FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 50 (2025) No. 4
ISSN 0867-6356
DOI: 10.2478/fcds-2025-0019 e-ISSN 2300-3405

A Transparent AI-Driven Multiclass Decision Support System for Thyroid Risk Prediction Using Machine Learning and Deep Learning Approaches

Siouar Ouartani *, Nora Taleb †

Abstract. Early and accurate diagnosis of thyroid disorders is essential due to their prevalence and health impact. To enhance interpretability in clinical settings, we propose a comprehensive workflow for transparent thyroid disease predic- tion using a multiclass classification problem with five diagnostic categories. A dataset of 9172 samples with 31 features was used to train various machine and deep learning models. A dual-layered framework combining Feature Selection (ETC, MI, RFE) and Explainable AI (SHAP, LIME) improved performance and transparency. Gradient Boosting achieved the highest accuracy (0.97). SHAP explained global feature influence, while LIME clarified individual predictions. Our approach supports interpretable, reliable AI-based diagnostic tools for thyroid disorder classification.

Keywords: Thyroid Disease Diagnosis, Medical Decision Support System (MDSS), Machine Learning, Deep Learning, Explainable AI (XAI), Feature Selection, Model Interpretability, LIME, SHAP.

1. Introduction

The thyroid gland, located in the anterior region of the neck, is a small, butterfly-shaped endocrine organ essential for regulating the body's metabolism through the secretion of thyroid hormones, primarily Free Triiodothyronine (FT3) and Free Thyroxine (FT4) [17]. Thyroid disorders are generally classified into two categories: hyperthyroidism and hypothyroidism. Hyperthyroidism is characterized by excessive production of thyroid hormones, leading to accelerated metabolic activity and increased energy consumption. In contrast, hypothyroidism is marked by insufficient

*Complex Systems Engineering Laboratory (LISCO), Computer Science Department, Badji Mokhtar University, Annaba, Algeria, siouar.ouartani@univ-annaba.dz
†Complex Systems Engineering Laboratory (LISCO), Computer Science Department, Badji Mokhtar University, Annaba, Algeria, talebnr@hotmail.fr

hormone production, resulting in reduced metabolic activity and energy utilization. Both conditions present significant health risks [17]. Despite advancements in diagnostic methods, the differentiation between these disorders can be challenging due to subtle variations in thyroid hormone levels. These minor discrepancies can potentially lead to misdiagnosis, even among experienced clinicians, which may result in inappropriate treatment and further complications.[5] To mitigate these risks, the emergence of Medical Decision Support Systems (MDSS) has revolutionized the field of thyroid disease prediction. These systems, powered by advancements in machine learning (ML), have provided clinicians with enhanced tools for analyzing complex datasets, leading to more accurate and reliable diagnostic outcomes. By leveraging ML algorithms, MDSS can identify patterns and correlations within vast amounts of patient data, thereby improving the predictive accuracy of thyroid disease diagnoses and supporting more informed decision-making processes in clinical settings [1]. Despite these advancements, traditional diagnostic approaches in thyroid disease prediction still face challenges, particularly in terms of transparency. Many conventional models operate as "black boxes", offering little insight into how predictions are made. This lack of interpretability can be a significant limitation in clinical practice, where understanding the rationale behind a diagnosis is often as important as the diagnosis itself. Clinicians need to trust the decisions made by these models, especially when they are used to inform critical healthcare decisions [9]. To address this issue, we propose an interpretable, multi-class classification framework for thyroid disease prediction that integrates machine learning and deep learning models with explainable artificial intelligence (XAI) techniques. The specific objectives of this work are to (1) apply and compare feature selection techniques to identify the most relevant biomarkers, (2) evaluate a wide range of ML and DL models on a large, real world thyroid dataset before and after feature selection, and (3) improve clinical interpretability of predictions using post-hoc XAI methods.

Following a rigorous preprocessing phase, various machine learning models including Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost (ADA), Gradient Boosting Machine (GBM), and Multi-Layer Perceptron (MLP) as well as deep learning models such as Deep Neural Networks (DNN), Long Short-Term Memory networks (LSTM), Convolutional Neural Networks (CNN), and hybrid CNN-LSTM models were evaluated to identify the most effective approach for this multi-class classification task. Gradient Boosting was selected for in-depth analysis due to its superior performance, achieving an accuracy rate of 97%. However, due to its complexity, it is critical to interpret and understand the features influencing its predictions.To enhance model interpretability, we adopted a dual XAI strategy. LIME (Local Interpretable Model-Agnostic Explanations) was incorporated to offer local interpretability by explaining why specific predictions were made for individual patients[11][8]. and , SHAP (SHapley Additive Explanations) was used to provide global interpretability, allowing us to understand how each feature contributes to model predictions across the entire dataset [8]. This distinction is crucial in medical settings, where clinicians need both an overview of what drives the model and a clear rationale for specific decisions [2].

By making the model more understandable and trustworthy [5], our method suc-

cessfully identified five distinct thyroid pathologies: Hashimoto's thyroiditis (primary hypothyroidism), increased binding protein, autoimmune thyroiditis (compensated hypothyroidism), non-thyroidal disease syndrome (NTIS), and Graves' disease (hyperthyroidism), along with the specific factors influencing each prediction.

The remainder of this paper is organized as follows: Section 2 reviews related works and recent developments in thyroid disease prediction and model interpretability. Section 3 introduces the core components of our methodology, including Medical Decision Support Systems (MDSS), explainable artificial intelligence (XAI), the black-box models used, and the interpretability techniques employed. Section 4 details the model construction process, covering dataset acquisition, preprocessing, feature selection, and performance evaluation before and after applying feature selection. Section 5 presents the interpretability analysis of the Gradient Boosting model using LIME and SHAP, and compares their strengths and limitations. Section 6 discusses our results in the context of existing studies, highlighting similarities and improvements. Finally, Section 7 concludes the paper with key findings and outlines potential directions for future research.

2. Related works

A variety of work has been done on the interpretation of machine learning (ML) models. Researchers have increasingly focused on explainable artificial intelligence (XAI) techniques to make ML models more transparent and understandable. These efforts aim to elucidate which features influence model predictions rather than merely improving accuracy, addressing the critical need for interpretability in clinical applications.

Hu and Sokolova[11] employed both post-hoc and ante-hoc Machine Learning explanation methods to analyze factors influencing behavior during the Covid-19 pandemic. They used six ML models, finding that Random Forest and Gradient Boosting achieved the highest accuracies 68.08% and 68.19% respectively . Using LIME, they explained model predictions and identified significant factors such as cannabis use, alcohol consumption, and Covid-19 diagnoses. Gini Importance corroborated these findings. Their study highlighted differences in functionality and usability between LIME and Gini Importance and suggested that larger samples and multiple labels could refine future analyses of mental health and behavioral factors.

In their study, Guler and his colleagues [8] conducted a comprehensive study on the early diagnosis and prediction of diabetes using various machine learning (ML) models and explainable artificial intelligence (XAI) techniques. Their research leveraged a diabetes dataset to evaluate the performance of ML models such as K-NN, SVM, Naive Bayes, CNN, Decision Tree, Random Forest, and XGBoost, with XGBoost achieving the highest accuracy at 98.91%. The study emphasized the importance of data preprocessing and visualization in analyzing diabetes-related features. To enhance model interpretability, they applied XAI methods, including SHAP and LIME, to elucidate the features most impactful on model outputs. Expert doctor commentary supported their analyses, highlighting the real-world potential of these models

in improving early diabetes diagnosis and patient outcomes.

Ghosh and Khandoker [7]investigate the critical issue of Chronic Kidney Disease (CKD), highlighting the necessity for early detection and treatment to mitigate its global impact. Utilizing explainable artificial intelligence (XAI) techniques, they analyze clinical data from 491 patients, employing five machine learning (ML) methods such as LR, RF, DT, and NB to develop predictive models. Among these, XGBoost emerges as the most effective, achieving an AUC of 0.9689 and an accuracy of 93.29%. Feature importance analysis reveals key variables influencing CKD prediction, while SHAP and LIME algorithms enhance model interpretability, providing valuable insights for clinicians in understanding individual predictions. This study underscores an interpretable ML approach for early CKD prediction, facilitating informed decision-making in clinical practice.

Junkang and all[12]. In their exploration, propose an innovative interpretable artificial intelligence (XAI) method to address the challenge of understanding predictions made by deep learning models, which often have complex parameters. Their approach, termed specific-input local interpretable model-agnostic explanations (LIME), interprets deep learning models applied to tabular data by utilizing feature importance and partial dependency plots (PDPs). This method enhances the stability of LIME interpretations and balances global and local interpretations by focusing on the features that models concentrate on and how these features influence predictions. The specific-input LIME method offers detailed explanations, making it particularly valuable in practical applications such as healthcare, where it can elucidate how machine learning models predict diseases and analyze the impact of individual features. Despite its effectiveness, the method has limitations related to sensitivity to perturbation size and direction, and high computational costs. Future work will focus on optimizing parameters to reduce these costs and improve robustness.

Mulwa and all [14] improve the adoption of machine learning (ML) in clinical settings for human gait analysis by addressing the instability of the local interpretable model-agnostic explanation (LIME) method. They propose GMM-LIME, which combines LIME with Gaussian mixture model (GMM) sampling to enhance stability. Their study shows that supervised ML models, including neural networks on the GaitRec dataset and Random Forest on the HugaDB datasets, achieved high accuracies of 95% and 99%, respectively. GMM-LIME provided more accurate, reliable, and consistently stable explanations than the original LIME. This advancement in explainable artificial intelligence (XAI) enhances the transparency of ML models, fostering greater trust and adoption in clinical settings for gait disorder diagnosis and care.

Settouti and Saïdi[18] propose a two-step framework using explainable artificial intelligence (XAI) to improve transparency in selecting chemotherapy treatments for multiple myeloma. They apply permutation feature importance, Kernel SHAP, and LIME to explain decisions from machine learning models such as Random Forest, SVM, XGBoost, and MLP. Their study reveals that Random Forest achieves the highest accuracy (98.63%), with XGBoost, MLP, and SVM following. Kernel SHAP is noted for its superior transparency in feature importance. The study highlights the importance of clear interpretation in XAI and suggests future enhancements for

explainable machine learning in therapeutic contexts.

Hajar Hakkoum and all.In their research[10], explore interpretability techniques in the context of breast cancer diagnosis using two feed-forward Artificial Neural Networks (ANNs): Multilayer perceptrons (MLPs) and Radial Basis Function Networks (RBFNs). They investigate three interpretation methods—Feature Importance, Partial Dependence Plot, and LIME—applied to these models trained on the Wisconsin Original dataset. Their study demonstrates that local explanations provided by LIME align with global interpretations from other techniques, indicating the potential for combined global and local interpretability to assess model trustworthiness. Despite RBFN's superior accuracy, MLP's interpretations better align with an interpretable decision tree classifier, suggesting the utility of interpretability in model selection. The investigation highlights the value of interpretability in understanding model behavior, enhancing trustworthiness, and aiding domain experts in discovering hidden patterns. Future work aims to explore additional interpretation techniques and improve MLP parameter tuning to further enhance model performance and interpretability. The study underscores the importance of ML explainability in bridging the gap between accuracy and interpretability, ensuring legal compliance and facilitating informed decision-making in medical applications.

Other new studies have explored XAI with different ML models for medical use For exampl, Aldughayfiq and all[2]. address the challenge of transparency in deep learning models for detecting retinoblastoma, a critical childhood eye cancer. They employ LIME and SHAP, two explainable AI techniques, to elucidate the decision-making process of a deep learning model based on InceptionV3 architecture trained on retinoblastoma and non-retinoblastoma fundus images. Using a dataset of 400 labeled images split into training, validation, and test sets, the model achieves 97% accuracy on the test set through transfer learning. LIME and SHAP successfully highlight the significant regions and features influencing the model's predictions, enhancing understanding and trust in its diagnostic capabilities. This study underscores the potential of combining deep learning with explainable AI to advance retinoblastoma diagnosis and treatment outcomes.

Islam and all[16]. proposed DCNN-GRU, a novel deep learning approach for the detection and classification of lung abnormalities, integrating Explainable AI (XAI) techniques such as LIME, SHAP, and Grad-CAM to improve model interpretability. The model combines a Deep Convolutional Neural Network (DCNN) for feature extraction with a Gated Recurrent Unit (GRU) to capture sequential dependencies in lung images. Evaluated on COVID-19 and lung cancer datasets, the model achieved 99.30% and 98.97% accuracy, respectively, outperforming existing approaches while reducing training time. The integration of XAI techniques provided clinically relevant insights, enabling better understanding and validation of the model's predictions. This study demonstrates the effectiveness of combining deep learning and XAI for improving diagnostic accuracy and decision transparency in medical imaging.

In their study, Arjaria and his colleagues[20], presented a thyroid disease diagnosis system that uses SHAP for model interpretability. The system employed a logistic regression model trained on the UCI KDD thyroid dataset and applied SHAP to explain the model's decisions both locally and globally. The study highlighted the

importance of features and the conditional dependencies among them, using methods like decision plots and force plots for local explanations, and summary plots for global insights. SHAP's model-agnostic approach makes it applicable to various machine learning models, improving trust and transparency in medical decision support systems.

Al Amin and colleagues[4], propose an innovative XAI framework aimed at enhancing transparency and trust in AI-driven medical applications, particularly within the domain of Artificial Intelligence of Medical Things (AIoMT). The framework integrates established explainability techniques such as LIME, SHAP, and Grad-CAM to provide both local and global interpretability for complex models employed in medical scenarios. The effectiveness of this framework is demonstrated through a brain tumor detection case study, illustrating how it explains the predictions of deep learning models in a clinically relevant manner. By combining these techniques, the authors address the pressing need for trustworthy and interpretable AI systems in healthcare, where decisions can have high-stakes consequences. The proposed framework not only improves decision-making but also enhances the acceptability of AI in healthcare by offering both visual and quantitative insights into the model's reasoning.

In their review, Qiyang Sun and all[21], provide a comprehensive analysis of recent advancements in explainable artificial intelligence (XAI) within the medical field. The authors emphasize the critical need for transparency in AI-driven medical decision-making, addressing the inherent challenges posed by the "black-box" nature of many AI models. They categorize and synthesize XAI techniques across visual, audio, and multimodal applications, aiming to guide future research and support healthcare professionals. The review highlights the importance of integrating XAI methods to enhance trust, accountability, and ethical standards in medical AI applications.

In [3] Johannes Allgaier and all. investigate the use of explainable artificial intelligence (XAI) methods in 450 medical machine learning (ML) use cases through a systematic literature review. They find that many ML applications lack explainability, though when used, model-agnostic methods like SHAP and Grad-CAM dominate for tabular and image data. The study highlights a trend toward better standardization in ML pipeline reporting but notes stagnation in data and code sharing. The authors emphasize the growing need for experts who bridge the gap between informatics and medicine to ensure the effective deployment of ML in healthcare.

Together, these studies highlight the growing importance of XAI techniques in improving the interpretability of diverse black box models for clinical applications, ensuring that AI-driven decisions are both accurate and transparent.

3. Materials and Methods

This study aims to create an effective prediction model for the early diagnosis of thyroid disease. In this section, we provide an overview of the materials and methods used in our study. We describe the dataset collection process and the machine learning models that were constructed and evaluated in our experiments, as well as the interpretability techniques applied to the best-performing models, which are explained in

detail later. The experimental workflow is visually depicted in Figure 1.

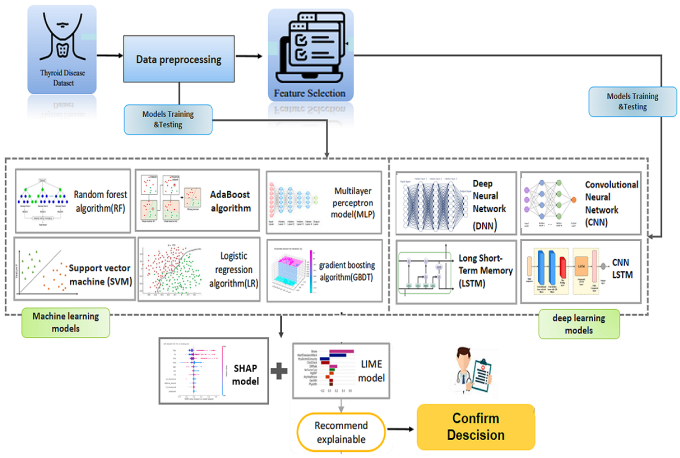

Figure 1: Process of the proposed methodology.

3.1. Medical Decision Support Systems and Explainable Artificial Intelligence (XAI)

Medical Decision Support Systems (MDSS) are essential tools in modern healthcare, providing clinicians with critical insights to enhance patient care and outcomes. MDSS are integral across all medical activities, including prevention, screening, diagnosis, and treatment, addressing a wide range of medical specialties, from chronic diseases to acute conditions [15]. These systems fall into two principal categories: knowledge-based decision support systems, which use predefined logic and expert knowledge, and data-based decision support systems, which commonly utilize machine learning algorithms to achieve high accuracy [5][15]. While effective, knowledge-based systems can be inflexible and limited by the comprehensiveness of the rules they are programmed with [15]. Furthermore, building and updating their knowledge bases can be a daunting, time-consuming, and costly task, significantly hindering their widespread adoption. On the other hand, data-based models face significant challenges, such as the need for large volumes of data for improved accuracy and their lack of interpretability, often being regarded as "black boxes"[15]. This is where Explainable Artificial Intelligence (XAI) comes into play. Figure 2 shows the general model of MDSS.

Explainable AI (XAI) methods are generally divided into intrinsic interpretability and post-hoc explanations [23]. Intrinsic methods ensure that models are inherently transparent by design, eliminating the need for additional interpretation. In contrast, post-hoc techniques provide explanations after a model has been trained, using feature attribution, textual descriptions, and visualizations to interpret complex black-box

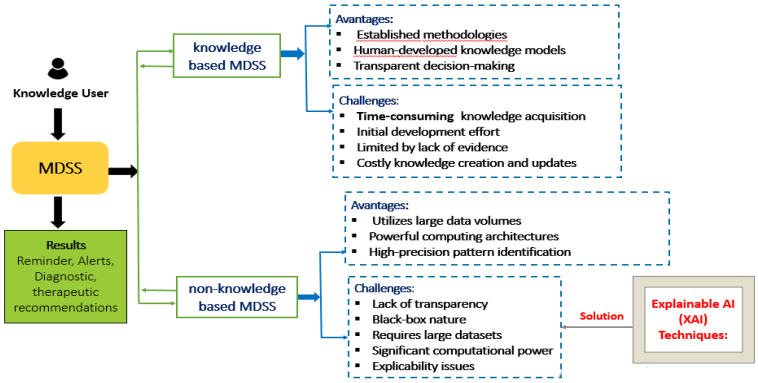

Figure 2: General model of MDSS.

models without exposing their inner workings [23].

Post-hoc methods can be further categorized into model-agnostic and model-specific approaches. Model-agnostic methods work independently of the underlying model architecture and analyze input-output relationships to generate explanations, while model-specific methods leverage internal model structures for interpretation. Additionally, explainability techniques can be classified as global, which provide an overview of how a model behaves across all predictions, or local, which focus on explaining individual predictions [3].

Among global model-agnostic methods, Partial Dependence Plots (PDPs), Permutation Feature Importance, and Leave One Covariate Out (LOCO) are widely used to assess overall feature influences on model behavior. In contrast, local model-agnostic techniques focus on explaining specific predictions, with popular examples including Individual Conditional Expectation (ICE), Locally Interpretable Model-Agnostic Explanations (LIME), Anchors (rule-based explanations), and Shapley Additive Explanations (SHAP) [3].

For global model-specific approaches, methods such as Mean Decrease Impurity (MDI), Testing Concept Activation Vectors (TCAV), Soft Decision Trees, SHAP, and TabNet leverage the internal structure of models to provide insights into overall decision-making patterns. Meanwhile, local model-specific methods, particularly gradient-based neural network explanations, focus on explaining individual predictions using gradient information within deep learning models. These diverse techniques enable different levels of interpretability [3]. This study specifically employs LIME and SHAP to enhance our model's interpretability. Overall, integrating XAI into MDSS not only enhances their transparency but also ensures more informed and confident medical decisions, ultimately improving patient outcomes. This transparency is critical for the clinical acceptance of AI, enabling healthcare providers to trust and effectively utilize these advanced systems [7].

3.2. Black box models adopted for the study

This study employs several machine learning models for thyroid disease detection, selected based on their ability to handle structured medical data effectively. The models include Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost (ADA), Gradient Boosting Machine (GBM), and Multi-Layer Perceptron (MLP). Each model was chosen to capture different patterns in the data, balancing interpretability, handling of non-linearity, and ensemble learning strength. The models were fine-tuned using a grid search approach with GridSearchCV to optimize their performance. This method evaluates various combinations of hyperparameters, such as solvers and regularization strengths for Logistic Regression and SVM, and estimators and tree depth for ensemble models like Random Forest and Gradient Boosting. The process uses cross-validation to select the best parameters, ensuring optimal model performance, measured by the weighted F1 score. Detailed hyperparameter settings are provided in Table 1.

- **Logistic Regression (LR)**: Selected for its simplicity and interpretability, LR serves as a benchmark model for binary classification. It is particularly useful for understanding the linear relationships between features and the target variable. Hyperparameters like the solver and the regularization parameter (C) were tuned to enhance performance[3].
- **Support Vector Machine (SVM)**: Chosen for its robustness in handling high-dimensional spaces and its effectiveness in classification tasks with clear decision boundaries. SVM was tuned with different kernel types (linear, polynomial, sigmoid) and regularization parameters (C) to optimize performance[19].
- **Random Forest (RF)**: An ensemble learning method that enhances predictive accuracy by aggregating multiple decision trees. RF is known for its ability to handle feature importance and manage overfitting effectively. Key hyperparameters such as the number of trees (n-estimators) and maximum depth (max-depth) were adjusted to balance performance and generalization[19].
- **Gradient Boosting Machine (GBM)**: Selected for its strength in handling complex patterns through sequential tree-building. GBM corrects the errors of its predecessors and often achieves superior predictive performance. Important hyperparameters include the number of trees (n-estimators), learning rate, and tree depth[19].
- **AdaBoost (ADA)**: This boosting method was included to evaluate its ability to improve weak classifiers iteratively. While it is generally more sensitive to noise than GBM, it was fine-tuned with the number of weak learners (n-estimators) and learning rate to maximize performance[19].
- **Multi-layer Perceptron (MLP)**: An artificial neural network that captures complex non-linear relationships. It was optimized using different network architectures, varying the number of layers, neurons per layer, and activation functions to enhance classification performance[7].

Table 1: Hyperparameters of ML Models.

Model	Hyperparameters
Logistic Regression	Solver: {liblinear, saga}; C: {1.0, 5.0, 8.0}
SVM	Kernel: {linear, poly}; C: {1.0, 5.0, 8.0}
Random Forest	n_estimators: {10, 200}; max_depth: {2, 20, 50}
Gradient Boosting	n_estimators: {10}; max_depth: {2}; learning_rate: {0.1, 0.5, 0.9}
AdaBoost	n_estimators: {10, 200}; learning_rate: {0.1, 0.5, 0.9}
MLP	hidden_layer_sizes: (100,); activation: relu; alpha: {0.0001, 0.001, 0.01}

In addition to traditional machine learning models, the study also evaluates the performance of deep learning models on the same dataset, employing various feature selection techniques (detailed in Section X). The deep learning models used include a Deep Neural Network (DNN), Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and a hybrid CNN-LSTM architecture [5]. These models were chosen based on their ability to capture different patterns in the data:

- **Deep Neural Network (DNN):** The DNN was selected as a baseline deep learning model to assess the benefits of deeper architectures compared to traditional machine learning approaches.
- **Long Short-Term Memory (LSTM):** The LSTM was chosen due to its ability to model sequential dependencies, which is relevant for time-series patterns or data with inherent temporal relationships [5].
- **Convolutional Neural Network(CNN):** The CNN was included because of its effectiveness in feature extraction, particularly for structured spatial patterns in input data [5].
- **CNN-LSTM:** The CNN-LSTM was tested as a hybrid model that combines CNN's feature extraction strengths with LSTM's sequential modeling ability, potentially enhancing predictive performance by leveraging both techniques [5].

Each model was designed with different configurations, varying the number of layers, positions of dropout layers, number of neurons, and activation functions. The hyperparameters were manually selected based on prior knowledge, experimentation, and empirical performance. Training was conducted with the categorical-crossentropy loss function and the 'Adam' optimizer, using a batch size of 16 and over 100 epochs to ensure robust learning. As shown in Table 2, the LSTM model used 128 units with a 0.5 dropout rate, while the CNN model included a Conv1D layer with 128 filters, max pooling, and dropout layers. The CNN-LSTM model combined a Conv1D layer with 128 filters, max pooling, and 100 LSTM units, and the DNN model consisted of two dense layers with 128 and 64 units, each followed by dropout layers. This manual tuning approach focused on adjusting layer configurations, dropout rates, and filter sizes to determine the best-performing model based on empirical evaluations.

Table 2: Hyperparameters of Deep Learning Models.

Model	Hyperparameters
LSTM	Units: 128; Dropout: 0.5
CNN	Filters: 128; Kernel Size: 5; Pool Size: 5; Dropout: 0.5
CNN-LSTM	Filters: 128; Kernel Size: 5; Pool Size: 5; LSTM Units: 100
DNN	Hidden Layers: [128, 64]; Dropout: 0.5

3.3. Interpretable Model Based on LIME and SHAP

LIME and SHAP are post-hoc interpretability techniques that enhance model transparency by providing localized insights into decision-making. LIME serves as a valuable tool in enhancing the interpretability of machine learning models by providing insights into the decision-making process at a local level, making complex models more understandable to users [12][14]. When applied to multiclass prediction tasks, such as our thyroid disease dataset, LIME explains predictions across multiple classes by dissecting the model's rationale and highlighting the features that significantly influence each prediction. This granular approach provides users with a clearer understanding of how the model operates across various class outcomes. SHAP, on the other hand, provides a more theoretically grounded explanation by distributing feature contributions based on cooperative game theory. It assigns each feature a Shapley value, representing its marginal contribution to the prediction relative to different feature combinations [21] . Unlike LIME, SHAP ensures consistency and additivity, meaning that features contributing more to a prediction will always have higher importance scores. This makes SHAP particularly valuable for both local and global interpretability, as it can explain individual predictions while also summarizing overall feature importance across the entire dataset. SHAP's reliance on game-theoretic principles makes it computationally expensive [21][3], but it provides robust, stable explanations compared to LIME.

By integrating both LIME and SHAP, we leverage their complementary strengths, LIME offers an intuitive, instance-level understanding of predictions, while SHAP ensures theoretically sound and consistent feature attributions.

4. Models Construction

4.1. Dataset Acquisition and Preprocessing

We acquired the thyroid disease dataset from the Kaggle platform [13]. This dataset includes diverse thyroid-related disease records categorized into multiple target classes. It contains 9172 samples, each characterized by 31 features encompassing Boolean, float, integer, and string data types, as detailed in Table 3. The initial dataset exhibited a significant class imbalance, with some target classes severely underrepresented.

Table 3: Overview of Dataset Features

N°	Attribute	Description	Data Type
1	Age	Patient's age	int
2	Sex	Gender the patient identifies as	str
3	On-thyroxine	Indicates if the patient is receiving thyroxine treatment	bool
4	Query on thyroxine	Indicates if the patient is queried for thyroxine treatment	bool
5	On antithyroid meds	Indicates if the patient is taking antithyroid medications	bool
6	Sick	Indicates if the patient has a sickness	bool
7	Pregnant	Indicates if the patient is pregnant	bool
8	Thyroid-surgery	Indicates if the patient has had thyroid surgery	bool
9	I131-treatment	Indicates if the patient is undergoing I131 therapy	bool
10	Query-hypothyroid	Indicates if the patient suspects hypothyroidism	bool
11	Query-hyperthyroid	Indicates if the patient suspects hyperthyroidism	bool
12	Lithium	Indicates if the patient is using lithium	bool
13	Goitre	Indicates if the patient has a goitre condition	bool
14	Tumor	Indicates if the patient has a tumor	bool
15	Hypopituitary	Indicates if the patient has a condition affecting the pituitary gland	bool
16	Psych	Indicates if the patient has a psychological condition	bool
17	TSH-measured	Indicates if TSH was tested in the bloodwork	bool
18	TSH	TSH concentration from blood test results	float
19	T3-measured	Indicates if T3 was tested in the bloodwork	bool
20	T3	T3 concentration from blood test results	float
21	TT4-measured	Indicates if TT4 was tested in the bloodwork	bool
22	TT4	TT4 concentration from blood test results	float
23	T4U-measured	Indicates if T4U was tested in the bloodwork	bool
24	T4U	T4U concentration from blood test results	float
25	FTI-measured	Indicates if FTI was tested in the bloodwork	bool
26	FTI	FTI concentration from blood test results	float
27	TBG-measured	Indicates if TBG was tested in the bloodwork	bool
28	TBG	TBG concentration from blood test results	float
29	Referral-source	Source of the patient's referral	str
30	Target	Patient's diagnosis related to hyperthyroidism	str
31	Patient-id	Unique identifier for each patient	str

To mitigate this issue, we filtered the dataset to retain only those target classes with more than 200 samples. This selection process resulted in five target classes, which represent different thyroid health conditions and their corresponding diagnostic categories [15], as illustrated in Table 4.

Table 4: Description of Target Classes

Condition	**Index**	**Specific Class**	**Distribution**
Hyperthyroid	(A)	Hyperthyroid	147
	(B)	T3 Toxicity	21
	(C)	Toxic Goiter	6
	(D)	Secondary Toxic	8
Hypothyroid	(E)	Hypothyroid	1
	(F)	Primary Hypothyroid	233
	(G)	Compensated Hypothyroid	359
	(H)	Secondary Hypothyroid	8
Binding Protein	(I)	Increased Binding Protein	346
	(J)	Decreased Binding Protein	30
General Health	(K)	Concurrent Non-Thyroidal Illness	436
Replacement Therapy	(L)	Underreplaced	111
	(M)	Replacement Therapy Consistent	115
	(N)	Overreplaced	11
Antithyroid Treatment	(O)	Antithyroid Medication	14
	(P)	I131 Therapy	5
	(Q)	Surgical Treatment	14
Miscellaneous	(R)	Discordant Assay Outcomes	196
	(S)	Elevated TBG	85
	(T)	Elevated Thyroid Hormones	0
No Condition	-	Normal	6771

Subsequently, we performed data balancing to address the skewed distribution of class samples. Initially, the dataset contained 9173 patient records, with 6771 records denoting normal thyroid function. The remaining records included 233 cases of primary hypothyroidism, 359 cases of compensated hypothyroidism, 346 cases with elevated binding proteins, and 456 cases of concurrent non-thyroidal illness. Given the predominance of the normal class, we implemented random undersampling to select 400 samples from the normal class[15], thereby ensuring a more balanced dataset for model training. Data preprocessing steps included handling missing values, standardizing numerical features, encoding categorical variables, and normalizing the data to ensure uniform feature scaling. These steps were crucial to improving model performance and training efficiency. We applied techniques such as label encoding for categorical features and StandardScaler for numerical features, ensuring that the data fed into the machine learning models was clean and well-prepared for the subsequent analysis.

4.2. Selection of Features and Dataset Partitioning

In the domain of machine learning, feature selection (FS) plays a pivotal role in building efficient and interpretable models. Within our dataset, which originally contained 30 attributes, not all variables contributed equally to model performance. To refine the dataset and retain only the most impactful features, we employed three distinct FS techniques: Extra Trees Classifier, Mutual Information, and Recursive Feature Elimination.

Extra Trees Classifier (ETC): ETC is an ensemble-based embedded technique that evaluates feature importance during the model training process. It constructs multiple decision trees and measures the contribution of each feature to decision making. In this study, ETC was configured with n estimators = 200 and max depth = 20. Features with an importance score exceeding 0.015 were selected, reducing the dataset to 11 key features [5]. Figure 3 illustrates the feature importance scores obtained using ETC. This process not only reduces data dimensionality but also prioritizes the most influential attributes, thus enhancing model efficiency [12].

Mutual Information (MI): MI (Filter Method) is a statistical FS technique that quantifies the dependency between each feature and the target variable. Unlike model-dependent approaches, MI assesses how much knowing the value of a feature reduces uncertainty about the target. In this study, MI was applied using SelectKBest with mutual info classif as the scoring function. The number of selected features was set to k = 10, meaning that the 10 most informative features were retained for model training [22].

Recursive Feature Elimination (RFE): RFE is a wrapper-based iterative approach that eliminates the least significant features in successive rounds. It starts with all features and ranks them based on their importance in a given predictive model. In our study, Logistic Regression was used as the base estimator. The RFE process was configured to select the top 10 most predictive features, iteratively eliminating the least relevant ones until only 10 key attributes remained [22].

By integrating these feature selection techniques, we refined our dataset while preserving crucial information. The final dataset comprised 1774 samples, with selected feature subsets varying based on the applied FS method.Figure 4 illustrates the number of selected features for each feature selection method.

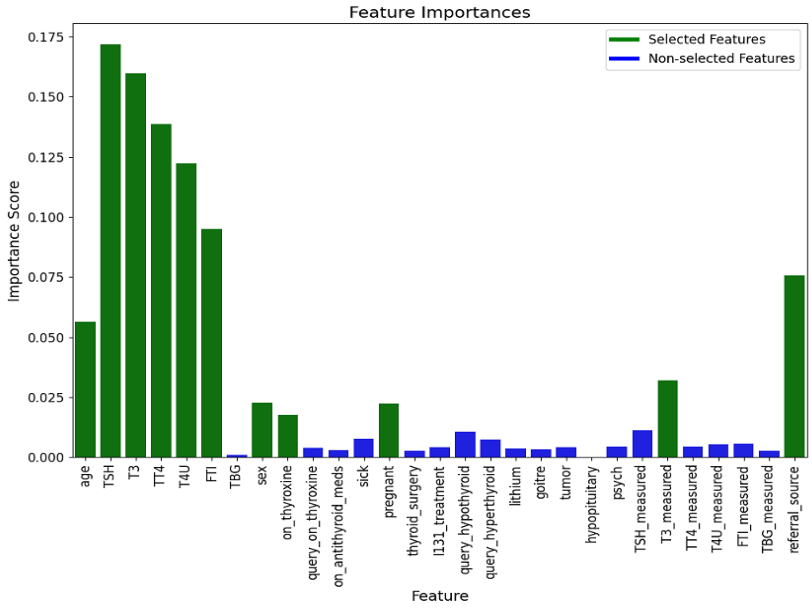

Figure 3: Feature Importance using ETC.

The data was subsequently divided into training and testing sets using an 80:20 split ratio. Specifically, 80% of the dataset was allocated for training our models, while the remaining 20% was reserved for evaluating the performance of our architecture.

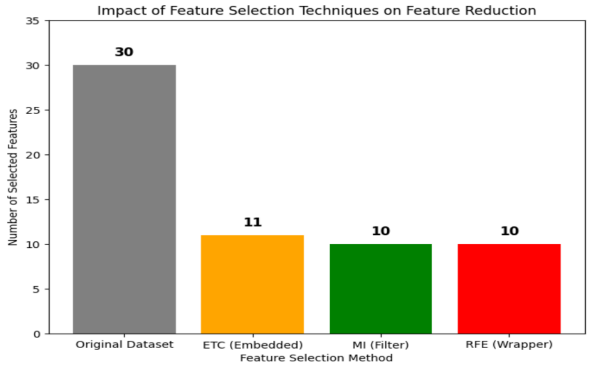

Figure 4: Impact of Feature Selection Techniques on the Number of Selected Features.

4.3. Performance of Black box Model Using Original Feature Set

4.3.1. Machine Learning Models

After splitting the data, we employed several machine learning models, including (LR), (SVM),(RF), (ADA), (GBM), and (MLP). The models were trained first using the original feature set, and subsequently with the important features identified through feature selection techniques, before being evaluated using 20% of the data reserved for testing. The performance of these models was assessed based on multiple metrics, including training accuracy, test accuracy, precision, recall, and F1 score. LR exhibited a high training accuracy of 93.02% and a test accuracy of 92.11%, indicating a well-generalized model with balanced precision and recall. The SVM model slightly outperformed Logistic Regression, achieving a training accuracy of 95.49% and a test accuracy of 93.24%, with improved precision and recall. Both Random Forest and Gradient Boosting achieved perfect training accuracy, which suggests potential overfitting. Despite this, they demonstrated strong generalization with a test accuracy of 97.75% and high precision and recall, making them robust candidates for this classification task. In contrast, the AdaBoost model showed lower performance, with a training accuracy of 76.39% and a test accuracy of 75.21%, along with imbalanced precision and recall, indicating it may not be the optimal choice for this dataset. The MLP model performed exceptionally well, with a training accuracy of 99.51% and a test accuracy of 94.08%, and maintained balanced precision and recall metrics.In summary, Random Forest and Gradient Boosting models demonstrate superior performance compared to other models, making them suitable choices for predicting outcomes using the original feature set.Table 5 shows the results of machine learning models and the Figure 4 illustrate the Confusion Matrices for Different Classifiers using the original feature set.

Table 5: ML Models Training Results

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
LR	0.9302	0.9211	0.9238	0.9211	0.9215
SVM	0.9549	0.9324	0.9336	0.9324	0.9325
RF	1.0000	0.9775	0.9793	0.9775	0.9774
GB	1.0000	0.9775	0.9793	0.9775	0.9774
AdaBoost	0.7639	0.7521	0.8732	0.7521	0.7213
MLP	0.9951	0.9408	0.9423	0.9408	0.9408

4.3.2. Deep Learning Models

The performance of deep learning models was also evaluated using the original feature set. The combined CNN-LSTM model and the Deep Neural Network (DNN) both achieved the highest accuracy of 94.37%, closely followed by the Convolutional Neural Network (CNN) with the same accuracy. Meanwhile, the Long Short-Term Memory

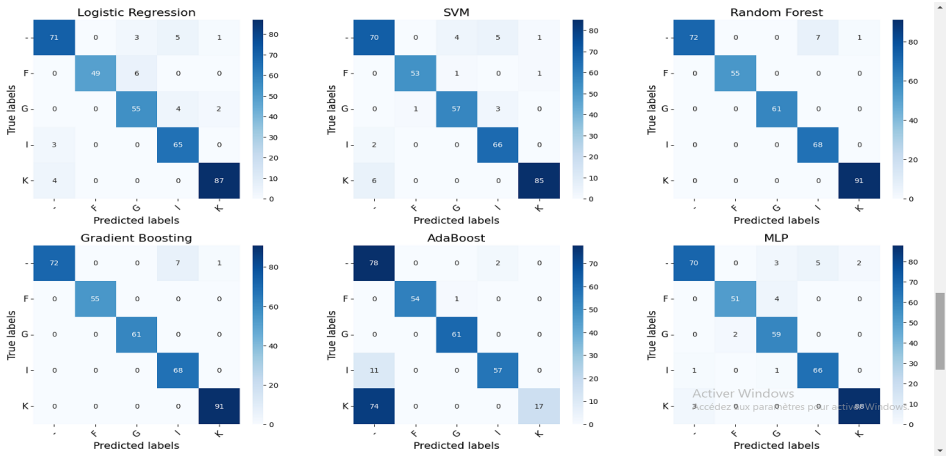

Figure 5: Confusion Matrices for Different Classifiers using original feature set.

(LSTM) model attained an accuracy of 91.27%, as shown in Table 6.

Table 6: Performance of Deep Learning Models Using Original Feature Set

Model	**Accuracy**
LSTM	0.9127
CNN	0.9437
CNN-LSTM	0.9493
DNN	0.9493

4.4. Performance of Black box Models using FS techniques

4.4.1. Machine Learning Models with selected set

Feature selection had varying effects on model performance. Using Extra Trees Classifier (ETC), Logistic Regression experienced a slight improvement in test accuracy, increasing from 0.921 to 0.927, while the Multi-Layer Perceptron (MLP) also saw a modest gain from 0.941 to 0.944. AdaBoost's performance improved when trained on ETC-selected features, whereas the Support Vector Machine (SVM) experienced a slight decrease in test accuracy from 0.932 to 0.924. Random Forest and Gradient Boosting remained stable. Mutual Information (MI) feature selection caused slight declines in the performance of Random Forest (0.972) and Gradient Boosting (0.969), while MLP (0.941) and SVM (0.924) accuracies remained consistent. However, AdaBoost's accuracy dropped significantly from 0.752 to 0.693 when using MI-selected features. RFE had the most positive effect on MLP, increasing its test accuracy from

0.941 to 0.961; AdaBoost also experienced a modest improvement. Other models exhibited minimal changes in performance. Overall, AdaBoost and MLP benefited the most from feature selection, whereas ensemble models such as Random Forest and Gradient Boosting remained robust across different feature subsets. Table 7 and Figure 6 present a performance comparison of machine learning models before and after feature selection.

Table 7: Performance Summary of Machine Learning Models After Feature Selection (with 95% Confidence Intervals)

Features	**Model**	**Train Acc.**	**Test Acc.**	**Precision**	**Recall**	**F1 Score**	**95% CI**
Original	LR	0.930	0.921	0.924	0.921	0.922	[0.904, 0.938]
	SVM	0.955	0.932	0.934	0.932	0.933	[0.917, 0.948]
	RF	1.000	0.978	0.979	0.978	0.977	[0.968, 0.987]
	GB	1.000	0.978	0.979	0.978	0.977	[0.968, 0.987]
	ADA	0.764	0.752	0.873	0.752	0.721	[0.725, 0.779]
	MLP	0.995	0.941	0.942	0.941	0.941	[0.926, 0.955]
ETC	LR	0.920	0.927	0.929	0.927	0.927	[0.911, 0.943]
	SVM	0.934	0.924	0.926	0.924	0.924	[0.907, 0.940]
	RF	1.000	0.975	0.976	0.975	0.975	[0.965, 0.984]
	GB	1.000	0.975	0.977	0.975	0.975	[0.965, 0.984]
	ADA	0.907	0.876	0.888	0.876	0.875	[0.856, 0.897]
	MLP	0.985	0.944	0.945	0.944	0.944	[0.929, 0.958]
MI	LR	0.921	0.924	0.927	0.924	0.924	[0.907, 0.940]
	SVM	0.928	0.924	0.927	0.924	0.924	[0.907, 0.940]
	RF	1.000	0.972	0.974	0.972	0.972	[0.962, 0.982]
	GB	1.000	0.969	0.970	0.969	0.969	[0.958, 0.980]
	ADA	0.717	0.693	0.834	0.693	0.626	[0.664, 0.722]
	MLP	0.986	0.941	0.941	0.941	0.941	[0.926, 0.955]
RFE	LR	0.920	0.921	0.923	0.921	0.922	[0.904, 0.938]
	SVM	0.936	0.932	0.934	0.932	0.932	[0.917, 0.948]
	RF	1.000	0.975	0.977	0.975	0.975	[0.965, 0.984]
	GB	1.000	0.972	0.976	0.972	0.972	[0.962, 0.982]
	ADA	0.895	0.887	0.905	0.887	0.888	[0.868, 0.907]
	MLP	0.969	0.961	0.963	0.961	0.960	[0.949, 0.973]

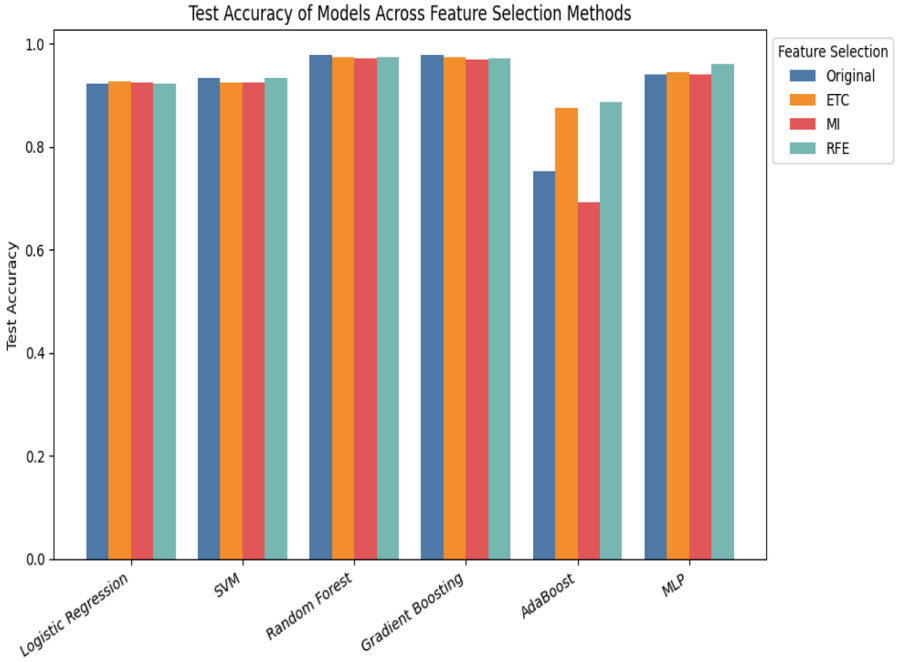

Figure 6: Comparison of Classifier Performance with Original Features vs. Selected Features.

Evaluating the performance of machine learning models is critical, but determining whether one model is significantly better than another requires more than just comparing raw accuracy or F1-scores. To assess this rigorously, we analyzed test accuracies using bootstrapped 95% confidence intervals (CIs), derived from the normal approximation to the binomial distribution. Confidence intervals provide a range in which we expect the true performance to lie with 95% certainty, accounting for sample variability[6]. This statistical approach helps determine whether observed performance differences reflect genuine improvements or are simply due to random chance[6].

The results clearly show that Random Forest (RF) and Gradient Boosting (GB), especially when paired with ETC or RFE feature selection, consistently perform best. For example, RF with RFE reached a test accuracy of 0.975 with a tight CI of [0.965, 0.984], far above models like AdaBoost, whose intervals fall much lower. While MLP also performed well, especially with RFE, its confidence interval slightly overlaps with RF and GB, making it strong but not clearly better. Overall, RF and GB stand out as statistically reliable top performers in this comparison Figure 7 shows the accuracy comparison across models, along with their bootstrapped 95% confidence intervals, highlighting which models are statistically superior.

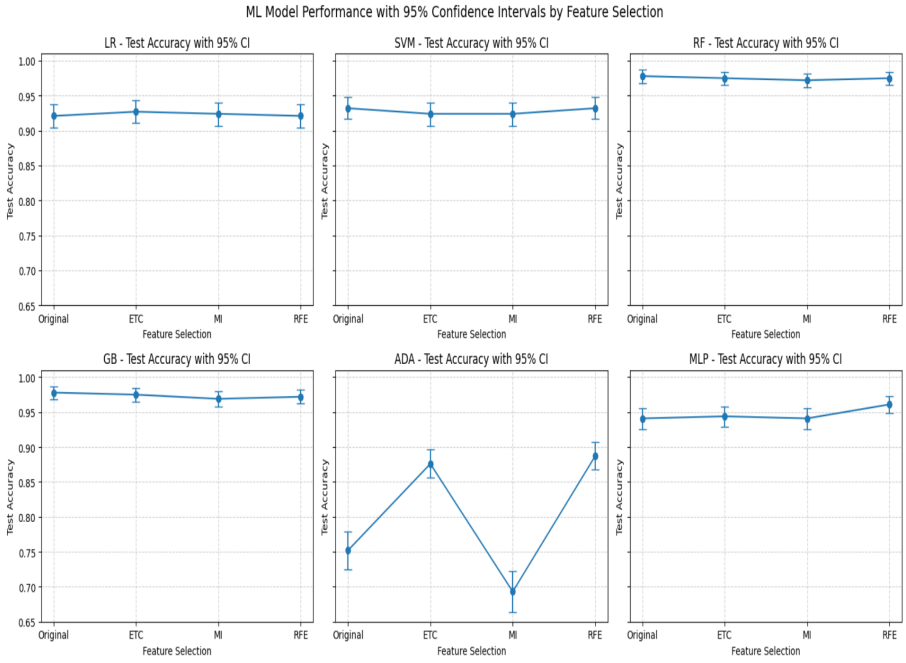

Figure 7: Accuracy Comparison with Bootstrapped Confidence Intervals .

4.4.2. Deep Learning Models with Selected Features

The impact of feature selection on deep learning model performance varied significantly depending on the architecture. As shown in Table 8, the DNN model consistently delivered high and stable performance across all feature selection techniques. Its accuracy remained above 0.949 regardless of the method used, with very tight 95% confidence intervals, suggesting that DNN can effectively adapt to reduced feature sets without compromising predictive power.

On the other hand, the LSTM and CNN models were more sensitive to feature reduction. For example, LSTM's performance dropped from 0.911 (CI: [0.892, 0.927]) with the original features to 0.812 (CI: [0.787, 0.835]) when RFE was applied. CNN also showed a significant decline when using MI (0.740, CI: [0.712, 0.766]) and RFE (0.790, CI: [0.764, 0.814]), indicating that these techniques likely removed important spatial or temporal patterns crucial to the models.

Interestingly, the CNN-LSTM hybrid model, which initially achieved the highest accuracy (0.952, CI: [0.937, 0.964]), also showed a noticeable decline when feature selection was applied, particularly with MI, where accuracy fell to 0.819. This suggests that even robust hybrid models are not immune to performance degradation when critical input information is excluded.

Overall, DNN proved to be the most resilient model across all conditions. Figure

8 illustrates the test accuracy of deep learning models with 95% confidence intervals, offering a visual comparison of both performance and reliability across architectures.

Table 8: Test Accuracy and 95% Confidence Intervals for Deep Learning Models with Feature Selection

Model	**Feature Selection**	**Accuracy**	**95% CI Lower**	**95% CI Upper**
LSTM	Original	0.911	0.892	0.927
	ETC	0.869	0.847	0.889
	MI	0.908	0.888	0.924
	RFE	0.812	0.787	0.835
CNN	Original	0.944	0.928	0.957
	ETC	0.916	0.897	0.932
	MI	0.740	0.712	0.766
	RFE	0.790	0.764	0.814
CNN-LSTM	Original	0.952	0.937	0.964
	ETC	0.919	0.900	0.934
	MI	0.819	0.794	0.842
	RFE	0.862	0.839	0.882
DNN	Original	0.946	0.930	0.958
	ETC	0.951	0.936	0.963
	MI	0.951	0.936	0.963
	RFE	0.949	0.934	0.961

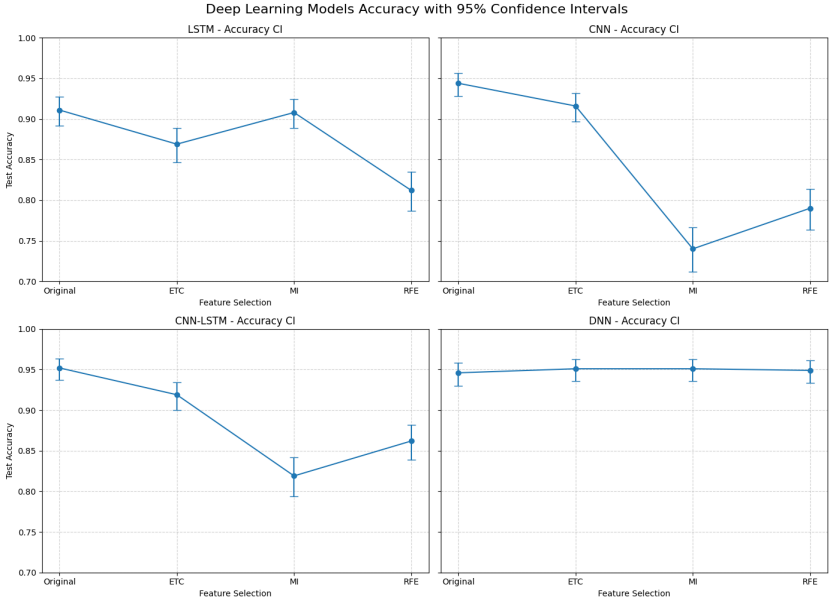

Figure 8: Test Accuracy of Deep Learning Models with 95% Confidence Intervals.

5. Model Interpretability: LIME and SHAP Analysis

As predictive performance alone is not sufficient in clinical applications, it is essential to understand how machine learning models make their decisions [11][7]. This section introduces an interpretation framework for the Gradient Boosting model using Explainable Artificial Intelligence (XAI) techniques. Specifically, we apply two popular post-hoc methods—LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations)—to gain insights into individual predictions and overall feature importance.

5.1. Model Interpretation Using LIME

LIME is a user-focused approach designed to bridge the gap between AI systems and human users. It concentrates on two fundamental aspects: model confidence and prediction confidence. By delivering explanations at the local level, LIME offers a distinctive and robust explainable AI system that elucidates individual predictions, thereby fostering trust and understanding [8][7]. Our proposed model aims to balance precision and interpretability through a multi-step approach. First, we used the Extra Trees Classifier to identify the most important features from the dataset. ETC provides a global perspective on feature importance, highlighting which features most influence the overall model's performance. Next, Random Forest and Gradient Boosting models stood out with their high accuracy rates and performance metrics. Given this, we selected the Gradient Boosting Classifier as our black-box model for explanation. However, the complexity of Gradient Boosting models necessitates understanding the attributes it emphasizes during predictions.

LIME complements this approach by providing local explanations for individual predictions. It approximates the complex model with a simpler, interpretable model to explain the contribution of specific features such as TSH, T3, TT4, T4U, and others. Figures 5–9 show the local explanations for the model that we chose. We interpret five prediction sample cases for each label ("Normal ","primary hypothyroid","compensated hypothyroid", "increased binding protein"and "non-thyroidal illness syndrome"). For each instance, we analyze four LIME outcomes: the classification probabilities for each label, the LIME explanation for the selected features, and a table organizing the features and their corresponding values.plus the graph that represents Individual parameter contribution plot of patient.

Based on the LIME prediction probabilities and feature values shown in Figure 6, the model classified the patient as "Normal" due to optimal TSH and T3 levels, lower T4U, and high FTI—strong indicators of normal thyroid function. Despite negative influences such as higher TT4 levels, older age, female sex, and being on thyroxine treatment, the positive features significantly outweighed these, resulting in a confident classification.

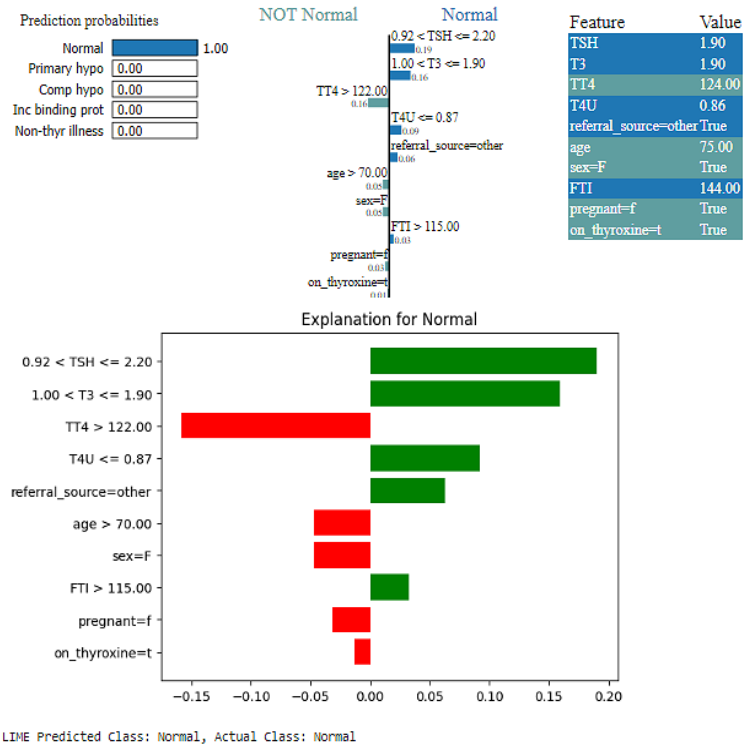

Figure 9: LIME for a case classified as “”Normal” by GBM model.

The LIME results for the prediction of ”Primary Hypothyroidism” highlight several features with positive and negative impacts. Features such as low TT4, low FTI, being female, older age (age>70), and not being on thyroxine significantly contribute to the model’s prediction, indicating a strong association with ”Primary Hypothyroidism”. On the other hand, features such as high TSH, moderate T3 levels, not being pregnant, T3 being measured, and referral source being other negatively impact the prediction, showing a weaker association with this condition. Figure 7 shows LIME results for a case classified as “Primary Hypothyroidism”.

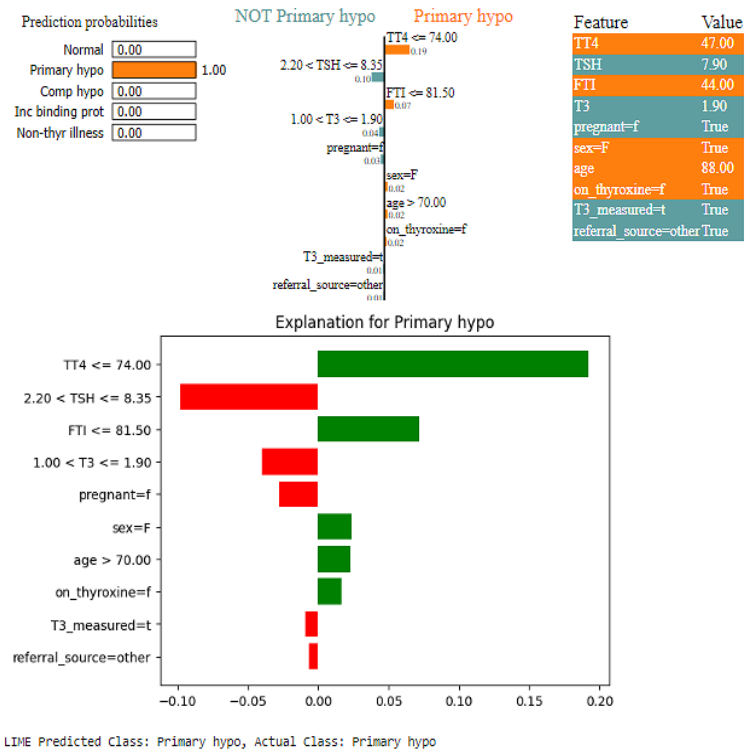

Figure 10: LIME for a case classified as “Primary Hypothyroidism” by GBM model.

In Figure 8, The LIME results for the model’s prediction of ”Compensated Hypothyroidism” highlight the positive and negative impacts of various features. Features such as T3, TT4, FTI, being male (sex=M), TSH levels, and T4U levels positively influence the prediction, indicating that they contribute significantly to the model’s confidence in classifying the condition as ”Compensated Hypothyroidism.” Conversely, features such as not being pregnant, having T3 measured, being older than 70, and not being on thyroxine negatively impact the prediction, suggesting that these factors are less associated with ”Compensated Hypothyroidism”.

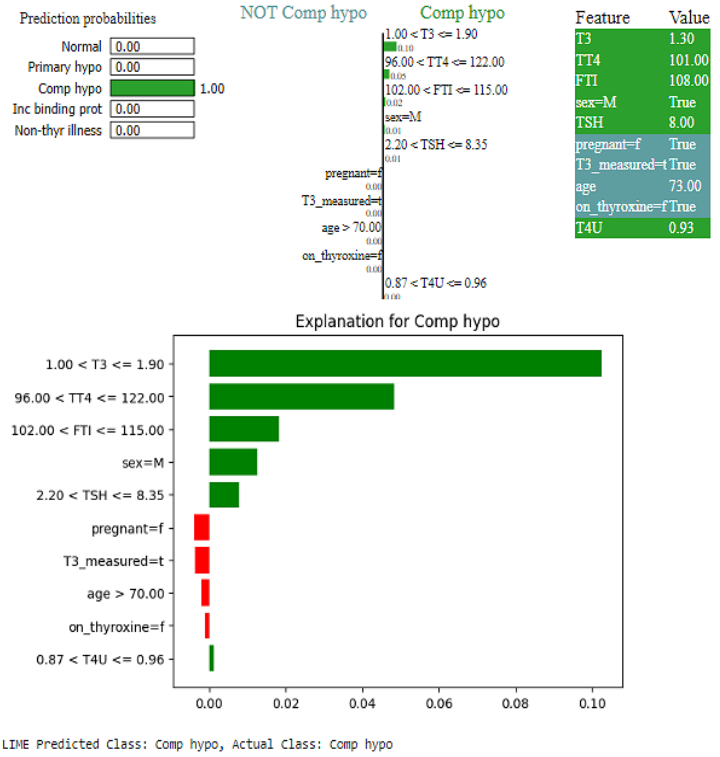

Figure 11: LIME for a case classified as "Compensated Hypothyroidism" by GBM model.

The fourth class, "increased binding protein", as shown in Figure 9, reveals positive and negative impacts of different features. Features such as high T4U, high TT4, low TSH, being female (sex = F), younger age ($\leq$ 37 years), reference source other than SVI, and unmeasured T3 positively influence the prediction, indicating their significant contribution to the model's confidence in classifying the condition as "increased binding protein." On the other hand, features such as high FTI, not being pregnant, and T3 within the specified range negatively impact the prediction, suggesting that these factors are less associated with "increased binding protein".

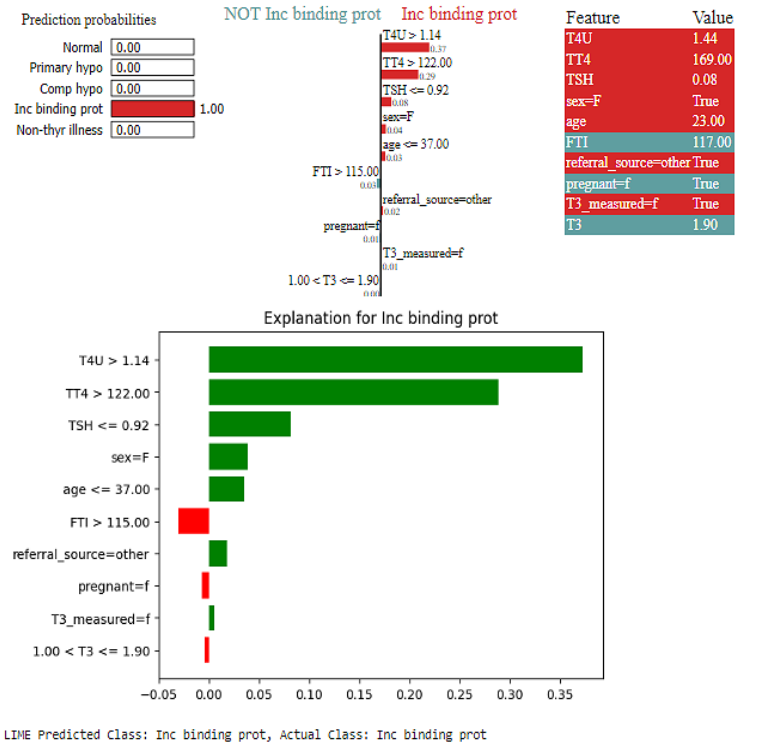

Figure 12: LIME for a case classified as “increased binding protein ” by GBM model.

Figure 10 shows that for the given instance, the model predicted ”Non-thyroidal illness” with absolute certainty. LIME’s local explanations reveal that features such as T3 $\leq$ 1.00 and referral-source = SVI positively influence this prediction, indicating their strong association with the ”Non-thyroidal illness” diagnosis. Conversely, features like FTI within the range of 81.5 to 102 and age between 37 and 57 exhibit negative impacts, suggesting they are less indicative of this condition. These insights are invaluable for clinicians as they align the model’s predictions with clinical knowledge, ensuring both accuracy and transparency in diagnosing thyroid conditions.

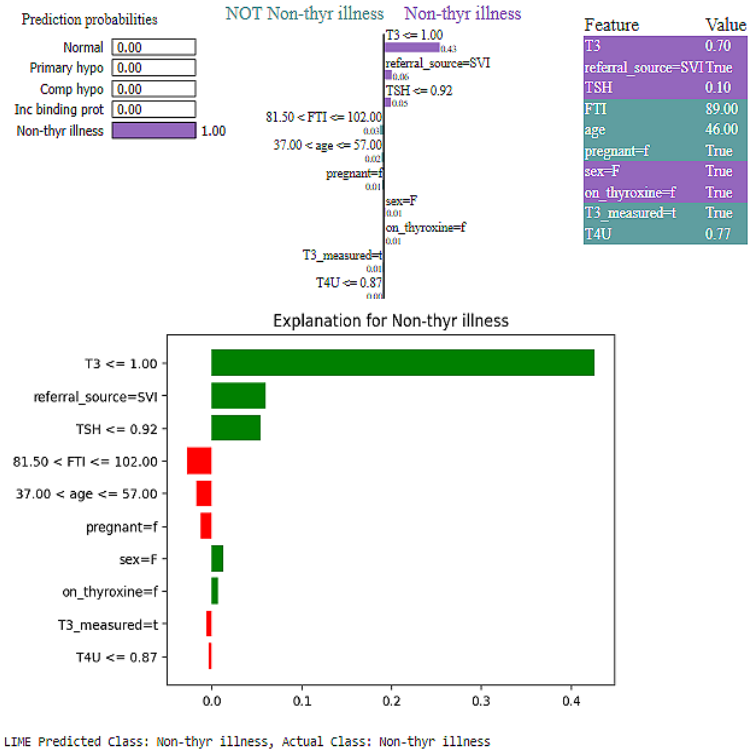

Figure 13: LIME for a case classified as "non-thyroid disease" by GBM model.

The LIME results illustrate that the ranking and importance of features change depending on the predicted thyroid condition. This dynamic nature ensures the model's rationale for each prediction is transparent and understandable. Figure 11 shows the explanation of the model's prediction for a random test instance using LIME. The model predicts this instance as "normal" with a high probability (0.86) and a notable probability (0.14) for "Inc binding prot". Other classes (primary hypothyroidism, complementary hypothyroidism, and non-thyroid disease) have very low probabilities.

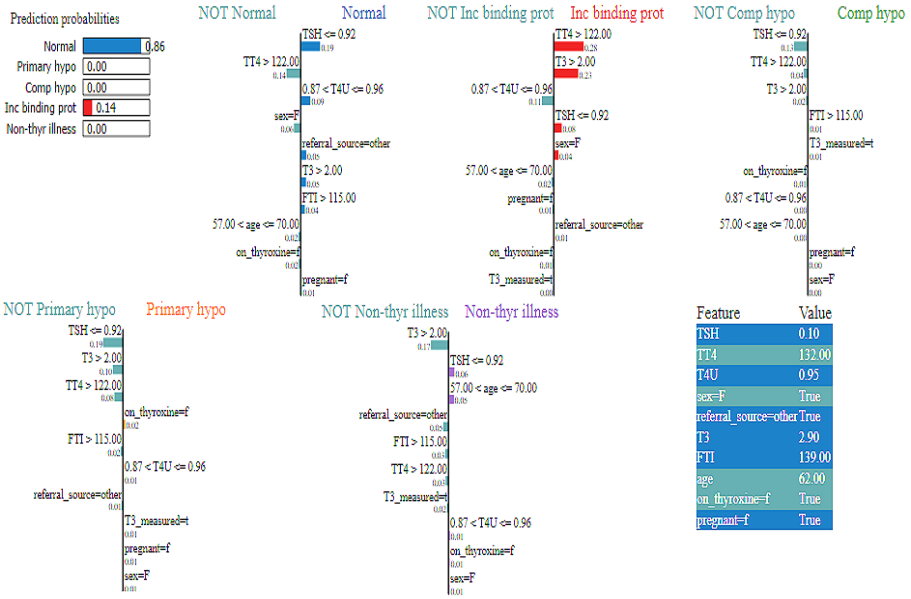

Figure 14: Holistic View of Feature Contributions Across All Classes Using LIME.

Examining feature contributions across all classes—not just the predicted one—provides a fuller understanding of how the model uses features to make predictions. This holistic view is essential for validating model decisions, especially in critical applications like healthcare. For example: - Features pushing the prediction toward normal (e.g., low TSH) - Features pushing toward primary hypothyroidism (e.g., low T4)

5.2. Model Interpretation Using SHAP

In addition to applying LIME, we used Shapley Additive Explanations (SHAP) to examine feature importance in our Gradient Boosting model and assess global contributions. SHAP systematically quantifies and visualizes how individual features influence predictions by measuring output changes when features are added or removed. For a comprehensive overview, we created a SHAP beeswarm plot to highlight the relative importance of features across the entire dataset. To further clarify model behavior, we developed five visualizations. Each shows how features impact predictions for a specific class, offering detailed insights into class-specific decision patterns.

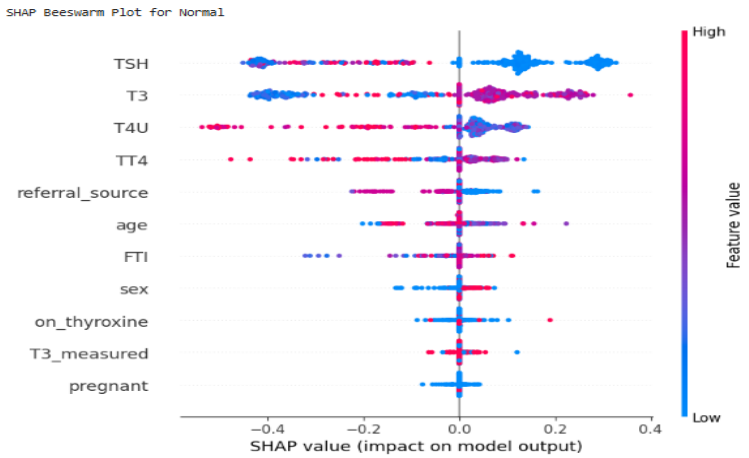

Figure 15: SHAP Beeswarm Plot Representing Feature Contributions for the Normal Class.

The SHAP beeswarm plot above provides a global explanation. Each point represents a SHAP value for an individual instance, with colors indicating feature values: red for high, blue for low. values. The x-axis represents the SHAP values, where positive values push the model toward predicting the "Normal" class, and negative values pull it away from this classification.

From Figure 14, we observe that higher values of T3 and T4U tend to have a positive impact on predicting the Normal class, pushing the model's output towards this classification. Conversely, TSH plays a significant role in the opposite direction, where lower TSH values strongly contribute negatively to the prediction of Normal, meaning that higher TSH levels reduce the likelihood of being classified as Normal. TT4 shows a mixed effect, where both high and low values impacting the classification in varying ways. Other features such as FTI, age, and referral source exhibit less pronounced but still noticeable contributions to the model's decision-making process. While their impact is relatively smaller, they follow a similar trend higher values tend to support the Normal classification, whereas lower values can lead the model toward alternative diagnoses.

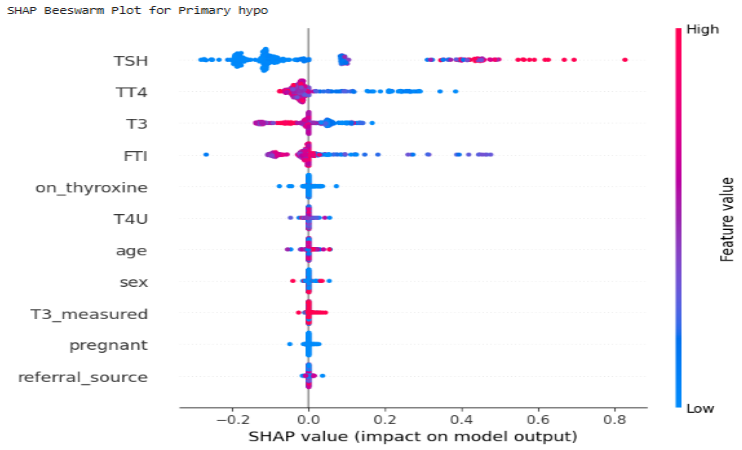

Figure 16: SHAP Beeswarm Plot Representing Feature Contributions for the Primary Hypothyroidism Class.

The SHAP beeswarm plot in Figure 15 highlights the key factors influencing the model's prediction of the "Primary Hypo" class. The most significant predictor is TSH, where higher values strongly increase the likelihood of a Primary Hypothyroidism diagnosis, while lower values decrease it. Conversely, lower levels of TT4, T3, and FTI contribute positively to this classification, whereas higher levels have a negative impact. These findings align with clinical knowledge, as primary hypothyroidism is typically associated with elevated TSH and reduced thyroid hormone levels. Other features, such as T4U, age, and referral source, play relatively minor roles in the model's decision-making process.

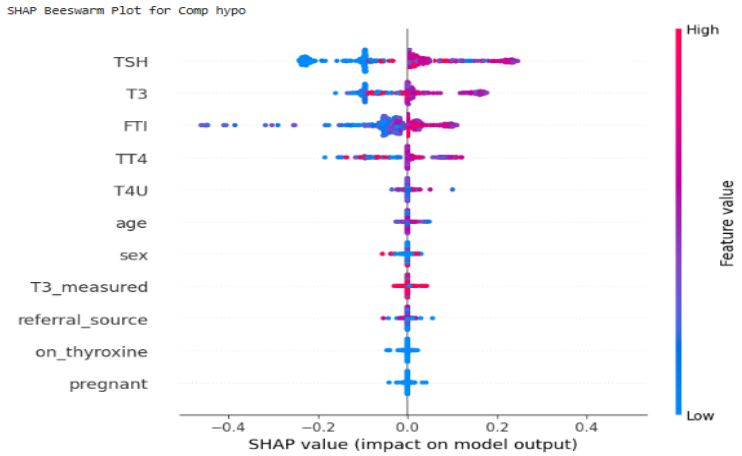

Figure 17: SHAP Beeswarm Plot Representing Feature Contributions for the Compensated Hypothyroidism Class.

The SHAP beeswarm plot for the Compensated Hypothyroidism class in Figure 16 shows that higher TSH levels strongly contribute to predicting this condition, while lower TSH values negatively impact the classification. Unlike Primary Hypo, where T3 exhibits a clear inverse relationship, its effect here appears more balanced, with both high and low values contributing to varying degrees. FTI follows a similar pattern, where lower values tend to support the classification, but its impact is not as strong as TSH. Other features, including T4U, age, referral source, and on thyroxine status, exhibit minimal influence on the prediction, indicating that the model relies primarily on TSH and thyroid-related hormone levels for predicting Compensated Hypothyroidism.

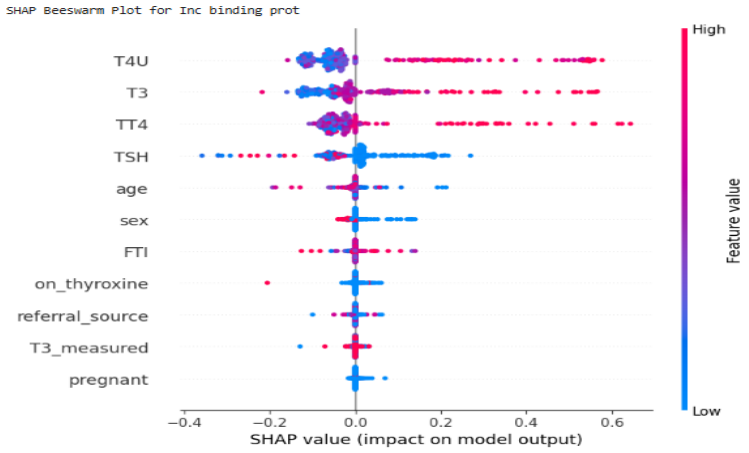

Figure 18: SHAP Beeswarm Plot Representing Feature Contributions for the Increased Binding Protein Class.

in Figure 17 The SHAP beeswarm plot explains the impact of different features on GB model predicting "Inc binding prot". Key features like T4U, T3, TT4, and TSH have the most influence, with higher values of T4U, T3, and TT4 pushing predictions higher. Lower TSH values tend to decrease the model's output. Less significant features include pregnant, referral source, and T3 measured.

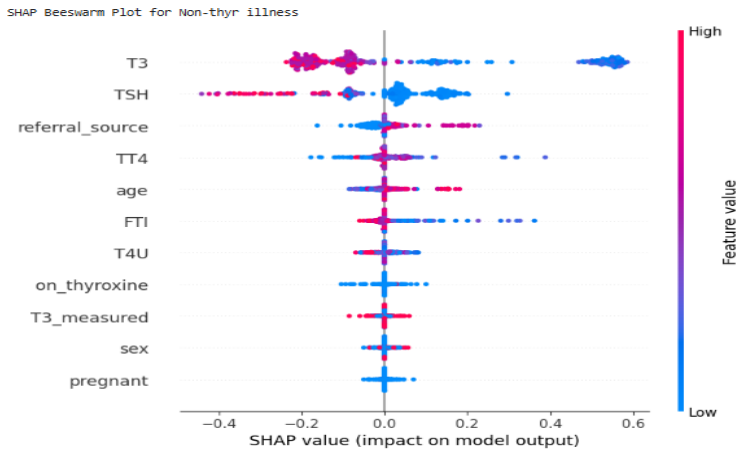

Figure 19: SHAP Beeswarm Plot Representing Feature Contributions for the Non-Thyroidal Illness Class.

The SHAP beeswarm plot in Figure 18 for the Non-Thyroidal Illness class shows that low T3 levels are the strongest predictor, significantly increasing the likelihood of this classification. High T3 values negatively impact the prediction, decreasing the chance of Non-Thyroidal Illness. Positively, mildly low TSH has a mixed influence, while TT4, FTI, and T4U show minimal impact.

5.3. Comparison Between LIME and SHAP

In this study on thyroid disorder prediction, both LIME and SHAP were employed to interpret the predictions of the Gradient Boosting Machine (GBM) model. LIME offered fast, case-specific explanations by locally approximating the model with linear models, making it suitable for real-time applications but less reliable due to variability and a lack of global context. In contrast, SHAP provided comprehensive, theoretically grounded global insights by fairly attributing importance to all features, though at a higher computational cost. This contrast is clearly illustrated in Table 11, where we compared LIME and SHAP. The study highlighted key trade-offs: LIME was simpler and faster but less stable, while SHAP was more robust and interpretable but computationally intensive. These methods jointly offer a complementary interpretability strategy for balancing local interpretability with global model understanding.

6. Discussion and Comparison with Existing Studies

Our study stands out due to its unique approach to enhancing interpretability in multi-class thyroid disease prediction. While other research also calculates model performance and incorporates interpretability, several distinctive aspects of our work

Table 9: Comparative Analysis of LIME and SHAP

Aspect	**LIME**	**SHAP**
Interpretation Level	Local (individual prediction)	Global (class-based behavior via beeswarm plots)
Purpose in Study	Explain why a specific instance was classified into a certain class	Understand how features influence class predictions across the dataset
Visualization Technique	Bar plot showing top features per instance	Beeswarm plot with feature contributions and distribution
Used Model	Gradient Boosting Classifier	Gradient Boosting Classifier
Backend Mechanism	Local surrogate with perturbed samples and linear approximation	Game-theoretic Shapley values using `KernelExplainer`
Features Considered	Features selected by Extra Trees Classifier (ETC)	All selected features with contribution magnitude
Explanation Scope	One instance at a time	Class-level, multi-instance aggregated
Output Format	Sorted top-N feature bars by influence	SHAP summary plots: direction, value, distribution
Complexity	Fast and efficient	Computationally intensive (especially for multiclass)
Usefulness	Case-level insights for domain experts	Reveals deeper global feature patterns
Integration in Study	Used for interpreting individual patient predictions	Used for understanding feature influence across diagnosis classes
Advantages	– Fast and intuitive – Focused on individual cases – Clinically useful	– Stable and consistent – Explains all features – Theoretically grounded
Disadvantages	– Results can vary due to randomness – Not suitable for global explanations	– No instance-level clarity in this study – Computationally expensive

contribute to its originality:

Dataset Size and Complexity: The dataset used in our study is substantial, with 9172 observations and 31 features. This scale and detail provide a robust foundation for training and evaluating models, potentially offering more generalizable insights compared to other studies[1], [19],[20] and [22] with smaller or less detailed datasets.

Focus on Multi-Class Classification: Unlike many studies [17], [1], [19], [20] and [22] that focus on binary or limited-class classification problems, this approach addresses the challenges of multi-class classification, distinguishing it from previous research where only (Chaganti and all., 2022; Ouartani and Taleb, 2024) makes use of the same database as our research. our study involves five distinct thyroid disease categories, reflecting a more nuanced understanding of thyroid disorders.

The Impact of Feature Selection on Model Performance: Numerous studies [5],[12], [23],[19] and [22] have shown that feature selection (FS) enhances model performance by reducing noise, improving training stability, and boosting generalization. Although some models perform better with all features, our objective was to balance accuracy with interpretability. We applied three FS methods ETC, MI, and RFE, with ETC proving especially effective in ranking feature importance and identifying key performance drivers. Overall, FS not only helped prevent overfitting but also aligned with prior research, confirming its essential role in building robust and interpretable machine learning models.

Machine Learning Evaluation: Our research evaluates a diverse set of machine learning models, including (RF), (LR), (SVM), (ADA), (GBM), (MLP), and advanced deep learning models like Deep Neural Networks (DNN), (LSTM), and (CNN). Unlike most existing studies that focus exclusively on either traditional machine learning [1] and [20] or deep learning approaches, with only the study [5] conducting a similarly comprehensive comparison This comprehensive evaluation compares not only accuracy with the Gradient Boosting model achieving 97% but also precision, recall, and F1 scores, thereby assessing model robustness across various scenarios.

LIME and SHAP evaluation: The evaluation of our model using LIME and SHAP confirmed the interpretability of predictions, with explanations judged understandable and clinically relevant by an endocrinologist. LIME successfully highlighted features such as TSH and FT4, aligning with standard diagnostic practice. Similar to [8] and [7], who demonstrated LIME and SHAP's utility in diabetes and CKD prediction, our findings support the effectiveness of XAI methods in medical contexts. However, unlike studies that introduced additional complexity such as GMM-LIME proposed by [14] our approach maintained simplicity while offering both global and local interpretability. Ultimately, this dual-layered interpretability approach FS for identifying globally important features, SHAP for feature influence, and LIME for patient-level justification not only enhances transparency but also builds greater confidence in the model's decisions. As shown in other recent studies [2, 16, 4], integrating explainable AI into medical diagnosis is not only a technical necessity but also a step toward safer, more accountable clinical decision-support systems.

7. Conclusion

In this work, we proposed a comprehensive and interpretable framework for multiclass thyroid disease prediction, combining machine learning models with robust feature selection and explainable artificial intelligence (XAI) techniques. Our approach was evaluated on a large and complex clinical dataset, with Gradient Boosting achieving the highest predictive accuracy. However, the strength of our work lies not only in predictive performance but in the transparency and clarity it offers to end users. To this end, we employed three feature selection methods: Extra Trees Classifier, Mutual Information, and Recursive Feature Elimination not to boost accuracy, but to reduce dimensionality and highlight the most relevant biomarkers for interpretability. To ensure that our models remain clinically actionable, we incorporated SHAP for global insights and LIME for local, instance-level explanations. Enabling transparency at the dataset level as well as for individual predictions. This dual-layered interpretability strategy strengthens trust in model outputs and aligns with the critical need for explainable AI in healthcare. Expanding the framework to other diseases with overlapping symptoms but differing etiologies would help evaluate its generalizability and further validate its usefulness in broader medical diagnostic applications.

Acknowledgments

The authors would like to thank the DGRSDT (General Directorate of Scientific Research and Technological Development) - MESRS (Ministry of Higher Education and Scientific Research), ALGERIA, for the financial support of LISCO Laboratory.

References

[1] Ahmad, W., Ahmad, A., Lu, C., A Novel Hybrid Decision Support System for Thyroid Disease Forecasting, Soft Computing, 22, 2018, 5377–5383.

[2] Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., Humayun, M., Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP, Diagnostics, 13, 2023, 1932, https://doi.org/10.3390/diagnostics13111932.

[3] Allgaier, J., Mulansky, L., Draelos, R. L., Pryss, R., How Does the Model Make Predictions? A Systematic Literature Review on the Explainability Power of Machine Learning in Healthcare, Artificial Intelligence in Medicine, 143, 2023, 102616, https://doi.org/10.1016/j.artmed.2023.102616.

[4] Amin, A., Hasan, K., Zein-Sabatto, S., Chimba, D., Ahmed, I., Islam, T., An Explainable AI Framework for Artificial Intelligence of Medical Things, Proc. IEEE Globecom Workshops (GC Wkshps), 2023, 2097–2102.

[5] Chaganti, R., Rustam, F., De La Torre Díez, I., V. Mazon, J. L., Rodríguez, C. L., Ashraf, I., Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques, Cancers, 14, 2022.

[6] du Prel, J. B., Hommel, G., Röhrig, B., Blettner, M., Confidence Interval or P-Value? Part 4 of a Series on Evaluation of Scientific Publications, Deutsches Ärzteblatt International, 2009, https://doi.org/10.3238/arztebl.2009.0335.

[7] Ghosh, S. K., Khandoker, A. H., Investigation on Explainable Machine Learning Models to Predict Chronic Kidney Diseases, Scientific Reports, 14, 2024, 3687.

[8] Guler, H., Avcı, D., Ulaş, M., Omma, T., Performance Comparison of Machine Learning Models Powered by SHAP and LIME Based Explainability Techniques on Diabetes Dataset, SSRN, 2024, https://ssrn.com/abstract=4713039.

[9] Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F., Factual and Counterfactual Explanations for Black Box Decision Making, IEEE Intelligent Systems, 34, 6, 2019, 14–23.

[10] Hakkoum, H., Idri, A., Abnane, I., Assessing and Comparing Interpretability Techniques for Artificial Neural Networks in Breast Cancer Classification, Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization, 2021, https://doi.org/10.1080/21681163.2021.1901784.

[11] Hu, Y., Sokolova, M., Explainable Multi-class Classification of the CAMH COVID-19 Mental Health Data, arXiv preprint, 2021, arXiv:2113.13430.

[12] Junkang, A., Zhang, Y., Joe, I., Specific-Input LIME Explanations for Tabular Data Based on Deep Learning Models, Applied Sciences, 13, 15, 2023, 8782.

[13] Kaggle, Thyroid Disease Dataset, https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-diseasedata.

[14] Mulwa, M. M., Mwangi, R. W., Mindila, A., GMM-LIME Explainable Machine Learning Model for Interpreting Sensor-Based Human Gait, Engineering Reports, 2024.

[15] Ouartani, S., Taleb, N., Decision Support System for Thyroid Disease Prediction Using Decision Tree Algorithm and Ontology, Artificial Intelligence Theory and Applications (AITA), 2024.

[16] Rahman, M., Ali, M., Mahim, M., Miah, S. M., Sipon, M., Enhancing Lung Abnormalities Detection and Classification Using a Deep CNN and GRU with Explainable AI, Machine Learning with Applications, 14, 2023, 100492, https://doi.org/10.1016/j.mlwa.2023.100492.

[17] Rawte, V., Royl, B., V. L., Thyroid Disease Diagnosis using Ontology-based Expert System, International Journal of Engineering Research Technology (IJERT), 4, 6, 2015.

[18] Settouti, N., Saidi, M., Preliminary Analysis of Explainable Machine Learning Methods for Multiple Myeloma Chemotherapy Treatment Recognition, Evolutionary Intelligence, 17, 2024, 513–533.

[19] Shiuh, T. L., Khai, W. K., Xin, Y. C., Wai, C. Y., Prediction of Thyroid Disease Using Machine Learning Approaches and Featurewiz Selection, JTEC, 15, 3, 2023, 9–16.

[20] Siddhartha, K., Abhishek, A., Gyanendra, S., Developing an Explainable Machine Learning-Based Thyroid Disease Prediction Model, International Journal of Business Analytics (IJBAN), 9, 3, 2022, 1–18.

[21] Sun, Q., Akman, A., Schuller, B. W., Explainable Artificial Intelligence for Medical Applications: A Review, arXiv preprint, 2024, arXiv:2412.01829.

[22] Tang, J., Alelyani, S., Liu, H., Feature Selection for Classification: A Review, in: C. C. Aggarwal (Ed.), Data Classification: Algorithms and Applications, Chapman and Hall/CRC, 2014, 37–64.

[23] Zacharias, J., von Zahn, M., Chen, J., Designing a Feature Selection Method Based on Explainable Artificial Intelligence, Electronic Markets, 2022, https://doi.org/10.1007/s12525-022-00608-1.

Received 04.09.2024, Accepted 01.06.2025