig. 4.

Each deep learning model was run for 20 epochs, and training and validation outcomes were assessed. The analysis reveals that, in the initial two epochs, both models exhibited high loss and low accuracy scores, attributable to the random weights assigned by each neural network model. Subsequently, each learning model updated the weights, improving performance scores and decreasing loss scores. The analysis concludes that both models achieved high accuracy scores of over 0.90 for both training and validation.

Performance metric scores for unseen test data from applied deep learning models are analyzed in Table 6. Performance analysis shows that the GRU method achieved an acceptable performance score of 0.89 but did not achieve the highest score. However, the applied LSTM achieved good accuracy scores of 0.93 and precision scores of 0.94. This analysis concludes that the deep learning models achieved acceptable scores, which could be further improved by training on a larger amount of data. Figure 5 visualizes the performance of various machine deep learning models for thyroid detection indicating the superior performance of the LGBM model with a 0.96 accuracy score.

In this analysis, the applied GRU and LSTM models underperform compared to LightGBM because they are designed for large datasets with complex temporal patterns. With only 3,772 observations, the models might not



(a) GRU (b) LSTM

Fig. 4 The time series-based performance analysis of deep learning approaches during the training process

Table 6 Performance analysis of applied deep learning methods for unseen test data

Technique	Accuracy	Target class	Precision	Recall	F1
GRU	0.89	Negative	0.82	1.00	0.90
		Sick	1.00	0.78	0.88
		Average	0.91	0.89	0.89
LSTM	0.93	Negative	0.95	0.91	0.93
		Sick	0.92	0.96	0.94
		Average	0.94	0.93	0.93
LGBM	**0.96**	**Negative**	**0.92**	**1.00**	**0.96**
		Sick	**1.00**	**0.91**	**0.95**
		Average	**0.96**	**0.96**	**0.96**

learn complex patterns and interdependencies and models might not be trained well. The LSTM and GRU have many parameters to optimize, which can be excessive for smaller datasets. LightGBM, being a tree-based model, typically requires fewer data to train effectively and can capture simpler patterns with a lower risk of overfitting. The LightGBM relies on structured feature input and can perform well with minimal preprocessing. Deep learning models often benefit from high-dimensional data with temporal dependencies. These factors can contribute to LightGBM's advantage in handling moderately sized, structured datasets over GRU and LSTM models.

Confusion matrix and histogram results analysis

The confusion matrix analysis is conducted to analyze the strengths and weaknesses of applied machine learning and deep learning approaches during diagnosis. Performance analysis using the confusion matrix is illustrated in Fig. 6. The analysis shows that a high error rate is observed, resulting from incorrect predictions, in the LR and LSVC methods. However, other methods achieved acceptable performance scores. The proposed LGBM demonstrated a minimum misclassification rate for unseen testing data, thereby validating the high performance of the proposed approach for thyroid diagnosis.

The comparisons of the accuracy performance of the applied deep learning and machine learning methods, based on histograms, are illustrated in Fig. 5. The analysis reveals that the linear models LR and LSVM achieved a lower performance compared to others. The proposed tree-based LGBM method achieved high-performance scores for the diagnosis of thyroid conditions. Figure 7 shows the receiver operating characteristic curve (ROC) that outperforms the LGBM model. The analysis illustrated that the proposed model achieved high ROC performance scores for the detection of thyroid syndrome.

Computational complexity analysis

In this section, we perform a computational complexity performance analysis of applied machine learning and deep learning methods to diagnose thyroid problems. Runtime computation scores are evaluated in seconds for each model built on the dataset, as shown in Table 7. The analysis reveals that machine learning-based LR and LSVM achieved low computational times; however, they also exhibited low-performance scores. Conversely, the deep learning methods demonstrated the highest computation scores. Consequently, in the comparative analysis, the LGBM method exhibited the best performance scores.

XAI results analysis

The results of the XAI analysis using the proposed method are illustrated in Fig. 8. This analysis determines

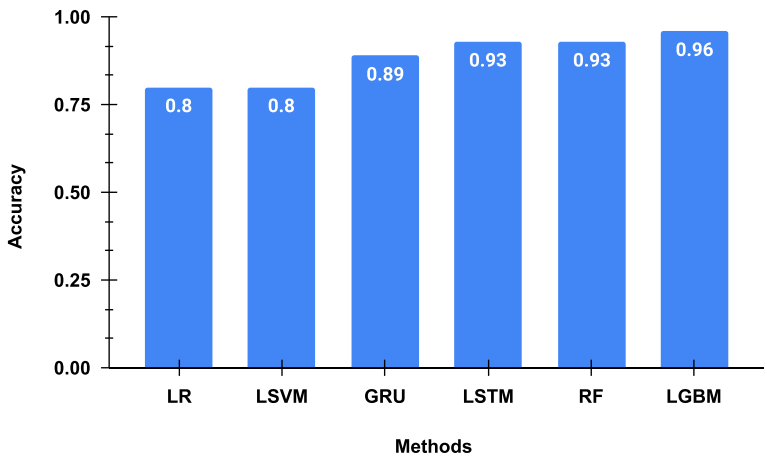


Fig. 5 Performance analysis of applied machine learning and deep learning

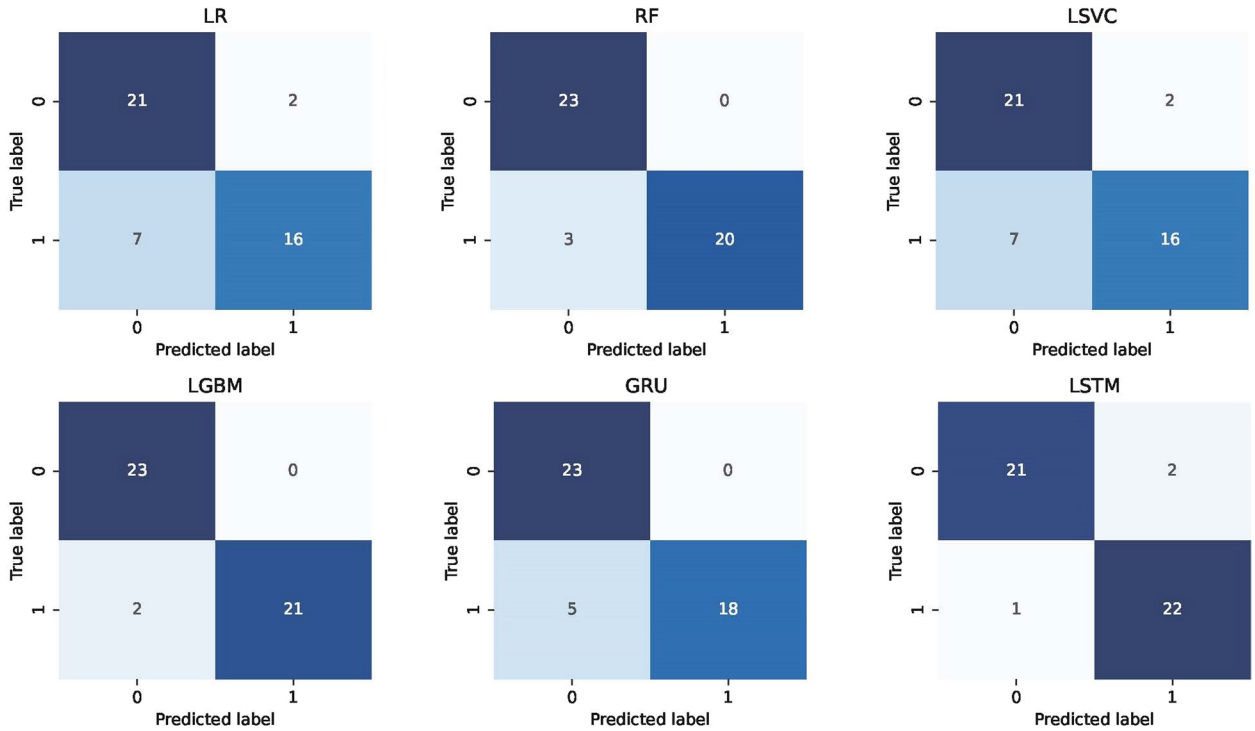


Fig. 6 The confusion matrix performance analysis of applied methods

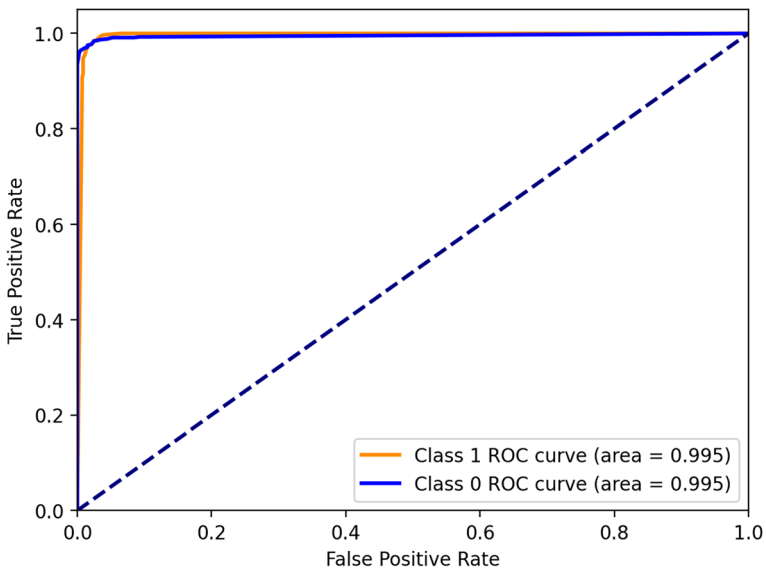


Fig. 7 The ROC curve analysis of outperformed LGBM model

the contribution of each dataset feature to the decisions made by the proposed method during thyroid diagnosis. The importance scores of dataset features are arranged in descending order in the SHAP chart analysis. This analysis reveals that features T3, referral_source, FTI, TT4, age, T4U, TSH, and on_thyroxine play a significant role in diagnosing thyroid disease using the proposed model.

The SHAP analysis identified T3, TT4, TSH, and FTI as key features for diagnosing thyroid disease, enhancing the model's predictive accuracy. T3 (triiodothyronine) is active, while TT4 (total thyroxine) shows total hormone production by the thyroid. Changes in T3 and TT4 levels signal hyperthyroid and hypothyroid conditions, affecting metabolic rates, heart rate, body temperature, and

Table 7 Computational complexity analysis of employed machine and deep learning approaches

Technique		Runtime computation (seconds)
ML	LR	0.713
	SVM	0.307
	RF	3.603
	LGBM	1.060
DL	GRU	43.17
	LSTM	88.17

energy levels. T3 is particularly valuable because it is the active form of thyroid hormone, directly affecting cellular metabolism and energy balance. Elevated T3 levels can signify hyperthyroidism, while decreased levels may indicate hypothyroidism, making it a critical marker in distinguishing between thyroid conditions. TT4, which encompasses both bound and unbound forms of thyroxine, provides a broader measure of thyroid output, helping to capture overall thyroid health and identify abnormalities. Combined with TSH (thyroid-stimulating hormone), which regulates thyroid function, and FTI (Free Thyroxine Index), these features together offer a comprehensive view of thyroid functionality. The model's reliance on these features aligns well with clinical diagnostic practices, supporting the model's interpretability

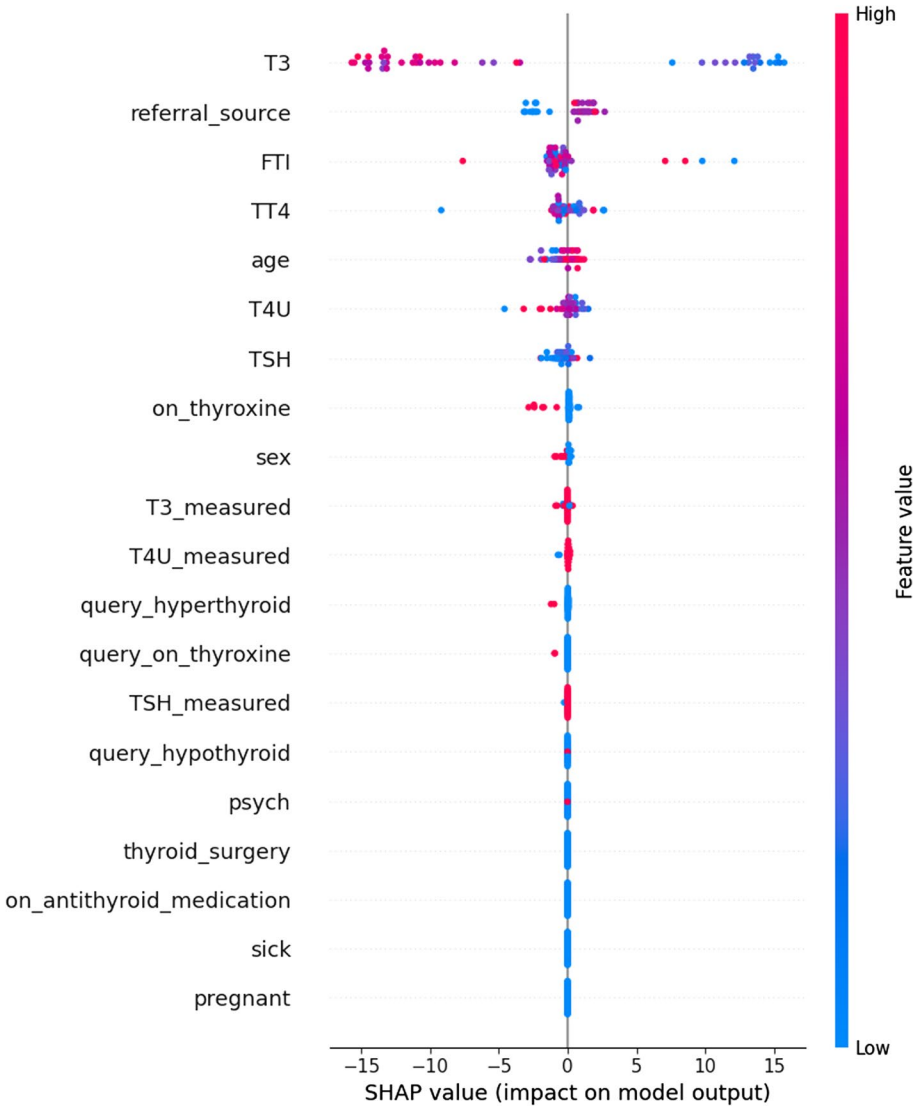


Fig. 8 The SHAP chart-based XAI analysis of the proposed method for diagnosis of the thyroid

for practitioners and adding relevance to its predictions in a clinical setting.

Performance fairness analysis of proposed model

This section analyzes the performance fairness analysis of an innovative SNL-proposed machine learning approach. Performance scores for male and female patients are presented in Table 8. Next, we have divided the data into two age groups: young and old, and evaluated the performance scores in Table 9. This analysis demonstrates that the proposed approach achieved high-performance scores in real-world scenarios.

Comparison with state-of-the-art studies

We have compared the performance of the proposed research approach with current studies and evaluated the results as reported in Table 10. The analysis concludes that the proposed approach outperformed state-of-the-art studies with high-performance scores for the diagnosis of thyroid disorders.

Study limitations

While this study provides valuable insights and demonstrates high accuracy in diagnosing thyroid disease, several limitations should be acknowledged to contextualize the results and identify areas for future improvement. First, the dataset used in this study was collected in 1987, which may limit its applicability to current clinical settings, as it may not fully represent modern trends in thyroid disease or reflect advances in medical understanding. Additionally, the dataset lacks detailed demographic information, such as ethnicity and socioeconomic background, which could affect thyroid disease presentation and progression. This limitation may impact the model's generalizability to diverse patient populations not represented in the dataset.

Another consideration is the absence of temporal data, as this study relies on static observations. Thyroid disease diagnosis and management often benefit from longitudinal data that captures changes in health indicators over time. Future work incorporating time-series data could better support diagnostic decisions by allowing the model to monitor disease progression across patient treatment timelines. Moreover, the model was evaluated using a single dataset without external validation of independent datasets. Such validation would be essential to verify the model's robustness and adaptability across different clinical settings, minimizing potential overfitting to the specific features of this dataset.

Finally, while SHAP-based explanations offer interpretability, the complexity of these explanations may still pose challenges for clinical use [47], as physicians may find it time-consuming to review the contributions of individual features for each prediction. Future research could focus on developing streamlined interpretability tools and user-friendly interfaces to facilitate robust and easier integration into clinical workflows [48]. Addressing these limitations in future studies will help improve the robustness, generalizability, and clinical usability of the proposed model, making it better suited for real-world diagnostic needs and enhancing its potential impact on patient care.

Potential integration into clinical workflows

The proposed diagnostic model aims to serve as a decision support tool for healthcare providers in diagnosing thyroid diseases, particularly in resource-limited or

Table 8 Performance fairness analysis of the proposed model for male and female patients

Male patients			
Target	Precision	Recall	F1
Negative	1.00	0.99	0.99
Sick	0.88	0.93	0.90
Average	**0.94**	**0.96**	**0.95**
Female patients			
Negative	1.00	1.00	1.00
Sick	0.94	0.89	0.92
Average	**0.97**	**0.95**	**0.96**

Table 9 Performance fairness analysis of the proposed model for young and adult age group patients

Young patients (age<=20)			
Target	Precision	Recall	F1
Negative	1.00	1.00	1.00
Sick	1.00	1.00	1.00
Average	**1.00**	**1.00**	**1.00**
Adult patients (age>20)			
Negative	1.00	0.99	0.99
Sick	0.99	1.00	1.00
Average	**0.99**	**0.99**	**0.99**

Table 10 Performance comparison of the proposed method with state-of-the-art studies

Ref.	Year	Proposed method	Performance accuracy
[25]	2023	Convolutional Neural Network	0.89
[18]	2023	Support Vector Machine	0.90
[28]	2022	Random forest	0.84
This Study	**2024**	**Proposed SNL**	**0.96**

high-volume settings. Its application could involve integration with electronic health record (EHR) systems, where the model processes patient data and flags possible thyroid abnormalities. This allows physicians to review predictions alongside patient histories and laboratory results. Integration could streamline preliminary diagnostic steps, help prioritize cases, and support early intervention strategies.

To ensure practical relevance, future work should include clinical validation trials in which the model's outputs are evaluated against physician diagnoses in a real-world setting. By embedding the model as a complementary tool in existing diagnostic workflows, we anticipate that it could expedite patient evaluations and help identify thyroid conditions early, particularly in asymptomatic cases where traditional screenings may be delayed. Physicians could also use this tool as part of a broader diagnostic framework, offering preliminary insights based on patient data, which are then further refined through clinical judgment and additional testing as needed. This collaborative approach between AI-based diagnostics and human expertise has the potential to enhance diagnostic accuracy and patient care quality in endocrinology clinics and primary care settings.

Conclusions and future work

This research proposes an effective artificial intelligence-based approach for the early diagnosis of thyroid illness, leveraging an open-access thyroid disease dataset with 3,772 patient observations. By uniquely combining the SMOTE-NC method with a fine-tuned LGBM technique, this study addresses class imbalance challenges and achieves a high diagnostic accuracy score of 0.96, outperforming current state-of-the-art methods. We further evaluated model performance by comparing it with four advanced machine learning techniques and two deep learning models, optimizing hyperparameters to enhance diagnostic accuracy. To support interpretability, we utilized the SHAP-based XAI mechanism, providing transparency into the model's decision-making process and facilitating clinician understanding of its predictions.

The potential applications of this model extend to clinical settings as a decision-support tool, with future goals of integration into EHR systems. This integration could assist healthcare providers by offering preliminary diagnostic insights and flagging potential cases for further evaluation, especially in high-volume or resource-limited environments. To fully realize its clinical potential, future work will focus on conducting real-world clinical validation trials, assessing the model's effectiveness in collaboration with physicians, and refining its utility within clinical workflows. Despite these promising outcomes, there are limitations that future research could address. The dataset's size and diversity, while substantial, could be further expanded to improve the model's robustness and generalizability across broader populations. Future studies may also explore incorporating additional biomarkers and features related to thyroid disease, as well as advanced methods like transfer learning to boost diagnostic performance. These enhancements will not only strengthen model accuracy but also support its applicability across diverse patient populations, ensuring greater reliability in clinical contexts. In addition, we plan to create a graphic user interface-based tool where medical specialists can input patient details, and the proposed framework will provide a real-time diagnosis of thyroid conditions in a clinical environment.

Acknowledgements
Not applicable.

Authors' contributions
AR conceptualization, data curation, writing manuscript. FE conceptualization, formal analysis, writing manuscript. ECM funding acquitistion, methodology, formal analysis. IDN investigation, project administration, visualization IA supervision, validation, writing & editing manuscript. All authors reviewed the manuscript.

Funding
This study is supported by the European University of Atlantics.

Data availability
The dataset, 'Thyroid Sickness Determination' used in this is publicly available at the following link:
https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 September 2024 Accepted: 22 November 2024
Published online: 29 November 2024

References

1. Economidou F, Douka E, Tzanela M, Nanas S, Kotanidou A. Thyroid function during critical illness. Hormones. 2011;10(2):117–24.
2. De Luca R, Davis PJ, Lin HY, Gionfra F, Percario ZA, Affabris E, et al. Thyroid hormones interaction with immune response, inflammation and non-thyroidal illness syndrome. Front Cell Dev Biol. 2021;8:614030.
3. Sinkó R, Mohácsik P, Kővári D, Penksza V, Wittmann G, Mácsai L, et al. Different hypothalamic mechanisms control decreased circulating thyroid hormone levels in infection and fasting-induced Non-Thyroidal Illness Syndrome in male Thyroid Hormone Action Indicator Mice. Thyroid. 2023;33(1):109–18.
4. Sipos JA, Ringel MD. Molecular testing in thyroid cancer diagnosis and management. Best Pract Res Clin Endocrinol Metab. 2023;37(1):101680.

5. Schneider SA, Tschaidse L, Reisch N. Thyroid disorders and movement disorders–a systematic review. Mov Disord Clin Pract. 2023;10(3):360–8.
6. Riis J, Kragholm K, Torp-Pedersen C, Andersen S. Association between thyroid function, nursing home admission and mortality in community-dwelling adults over 80 years. Arch Gerontol Geriatr. 2023;104:104806.
7. Purohit J, Barjatya R, Kataria SK. Evaluation of Hyperprolactinemia and Thyroid Disorder among Women with Dysfunctional Uterine Bleeding at Tertiary Care Hospital of western Rajasthan. Sch Int J Anat Physiol. 2023;6(5):61–3.
8. Zhang X, Lee VC, Rong J, Liu F, Kong H. Multi-channel convolutional neural network architectures for thyroid cancer detection. PLoS ONE. 2022;17(1):e0262128.
9. Fiorentino V, Pizzimenti C, Franchina M, Micali MG, Russotto F, Pepe L, et al. The minefield of indeterminate thyroid nodules: could artificial intelligence be a suitable diagnostic tool? Diagn Histopathology. USA: Elsevier; 2023.
10. Aversano L, Bernardi ML, Cimitile M, Maiellaro A, Pecori R. A systematic review on artificial intelligence techniques for detecting thyroid diseases. PeerJ Comput Sci. 2023;9:e1394.
11. Imans D, Abuhmed T, Alharbi M, El-Sappagh S. Explainable Multi-Layer Dynamic Ensemble Framework Optimized for Depression Detection and Severity Assessment. Diagnostics. 2024;14(21):2385.
12. Saleh H, El-Rashidy N, Abuhmed T, El-Sappagh SLSTM, deep learning model for Alzheimer's disease prediction based on cost-effective time series cognitive scores. In: 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES). IEEE; 2023. pp. 1–6.
13. Rahim N, El-Sappagh S, Rizk H, El-serafy OA, Abuhmed T. Information fusion-based Bayesian optimized heterogeneous deep ensemble model based on longitudinal neuroimaging data. Appl Soft Comput. 2024;162:111749.
14. Rani CP, Nagaraju T, Vardhan NSH, Teja PN, Charishma P. Machine Learning Model for Accurate Prediction of Thyroid Disease. In: 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2023. pp. 1–7. https://doi.org/10.1109/ACCAI58221.2023.10199375.
15. Dixit R, Tayal MA, Bedi S, Saxena S. Thyroid Disorder Classification using Machine Learning. In: 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), 2023. pp. 1–5. https://doi.org/10.1109/ICETET-SIP58143.2023.10151522.
16. Nandi-Munshi D, Taplin CE. Thyroid-related neurological disorders and complications in children. Pediatr Neurol. 2015;52(4):373–82.
17. Hossain MB, Shama A, Adhikary A, Raha AD, Uddin KA, Hossain MA, et al. An Explainable Artificial Intelligence Framework for the Predictive Analysis of Hypo and Hyper Thyroidism Using Machine Learning Algorithms. Hum-Centric Intell Syst. 2023;3:1–21.
18. Priya VV, Subashini R, Priya SH. Thyroid Disease Prediction using Random Forest Algorithm. In: 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), 2023. pp. 794–799. https://doi.org/10.1109/ICCMC56507.2023.10083592.
19. Islam SS, Haque MS, Miah MSU, Sarwar TB, Nugraha R. Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. PeerJ Comput Sci. 2022;8:e898.
20. AYINDE B. Thyroid Sickness Determination | Kaggle, 2022. https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination. Accessed 8 May 2023
21. Gök EC, Olgun MO. SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. Neural Comput & Applic. 2021;33(22):15693–707.
22. A S, A BA, E S. Balancing of an imbalanced dataset by applying SMOTE variants and predicting neonatal mortality using ensemble learning techniques. In: 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), 2022. pp. 1–6. https://doi.org/10.1109/ICITIIT54346.2022.9744204.
23. Wibowo W, Muhaimin A, Abdul-Rahman S. Predicting Internet Usage for Digital Finance Services: Multitarget Classification Using Vector Generalized Additive Model with SMOTE-NC. In: The International Conference on Data Science and Emerging Technologies. Springer; 2022. pp. 494–504.
24. Chen Jh, Zhang YQ, Zhu Tt, Zhang Q, Zhao Ax, Huang Y. Applying machine-learning models to differentiate benign and malignant thyroid nodules classified as C-TIRADS 4 based on 2D-ultrasound combined with five contrast-enhanced ultrasound key frames. Front Endocrinol. 2024;15:1299686.
25. Brindha V, Muthukumaravel A. Efficient Method for the prediction of Thyroid Disease Classification Using Support Vector Machine and Logistic Regression. In: Computational Intelligence for Clinical Diagnosis. Springer; 2023. pp. 37–45.
26. Jakkulla PK, Ganesh KM, Jayapal PK, Malla SJ, Chandanapalli SB, Sandhya E. Selection of Features Using Adaptive Tunicate Swarm Algorithm with Optimized Deep Learning Model for Thyroid Disease Classification. Ingenierie Systemes Inf. 2023;28(2):299.
27. Nayak C, Ajalkar D, Shinde JP, Barik SS. Machine Learning Thyroid Model for Prediction System. In: 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), 2023. pp. 602–607. https://doi.org/10.1109/DICCT56244.2023.10110065.
28. Alyas T, Hamid M, Alissa K, Faiz T, Tabassum N, Ahmad A. Empirical method for thyroid disease classification using a machine learning approach. BioMed Res Int. 2022;2022:932–80.
29. Rossi RA, Ahmed NK. The Network Data Repository with Interactive Graph Analytics and Visualization. In: AAAI, 2015. https://networkrepository.com. Accessed 6 Mar 2024.
30. Pal M, Parija S, Panda G. Enhanced Prediction of Thyroid Disease Using Machine Learning Method. In: 2022 IEEE VLSI Device Circuit and System (VLSI DCS), 2022. pp. 199–204. https://doi.org/10.1109/VLSIDCS53788.2022.9811472.
31. Junaid M, Ali S, Eid F, El-Sappagh S, Abuhmed T. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease. Comput Methods Prog Biomed. 2023;234:107495.
32. Bini F, Pica A, Azzimonti L, Giusti A, Ruinelli L, Marinozzi F, et al. Artificial intelligence in thyroid field–a comprehensive review. Cancers. 2021;13(19):4740.
33. Raza A, Munir K, Almutairi M. A novel deep learning approach for deepfake image detection. Appl Sci. 2022;12(19):9820.
34. Rehman A, Raza A, Alamri FS, Alghofaily B, Saba T. Transfer Learning-Based Smart Features Engineering for Osteoarthritis Diagnosis From Knee X-Ray Images. IEEE Access. 2023;11:71326–38. https://doi.org/10.1109/ACCESS.2023.3294542.
35. Qadri AM, Raza A, Munir K, Almutairi MS. Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning. IEEE Access. 2023;11:56214–24. https://doi.org/10.1109/ACCESS.2023.3281484.
36. Ishtiaq A, Munir K, Raza A, Samee NA, Jamjoom MM, Ullah Z. Product Helpfulness Detection With Novel Transformer Based BERT Embedding and Class Probability Features. IEEE Access. 2024;12:55905–17.
37. Khalid M, Raza A, Younas F, Rustam F, Villar MG, Ashraf I, et al. Novel Sentiment Majority Voting Classifier and Transfer Learning-based Feature Engineering for Sentiment Analysis of Deepfake Tweets. IEEE Access. 2024;12:67117–29.
38. Younas F, Raza A, Thalji N, Abualigah L, Zitar RA, Jia H. An efficient artificial intelligence approach for early detection of cross-site scripting attacks. Decis Anal J. 2024;11:100466.
39. Darawsheh SR, Al-Shaar AS, Haziemeh FA, Alshurideh MT. Classification Thyroid Disease Using Multinomial Logistic Regressions (LR). In: The Effect of Information Technology on Business and Marketing Intelligence Systems. Springer; 2023. pp. 645–659.
40. Raza A, Siddiqui HUR, Munir K, Almutairi M, Rustam F, Ashraf I. Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction. PLoS ONE. 2022;17(11):e0276525.
41. Chen Z, Ying TC, Chen J, Wang Y, Wu C, Su Z. Assessment of Renal Fibrosis in Patients With Chronic Kidney Disease Using Shear Wave Elastography and Clinical Features: A Random Forest Approach. Ultrasound Med Biol. 2023;49(7):1665–71.
42. Mohi Uddin KM, Biswas N, Rikta ST, Dey SK, Qazi A. XML-LightGBMDroid: A self-driven interactive mobile application utilizing explainable machine learning for breast cancer diagnosis. Eng Rep. 2023;11:e12666.
43. Merkelbach K, Schaper S, Diedrich C, Fritsch SJ, Schuppert A. Novel architecture for gated recurrent unit autoencoder trained on time series from electronic health records enables detection of ICU patient subgroups. Sci Rep. 2023;13(1):4053.
44. Wu X, Wang HY, Shi P, Sun R, Wang X, Luo Z, et al. Long short-term memory model-a deep learning approach for medical data with

irregularity in cancer predication with tumor markers. Comput Biol Med. 2022;144:105362.
45. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4793–813. https://doi.org/10.1109/TNNLS.2020.3027314.
46. Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470.
47. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): what we know and what is left to attain Trustworthy Artificial Intelligence. Inf Fusion. 2023;99:101805.
48. Javed H, El-Sappagh S, Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artif Intell Rev. 2024;58(1):12.

Publisher's Note