Artificial Intelligence Meets Endocrinology: A Machine Learning-Based Approach to Thyroid Disease Diagnosis Using Feature Selection Methods

Aftab Ahmad Khan[1], Bakhtiar Khan[1], Muhammad Arif[1], Waseel ud Din[2], Wahab Khan[1], Yasir Tayyab Khayyam[4] Ashraf Ullah[1], Kalim Ullah[3]
[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[2]Department of Computer Science, Birmingham City University, United Kingdom
[3]Department of Electrical Engineering, University of Science & Technology Bannu, Pakistan
[4]Gomal Research Institute of Computing (GRIC), Faculty of Computing, Gomal University, D.I. Khan, K.P.K, Pakistan
*** Correspondence:** Aftab Ahmad Khan; Email: aftabaak7@gmail.com

Citation | Khan. A. A., Khan. B, Arif. M, Din. W, Khan. W, Khayyam. Y. T, Ullah. A, Ullah. K, "Artificial Intelligence Meets Endocrinology: A Machine Learning-Based Approach to Thyroid Disease Diagnosis Using Feature Selection Methods", IJIST, Vol. 07 Issue. 04 pp 2383-2398, October 2025

Received| August 31, 2025 **Revised|** October 08, 2025 **Accepted|** October 10, 2025 **Published|** October 12, 2025.

Thyroid Disease (TD) arises when the thyroid gland either grows abnormally or does not generate enough thyroid hormones, and might cause serious health issues and consequences. Early and efficient identification of thyroid disease is important for improved clinical intervention and disease management. By combining sophisticated and advanced machine learning models with a range of advanced feature selection strategies, this research study aims to enhance the classification of thyroid disease based on a machine learning based diagnostic system. The preprocessed dataset used in this study and the trials were taken from the machine learning repository at the University of California, Irvine (UCI). We employ two popular feature selection techniques- Chi-Square, and Recursive Feature Elimination, and a dimensionality reduction technique Linear Discriminant Analysis (LDA), and to choose the best features from the dataset for experiments. After selecting the most suitable features, they were then used to train and test the machine learning models: Multi-Layer Perceptron (MLP), Gradient Boost (GB), and Recurrent Neural Network (RNN). Evaluation matrices, accuracy, precision, recall, and F1-score were used to assess models' performance. The experimental results show that the machine learning model Gradient Boost (GB) outperformed the other models and yielded an accuracy of 99%, indicating its ability to classify the Thyroid Disease (TD) accurately. The proposed research work helps to create an intelligent decision-support system for medical diagnostics by offering an understandable and reliable framework for Thyroid Detection.

Keywords: Thyroid Disease, Hormones, Feature Selection, Linear Discriminant Analysis, Chi-Square, Recursive Feature Elimination, Machine Learning Based Diagnostic System

Introduction:

The thyroid, one of the body's most vital organs, produces hormones essential for numerous physiological processes. Once released into the bloodstream, these hormones circulate throughout the body, regulating growth and metabolism. This indicates that the thyroid directly influences how the body utilizes energy and sustains overall physiological balance. The thyroid gland is perfectly positioned to perform its functions because it is located in the neck, just below the Adam's apple. Understanding thyroid function is essential, as it serves as a key basis for investigating and diagnosing various disorders linked to hormonal imbalances. By producing hormones necessary for growth and metabolism, the thyroid ensures that the body's energy needs are met, promoting overall health and well-being [1]. Thyroid disease is ranked 2nd after diabetes by the prominent healthcare organization WHO (World Health Organization). Iodine-deficient areas are home to almost one-third of the world's population. In regions where iodine intake falls below 25 micrograms per day, congenital hypothyroidism is common, while areas with daily iodine consumption under 50 micrograms typically experience a higher prevalence of goitre. [2]. The thyroid gland is composed of two basic hormones known as triiodothyronine (T3) and levothyroxine (T4). T3 and T4 are the iodine-rich hormones in vertebrates, produced by the iodine consumed with food, which control the blood pressure, body temperature, and heartbeat. The inside pituitary gland synthesizes the serum thyrotropin (TSH), which in turn is responsible for regulating the production of T3 and T4 hormones. [3][4][5].

Iodine deficiency is the leading cause of thyroid disease, though other factors may also contribute. Thyroid conditions generally occur in three states: euthyroidism, hyperthyroidism, and hypothyroidism. Euthyroidism refers to normal hormone production, hyperthyroidism to excessive hormone production, and hypothyroidism to insufficient or impaired hormone production in the body. The primary factor of thyroid disease is the deficiency of iodine; however, other factors also cause thyroid disease. Three states of thyroid disease are "Euthyroidism", "hyperthyroidism", and the last one is "hypothyroidism", where Euthyroidism indicates normal production of hormones, hyperthyroidism indicates more production of hormones, while hypothyroidism indicates faulty production of thyroid hormones in the human body. In iodine-sufficient populations, women are up to ten times more likely than men to develop hyperthyroidism, affecting approximately 0.5–2% of the female population [2]. The primary causes of Hypothyroidism are inadequate thyroid hormone production and inadequate alternative therapy. Between 1% and 2% of people in iodine-depleted areas have hypothyroidism, which is more prevalent in elderly women and ten times more likely in women than in males [2]. The Thyroid disease is a lifelong disease, but Transient thyroid disease is a short-term disease and lasts for several months or less than one year [6]. Studies show that around 12% of individuals experience thyroid disease at some point in their lifetime, with women being 5–8 times more likely than men to be affected. This higher prevalence among women is partly due to pregnancy, during which the thyroid gland enlarges by about 10% in iodine-sufficient regions and by 20–40% in iodine-deficient areas [4].

The integration of machine learning and deep learning into the healthcare industry offers promising results by providing automatic and improved detection accuracy. A machine learning model can learn to find subtle patterns that could point to a disease based on clinical data, which has complex and nonlinear connections and relationships. However, the high-dimensional nature of medical data, which frequently contains duplicate and irrelevant features, poses substantial issues and potentially leads to model overfitting and increases the computational cost. Although earlier research has demonstrated that individual models such as Random Forest or XG-Boost are effective for thyroid diagnosis, a thorough examination of the interactions between various feature selection techniques and various model architectures, from ensemble approaches to deep neural networks, remains understudied. By

putting forth a strong machine learning (ML)-based detection framework that methodically examines the synergy between sophisticated feature selection methods and competent machine learning algorithms, this study seeks to close this gap. The goal of our work is to construct intelligent clinical decision-support systems by identifying the most effective and interpretable pipeline for thyroid disease detection, in addition to achieving high accuracy.

Related Work:

The primary cause of thyroid disease is the abnormal growth of thyroid tissues. Disorders occur when the gland produces either an excess or a deficiency of hormones relative to the normal range. In this study, the authors analyzed the dataset using L1- and L2-based feature selection methods, achieving 100% accuracy with Naïve Bayes and Logistic Regression, while KNN attained an accuracy of 97.8% [3]. Machine learning plays a vital role in the timely detection of thyroid disease, and timely detection indeed leads to timely treatment. The authors [1] classify the data based on sampled and unsampled datasets and achieve an accuracy of 94.8% with the random forest ensemble method.. The thyroid gland, one of the largest endocrine organs in the human body, plays a key role in regulating metabolism, and its early disease detection can significantly reduce mortality rates. The CNN generates an accuracy of 97% with medical images, 94% with ultrasound images, and CT scan images [7]. Human health is important, and society tries to care for the patient as quickly as possible. Thyroid is one of the diseases that affects the global population. Artificial intelligence offers effective methods for thyroid disease detection; for instance, applied XGBoost and achieved an accuracy of 99%, demonstrating outstanding results

Thyroid disease results from the abnormal production of TSH, T3, and T4 hormones, and many patients remain untreated due to delayed or missed detection. Machine learning assists healthcare by detecting such diseases earlier; for this purpose, KNN, Naïve Bayes, Random Forest, and other algorithms are used [5]. Thyroid disease increases with the passage of time since 1990, thyroid cancer become another rising problem, deep CNN is used for thyroid cancer detection due to its prominent results, and applied by the author with the accuracy rate of 90.8% for female patients and 90.1% for male patients [8]. The use of machine learning and artificial intelligence in healthcare has expanded rapidly, with early-stage disease detection being one of its key applications. Automated systems for detecting thyroid disease are particularly crucial, as they can save both lives and healthcare costs. [9]. Thyroid disease is increasing rapidly worldwide and affects people in India as well. This study applies various methods, including KNN, XGBoost, Logistic Regression, and Decision Tree, with XGBoost achieving the highest accuracy of 98.5% [10]. In the Human body, all metabolic processes and reproductive activities that are necessary for neuron development and growth of the human body depend on thyroid hormones; a deficiency might affect the population, as well as have some negative effects. For early detection, the authors [11] use Type-2 Fuzzy SVM with various optimization techniques and achieve 99% accuracy. Machine learning models are widely used as a tool for thyroid disease detection. The authors employed filter-based feature selection combined with a stacked ensemble approach, ultimately achieving an accuracy of 99.9% in their experiments [12].

Despite the promising results, our proposed work aims to address a number of limitations in the current literature. First, without doing a thorough analysis of the effects of feature selection tactics/techniques, many researchers describe performance mainly in terms of accuracy only. For instance, the study conducted by compares several algorithms, but they don't focus on feature selection methods. Furthermore, a comparative examination of the numerous FS method types (filter, wrapper, and embedding) applied to the identical dataset and models is inadequate. Although they employed L1/L2 regularization, a type of embedded FS, they failed to contrast it against alternative strategies such as Chi-Square, a filter technique, or Recursive Feature Elimination (RFE, a wrapper method).

Objectives:

This research is driven by the hypothesis that the integration of machine learning models with feature selection and dimensionality reduction methods can significantly enhance the diagnosis of thyroid disease. The research study objectives are:

To validate whether the feature selection methods like RFE and Chi-Square, and the dimensionality reduction method LDA, can improve the performance of RNN, GB, and MLP by reducing computational time and complexity, and overfitting.

To investigate which feature selection method, RFE or Chi-Square, is most effective in identifying critical biomarkers for detection across different model architectures.

To establish a benchmark for thyroid disease detection by comparing the efficiency of each model before and after selection of the most appropriate feature from the dataset.

Novelty Statement:

The novelty of our research study lies in the architecture, a comparative framework that rigorously evaluates two feature selection methods, Chi-Square and RFE, and a dimensionality reduction method, LDA, across a suite of models, including MLP, GB, and RNN. This approach allows us not only to identify the best performance model but also the most effective feature selection integration, providing a nuanced understanding that is currently missing from the literature.

Methodology:

Figure 1 illustrates the proposed methodology of this study, encompassing data preprocessing, feature selection, model implementation, and evaluation.

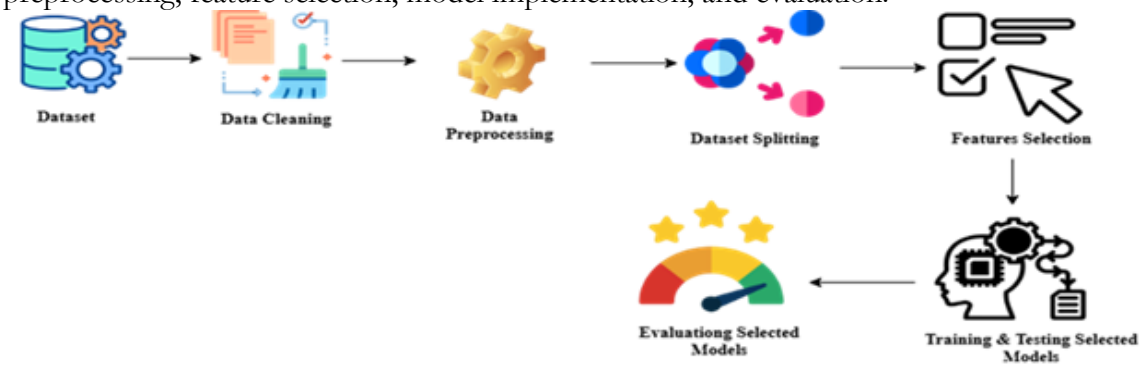

Figure 1. Proposed working methodology for the experiments

Dataset Details:

The dataset used in the experiments was sourced from the Garvan Institute repository in Sydney, Australia. It comprises 3,772 instances with 30 distinct features. Of the 30 features, 20 are categorical and the remaining are continuous. Most categorical features are binary (True/False), including: query on thyroxine, on antithyroid medication, sick, pregnant, thyroid surgery, I131 treatment, query hypothyroid, query hyperthyroid, lithium, goiter, tumor, hypopituitary, psych, TSH measured, T3 measured, TT4 measured, T4U measured, FTI measured, TBG measured, and referral. Among the 30 attributes in the dataset, six are continuous variables: age, TSH, T3, TT4, T4U, and FTI. These variables are summarized in Table 1.

Table 1. Features in the dataset with continuous values

S No.	Feature Name	Description
1	age	Patient's age
2	TSH	A hormone that stimulates the Thyroid
3	T3	Triiodothyronine
4	TT4	Thyroxine Hormone level in the blood bloodstream
5	T4U	Thyroxine Utilization Rate
6	FTI	T4 index for diagnosing Thyroid Disease

One of the attributes in the dataset contains a "?" value, which caused an inaccurate result. The feature name is "TBG," and the code fully ignores this feature during the implementation of models in the experiments. Some feature also has missing values represented by "?".

Data Preprocessing:

Several data preparation techniques were applied in the experiment, as the dataset contained missing values, noise, and outliers. Such issues render raw data unsuitable for analysis; therefore, effective preprocessing is essential to enhance the efficiency and accuracy of large datasets. The following preprocessing steps are applied to the dataset used in the experiments:

Filling the Missing Values:

In this step, the placeholder value "?" was replaced with NaN values using Python's replace () method. Converting the placeholder values to NaN marked them as missing data, making them easier to handle during subsequent processing. Without changing, the placeholder values present in the dataset might cause errors or inaccurate results during training and testing the models, but also in mathematical calculations. Table 2 shows all the features in the dataset used in this research study and their frequency.

Table 2. Frequency of features in the dataset

Feature	count	unique	Top	Frequency
Age	3772	94	59	95
Sex	3772	3	F	2480
on thyroxine	3772	2	F	3308
Query on thyroxine	3772	2	F	3722
on antithyroid medication	3772	2	F	3729
Sick	3772	2	F	3625
Pregnant	3772	2	F	3719
thyroid surgery	3772	2	F	3719
I131 treatment	3772	2	F	3713
query hypothyroid	3772	2	F	3538
query hyperthyroid	3772	2	F	3535
Lithium	3772	2	F	3754
Goitre	3772	2	F	3738
Tumor	3772	2	F	3676
Hypopituitary	3772	2	F	3771
Psych	3772	2	F	3588
TSH measured	3772	2	T	3403
TSH	3772	288	?	369
T3 measured	3772	2	T	3003
T3	3772	70	?	769
TT4 measured	3772	2	T	3541
TT4	3772	242	?	231
T4U measured	3772	2	T	3385
T4U	3772	147	?	387
FTI measured	3772	2	T	3387
FTI	3772	235	?	385
TBG measured	3772	1	F	3772
TBG	3772	1	?	3772
referral source	3772	5	Other	2201
binaryClass	3772	2	P	3481

Finding and Encoding Non-Numeric Columns:

Some of the columns in the data corpus have a non-numeric type, such as a categorical variable, so the label encoder ( ) method of Python is used to convert them to numeric values. A label encoder assigns a unique numerical value to each field in a dataset. This step is crucial because some of the machine learning models and neural networks cannot work with string or categorical values. Label encoders preserve the categorical information as well and make sure that each column has a numerical value that is appropriate for calculation.

Handling Missing Values:

To address the issue of missing values in the dataset, they were replaced with the corresponding column's mean. This technique ensures a practical approach for handling missing data during the training process. Mean imputation is an efficient and straightforward method without adding biases and preventing rows and columns from incomplete data in the dataset, which might cause errors in later steps.

Maintaining Numerical Features:

The dataset is divided into two portions: Features normally represented with X in the Python code, and the target variable represented by Y. All the features' columns are converted to Numeric values by using the two numeric method of Python. This process ensures to convert features are converted to numbers or NaN values, and this step makes the data feasible to be input to machine learning and deep learning models.

Feature selection:

The rapid growth of technology and internet applications generates massive amounts of data, and managing such vast volumes presents a significant challenge [13]. Feature selection is an effective way to remove redundant features and irrelevant features to improve model accuracy and reduce computational time. The redundant and irrelevant features lead to low performance of the models, as well as the overfitting problem. These unnecessary features seriously impact the learning process and training speed. The model's performance is determined by the feature selection technique and algorithm, and the quantity/number of features selected from the feature set [14]. The experiments employed three feature selection methods: Chi-Square, Linear Discriminant Analysis (LDA), and Recursive Feature Elimination (RFE).

Handling Class Imbalance Problem:

The distribution of the target variable in the dataset is highly imbalanced, with the majority class representing 92% of the instances. In order to ensure that our proposed models did not develop biased results toward the majority class, we implemented a two-stage strategy. First, all performance metrics (Precision, Recall, F1-Score) were prioritized over raw accuracy to provide a more realistic examination of model performance on the minority class. Second, we employed the Synthetic Minority Over-sampling Technique (SMOTE) during the training stage for the deep learning models used in the experiment (LSTM and MLP), which are notably sensitive to imbalanced data. SMOTE produces synthetic samples for the minority class to create a balanced training set. The ensemble method (Gradient Boost) is examined both with and without class weight adjustment, and its inherent robustness carried out the imbalance more efficiently without SMOTE.

Chi-Square:

A statistic-based method known as Chi-Square is used to assess and examine the relationship between the target variable and categorical variables of the dataset [15]. The Chi-Square method assesses whether features are dependent on the class label. As a non-parametric statistical technique, it helps identify relevant features and is widely applied in analytical tasks [15]. Chi-Square determines whether the occurrence of a particular class and a particular phrase is independent. Formally, the quantity for each phrase in a given document DDD is estimated, and its score is used to rank them [16]. The features are ranked by calculating the weights for

each feature based on their importance and are working for data distribution [17]. The mathematical equation for the algorithm is:

$$\chi 2 = \sum \frac{(Oij-Eij)2}{Eij} \quad \text{(i)}$$

where Oij and Eij are the observed and expected frequencies at the ith row and jth column, respectively. Eij is computed as:

$$Eij = \frac{(\text{Total Rows}) \text{ x } (\text{Total Columns})}{\text{Grand Total}} \quad \text{(ii)}$$

Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis in the study is employed as a dimensionality reduction method, rather than a feature selection method. Unlike feature selection methods like RFE and Chi-Square used in this research study, which select a subset of the original features of a dataset, LDA creates a new, smaller set of features that are linear combinations of the original input features. LDA aims to project the data onto a lower-dimensional space that improves the separability between the binary classes. In our experiments, the dataset is transformed using LDA, and the resulting components are used to train and evaluate the models. This proposed approach of LDA is similar to Principal Component Analysis (PCA) and is supervised, as it uses class labels to define the axes of maximum discrimination. Within the class scatter matrix is Sw, and within the class matrix, Sb, are calculated as:

$$S_b = \sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T \quad \text{(iii)}$$

$$S_w = \sum_{i=1}^{C} \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T \quad \text{(iv)}$$

Recursive Feature Elimination (RFE):

Recursive Feature Elimination (RFE) helps to identify the most relevant and important features for machine learning models and eliminate the less important features iteratively. It evaluates model performance across different feature subsets using classification methods [18]. Although RFE aims to discard weak features, some of these may become useful when combined with others. The algorithm leverages the generalization capability of Support Vector Machines (SVM) to enhance feature selection [19]. The following are some equations that RFE uses for feature selection and feature ranking.

$$\hat{y} = \sum_{j=1}^{p} w_j x_j + b \quad \text{(v)}$$

Where Wj represents the weight of the i-th feature in the dataset. In RFE, the features are based on their importance score:

$$R(j) = \text{importance}(x_j) \quad \text{(vi)}$$

Where R(j) is the rank of the feature. Some features are removed based on the lowest importance score they achieve; for this, the RFE uses the equation.

$$X_{\text{new}} = X_{\text{old}} \setminus \{x_{\text{least important}}\} \quad \text{(vii)}$$

Models used in the Experiments:

Recurrent Neural Network (RNN):

The concept of RNN dates back to 1970, when Werbos introduced the concept of backpropagation through time (BPTT), which became the foundation for RNN [20]. A Recurrent Neural Network (RNN) is a deep learning model designed to process sequential or time-series data by learning patterns that evolve. Similar to other neural networks, an RNN is composed of processing units called neurons, which are organized into layers [21]. Because of high-dimensional hidden vectors and non-linear dynamics, RNNs can store and process previous information [22]. The architecture of an RNN typically consists of three layers: the input layer, the hidden layer, and the output layer. Each layer contains neurons, with recurrent connections that enable information to be iteratively passed through the network. In the hidden layer, perform the calculation by combining the current input and the previous hidden state. The choice of activation function also affects the RNN; activation functions are Sigmoid, ReLu, Identity, and Tanh [20]. Figure 2 illustrates the internal structure of an RNN, showing

how information flows sequentially from one layer to the next. During backpropagation, the network traces errors backward through previous steps before generating the final output.

The RNN used in this research experiment consists of an LSTM input layer with 64 neurons and with Tanh activation function, to handle overfitting, a dropout layer with a rate of 0.3 is added, a fully connected dense layer with 32 neurons and ReLu activation function, a second dropout layer with a rate of 0.3 is also added. And a final dense output layer with only one neuron and a Sigmoid activation function for binary classification is added. The RNN is compiled with Adam Optimizer, trained with Binary Cross-Entropy as a loss function and accuracy as the evaluation metric. It is trained with 100 epochs, and a batch size of 32 is selected.

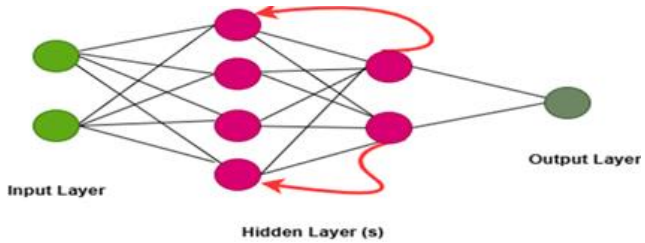

Figure 2. Working Mechanism of Recurrent Neural Network (RNN)

Gradient Boost (GB):

Boosting algorithms are ensemble methods that combine weak learners, such as decision trees, to build a stronger and more accurate predictive model. While weak learners perform poorly on their own, boosting enhances their proficiency by iteratively improving their performance [23]. Gradient boost is a boosting algorithm used for classification and regression problems by finding a proximation [24]. GB represents a decision tree for large and complicated datasets and combines a weak prediction model to perform well. It arranges the model in such a way that the next model learns from the error or loss function of the previous model [25]. Gradient Boosting is a powerful ensemble method available in several variants, including AdaBoost, XGBoost, and LightGBM, among others [26]. and is used in many applications like multi-class classification, ranking, and in click prediction [27].

Gradient Boosting is highly adaptable and can be tailored to specific data-related tasks. It allows flexibility in designing and selecting loss functions, making boosting algorithms easy to implement and experiment with across different models [26]. Figure 3 depicts the internal working mechanism of Gradient Boost, where multiple weak learner algorithms are sequentially trained to reduce the prediction error.

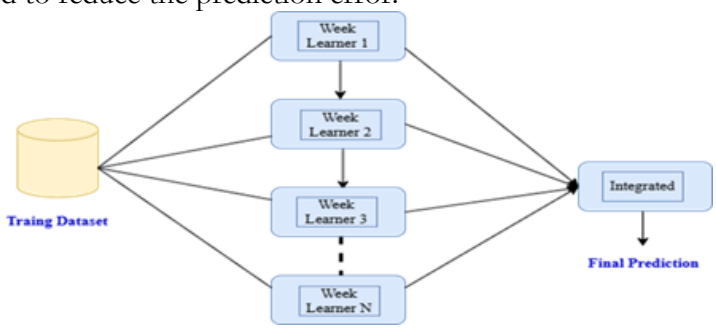

Figure 3. Working Mechanism of the Ensemble Method: Gradient Boost

The model Gradient Boost is used in experiments with Max-Depth 3, the number of weak learners (n_ n_estimator) is 100, with a learning rate of 0.8, and the subsample size is 0.8 to improve randomness and improve robustness.

Multi-Layer Perceptron (MLP)Artificial Neural Networks are inspired by the living organisms' neuron system. They are widely used to address complex problems, particularly in situations where statistical mapping is required to transform input data into the desired output [28]. MLP is a feedforward neural network that converts a collection of input data into output

and is the modified version of the standard linear perceptron with layers of neurons and a nonlinear activation function. MLP consists of multiple layers of neurons; the complexity of the network changes when the number of layers or the number of neurons in each will changes. MLP at least has three layers of neurons: these are, input layer, one or more hidden layers, and the output layer [29]. The output layer of the model has a single neuron that presents the output of the model. MLP calculates the value of neurons at the current layer with the help of an activation function. The backward propagation method first assigns random weights to the neurons' connections, which are then modified. Figure 5 shows the internal architecture of an MLP, which is made up of several layers of linked neurons that move input data from one layer to the next before producing an output.

The MLP used in Thyroid disease detection experiments is composed of two hidden layers with 128 neurons in the first layer, while the 2nd layer consists of 64 neurons. The ReLu activation function is used with each hidden layer, and the dropout rate is 0.3 used with each hidden layer for regularization. The MLP is compiled with Adam Optimizer, 100 epochs, batch size is 40, and binary cross-entropy for binary classification.

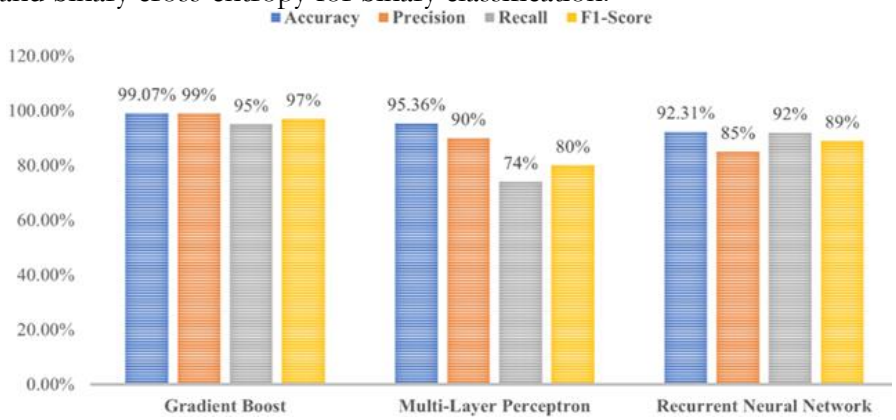

Figure 4. Experimental results of GB, RNN, MLP with all features

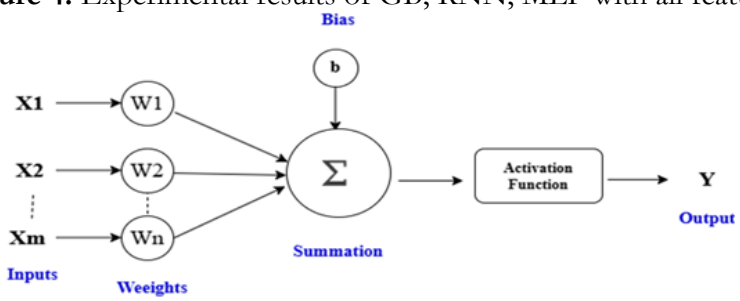

Figure 5. Working mechanism of Multi-Layer Perceptron (Input, Hidden, and Output Layers)

Results and Discussion:

In this study, we employed three well-known algorithms: two deep learning models, RNN and MLP, and Gradient Boost, an ensemble method. These models were used to illustrate different perspectives discussed in the table. The experiments were conducted twice: first using all features in the dataset, and then using feature selection methods such as Chi-Square, LDA, and RFE to extract the most relevant features, after which the experiments were repeated with this reduced feature set.

Experiments with all Features:

The dataset used in the experiments consists of 30 features, and RNN, MLP, and GB are trained and tested with all features. The accuracy of all these models is satisfactory, where GB performs outclass with the accuracy of 99.07%, the accuracy of MLP is 95.36%, and finally RNN generates 92.31% of accuracy. Table 3 summarizes the performance outcomes of three models, GB, MLP, and RNN, using all features in the dataset.

Table 3. Performance comparison of GB, MLP, and RNN with all features

Model Name	Accuracy	Precision	Recall	F1-Score
GB	99.07%	99%	95%	97%
MLP	95.36%	90%	74%	80%
RNN	92.31%	85%	92%	89%

Gradient Boost achieved 99% precision, 95% recall, and a 97% F1-score. The Multi-Layer Perceptron (MLP) achieved 90% precision, 74% recall, and an 80% F1-score, while the Recurrent Neural Network (RNN) achieved 85% precision, 92% recall, and an 89% F1-score, respectively. In the trial with every feature, the Gradient Boost ensemble approach performs admirably. All of the model's performance evaluations are shown in the table 3. Figure 4 represents the comparison of evaluation matrices, Accuracy, Precision, Recall, and F1-Score for the three models used in this study: Gradient Boost, Recurrent Neural Network, and Multi-Layer Perceptron, as well as the results generated by these models using all features of the dataset.

Experiments with the Recursive Feature Elimination (RFE) feature selection technique:

To extract the most prominent features from the dataset used in our experiments for thyroid disease detection, we implement the Recursive Feature Elimination (RFE) technique. By eliminating irrelevant or redundant features, this feature selection method improves computational efficiency and mitigates overfitting, thereby enhancing overall model performance. RFE select age, sex, on thyroxine, TSH measured, TSH, T3, TT4, T4U, FTI, and referral source, and ignore the other features of the dataset.

Since thyroid disease is often influenced by factors such as age and hormone levels (FTI, T3, TT4, TSH, and T4U), these features were selected based on medical relevance. Choosing the most appropriate features and removing the redundant features from the dataset shows the improved models. Figure 6 illustrates the comparison of evaluation matrices, Accuracy, Precision, Recall, and F1-Score, for GB, MLP, and RNN after applying the Recursive Feature Elimination method, and shows how the feature selection technique impacts the performance of models.

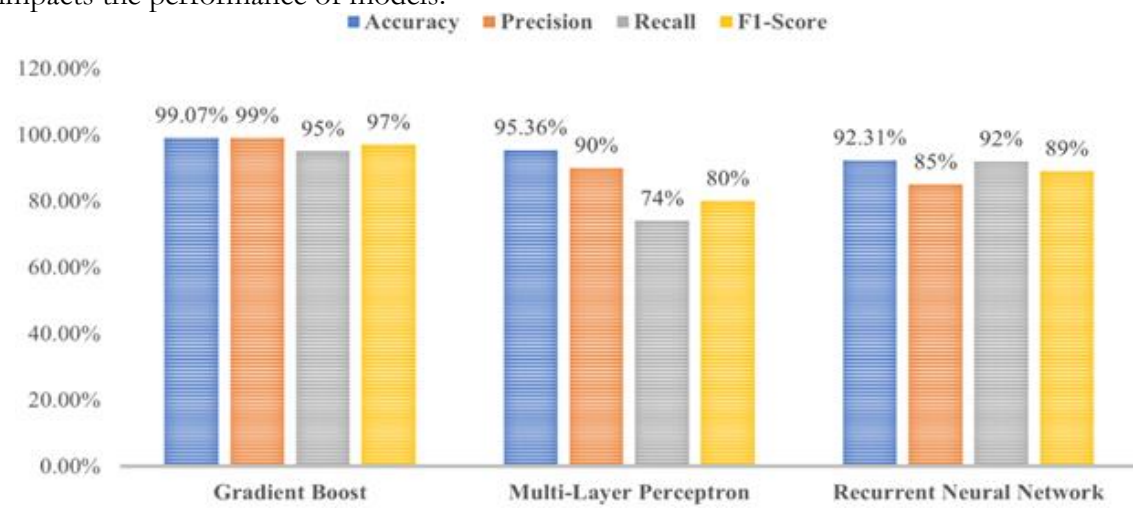

Figure 6. Experimental results of GB, RNN, and MLP with the RFE feature selection method

Table 4. Performance comparison of GB, MLP, and RNN using the RFE method

Model Name	Accuracy	Precision	Recall	F1-Score
GB	99%	96%	88%	92%
MLP	95%	96%	98%	97%
RNN	94.54%	94.54%	99.86%	97.14%

Experiments using Gradient Boost combined with RFE achieved an accuracy of 99%, precision of 96%, recall of 88%, and an F1-score of 92%; however, the results indicate only a marginal improvement. The use of RFE with the Recurrent Neural Network (RNN) enhanced model performance, yielding an accuracy of 94.57%, precision of 94.54%, recall of 99.86%, and an F1-score of 97.14%. The result of MLP with RFE is: accuracy 95%, Precision 96%, recall 98%, and F1-score 97%. Table 4 depicts the performance evaluation of the three models: GB, MLP, and RNN, and applying the RFE feature selection method.

Experiments with the Chi-Square Feature Selection Technique:

For thyroid disease detection, we employed the Chi-Square feature selection method to pinpoint the features most statistically relevant for accurate classification. Chi-square is a well-known statistical test for feature selection since it measures the association between features and the target variable, especially when the data is categorical. After applying the chi-square tool, it just selects the features sex, on thyroxine, pregnant, query hypothyroid, psych, TSH measured, TSH, T3, TT4, and FTI, and discards the rest of the features. The selection of these features shows that they are statistically associated with thyroid disease. Figure 7 illustrates the Accuracy, Precision, Recall, and F1-score for GB, RNN, and MLP following the application of the feature selection method Chi-Square, which reduces features statistically and influences the reliability and accuracy.

Using Chi-Square with Gradient Boost resulted in high metrics—99% accuracy, precision, recall, and F1-score—but showed only marginal improvement. Better performance was observed with Chi-Square combined with Recurrent Neural Networks, where the RNN achieved 92% accuracy, 92% precision, 100% recall, and a 96% F1-score. The Multi-Layer Perceptron (MLP) with Chi-Square achieved 94% accuracy, 96% precision, 98% recall, and a 97% F1-score. Table 5 represents the evaluation performance of models GB, MLP, and RNN, and uses Chi-Square as a feature selection method.

Table 5. Performance comparison of GB, MLP, and RNN using the Chi-square method

Model Name	Accuracy	Precision	Recall	F1-Score
GB (Gradient Boosting)	99%	96%	88%	92%
MLP (Multilayer Perceptron)	95%	96%	98%	97%
RNN (Recurrent Neural Network)	94.54%	94.54%	99.86%	97.14%

Experiments with the Linear Discriminant Analysis (LDA) feature selection technique:

Linear Discriminant Analysis (LDA) is largely used for dimensionality reduction rather than a feature selection technique like RFE and Chi-Square. Instead of selecting an appropriate feature, LDA converts the dataset into lower lower-dimensional space. The newly transformed feature space is a combination of the existing features; therefore, LDA does not select individual features or discard others. Figure 8 represents the Accuracy, Precision, Recall, and F1-Score for GB, RNN, and MLP after Linear Discriminant Analysis method for dimensionality reduction, which converts the features into low low-dimensional space to maximize class separability to maintain model performance.

When the experiments were conducted by using LDA, the Gradient Boost algorithm produced 94% accuracy, 95%, 98%, and the F1-Score with 97%. The MLP generates an accuracy of 92%, Precision 92%, Recall 100%, and F1-Score 96%, and the Recurrent Neural Network with an accuracy rate of 94.64%, Precision with 95%, Recall 98%, and F1-Score produces 97%. Table6 represents the evaluation performance of models GB, MLP, and RNN, and LDA as feature selection methods.

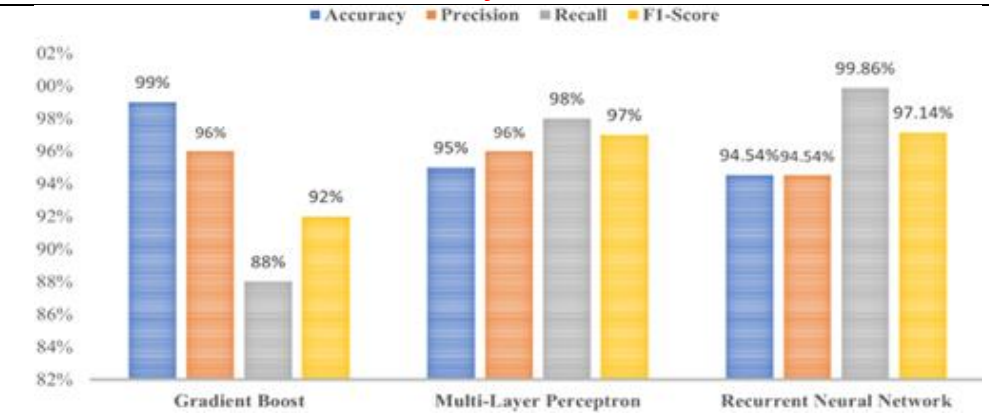

Figure 7. Experimental results of GB, RNN, and MLP with the Chi-Square feature selection method

Table 6. Performance comparison of GB, MLP, and LDA feature selection methods

Model Name	Accuracy	Precision	Recall	F1-Score
GB	94%	95%	98%	97%
MLP	92%	92%	100%	96%
RNN	94.64%	95%	98%	97%

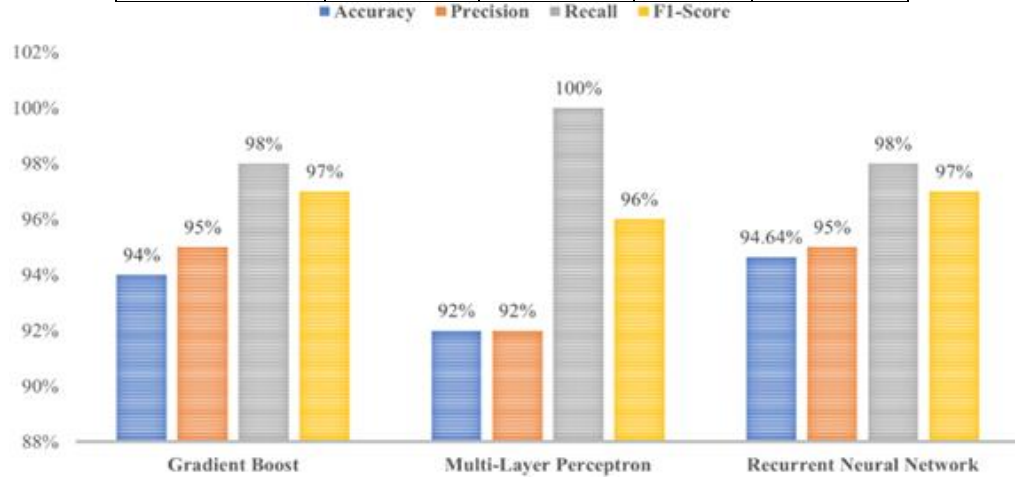

Figure 8. Experimental results of GB, RNN, and MLP with the LDA feature selection method

Discussion:

Our research study encompasses a comprehensive examination and comparative use of two feature selection methods, like RFE and Chi-Square, and the dimensionality reduction algorithm LDA, with the combination of MLP, GB, and RNN. Our study offers a straightforward statistical comparison of filter (Chi-Square), a rapper (RFE), and an embedded (LDA) technique within the same experimental setup.

Statistical Significance Analysis

We perform a statistical significance test to go beyond a descriptive comparison and statistically evaluate our results. The predictions of the two top-performing configurations the LSTM model optimized using the RFE feature selection technique and the default Gradient Boost model were subjected to a McNemar's test. Their binary predictions on the test set were subjected to the test. The statistical significance of the variance in classification performance between these two models is indicated by the resulting p-value, which was $p < 0.05$. This result statistically demonstrates that although both models attain high accuracy, the LSTM+RFE model's greater recall is a result of its significantly altered prediction patterns and distinct error profiles.

Table 7. Propose research work with features

Study / Method	Dataset & Features	Feature Selection	Model(s) Used	Best Accuracy	Key Contribution / Limitation
Our Proposed research work	UCI Thyroid dataset with 30 features	RFE, Chi-Square, LDA	GB, MLP, RNN	**99%**	Combines two feature selection and a dimensionality reduction method with multiple ML models to optimize accuracy and reduce complexity.
[3]	UCI Thyroid dataset	L1/L2 regularization	Naïve Bayes, Logistic Regression	100%	Achieved perfect accuracy with simpler models; lacks deep learning integration.
[1]	UCI dataset (sampled/unsampled)	Nill	Random Forest	94.8%	Focused on data sampling effects; lower accuracy than our proposed work
[30]	Multi-class thyroid data	Nill	XGBoost	99%	Strong boosting model performance, but no feature selection analysis.
[11]	UCI Thyroid dataset	Hybrid optimization	Type-2 Fuzzy SVM	99%	Uses fuzzy SVM with hybrid optimization; computationally complex.
[12]	UCI Thyroid dataset	Filter-based selection	Stacked Ensemble	99.9%	Highest reported accuracy; limited evaluation of deep neural models.
[7]	Medical imaging (Ultrasound/CT)	Nill	Deep CNN	94–97%	Image-based detection requires expensive imaging data.
[10]	Indian thyroid dataset	Nill	XGBoost, Logistic Regression	98.5%	Demonstrated robust boosting performance without feature optimization.
Our Proposed research work	UCI Thyroid dataset with 30 features	RFE, Chi-Square, LDA	GB, MLP, RNN	**99%**	Combines two feature selection and a dimensionality reduction method with multiple ML models to optimize accuracy and reduce complexity.

Conclusion and Future Plan:

In this research study, we proposed an effective and accurate Machine learning based detection system for thyroid disease by using well-known machine learning models like Gradient Boost, Recurrent Neural Network, and Multi-Layer Perceptron. Additionally, we utilized feature selection algorithms like chi-square, Redundant Feature Elimination, and Linear Discriminant Analysis. The finding demonstrates that the selected features by using the mentioned method greatly affect the results, and hence, Gradient Boost achieves the highest accuracy of 99% and shows dominance in the detection process. This research will help healthcare professionals in the timely and early detection of dangerous thyroid disease and will help them in quick decision-making to decrease the chance of the situation worsening. We admit that this research study possesses some limitations as well. The big problem is the imbalanced dataset, which might affect the accuracy of the models used. The second problem is the dataset size and sample size, as the size of the dataset is limited to a few thousand cases. If we improve the size of the dataset, algorithms will perform well. Another big problem is the missing values in the dataset. In the future, we will improve the dataset and will try to remove or reduce the missing values.

References:

[1] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "[Retracted] Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," *Biomed Res. Int.*, vol. 2022, no. 1, p. 9809932, Jan. 2022, doi: 10.1155/2022/9809932.

[2] M. P. J. Vanderpump, "The epidemiology of thyroid disease," *Br. Med. Bull.*, vol. 99, no. 1, pp. 39–51, Sep. 2011, doi: 10.1093/BMB/LDR030.

[3] H. Abbad Ur Rehman, C. Y. Lin, Z. Mushtaq, and S. F. Su, "Performance Analysis of Machine Learning Algorithms for Thyroid Disease," *Arab. J. Sci. Eng.*, vol. 46, no. 10, pp. 9437–9449, Oct. 2021, doi: 10.1007/S13369-020-05206-X/TABLES/4.

[4] A. Sultana and R. Islam, "Machine learning framework with feature selection approaches for thyroid disease classification and associated risk factors identification," *J. Electr. Syst. Inf. Technol. 2023 101*, vol. 10, no. 1, pp. 1–23, Jun. 2023, doi: 10.1186/S43067-023-00101-5.

[5] S. Verma, R. Popli, H. Kumar, and A. Srivastava, "Classification of thyroid diseases using machine learning frameworks," *Int. J. Health Sci. (Qassim).*, vol. 6, no. S1, pp. 7552–7566, Apr. 2022, doi: 10.53730/IJHS.V6NS1.6603.

[6] F. Monaco, "Classification of thyroid diseases: suggestions for a revision," *J. Clin. Endocrinol. Metab.*, vol. 88, no. 4, pp. 1428–1432, Apr. 2003, doi: 10.1210/JC.2002-021260.

[7] X. Zhang, V. C. Lee, J. Rong, J. C. Lee, and F. Liu, "Deep convolutional neural networks in thyroid disease detection: A multi-classification comparison by ultrasonography and computed tomography," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, doi: 10.1016/J.CMPB.2022.106823.

[8] X. Zhang, V. C. S. Lee, J. Rong, J. C. Lee, J. Song, and F. Liu, "A multi-channel deep convolutional neural network for multi-classifying thyroid diseases," *Comput. Biol. Med.*, vol. 148, p. 105961, Sep. 2022, doi: 10.1016/J.COMPBIOMED.2022.105961.

[9] L. Aversano, M. L. Bernardi, M. Cimitile, A. Maiellaro, and R. Pecori, "A systematic review on artificial intelligence techniques for detecting thyroid diseases," *PeerJ Comput. Sci.*, vol. 9, p. e1394, Jun. 2023, doi: 10.7717/PEERJ-CS.1394.

[10] S. Sankar, A. Potti, G. Naga Chandrika, and S. Ramasubbareddy, "Thyroid Disease Prediction Using XGBoost Algorithms," *J. Mob. Multimed.*, vol. 18, no. 3, pp. 917–934, Feb. 2022, doi: 10.13052/JMM1550-4646.18322.

[11] V. Sureshkumar, S. Balasubramaniam, V. Ravi, and A. Arunachalam, "A hybrid

optimization algorithm-based feature selection for thyroid disease classifier with rough type-2 fuzzy support vector machine," *Expert Syst.*, vol. 39, no. 1, Jan. 2022, doi: 10.1111/EXSY.12811.

[12] G. Obaido *et al.*, "An Improved Framework for Detecting Thyroid Disease Using Filter-Based Feature Selection and Stacking Ensemble," *IEEE Access*, vol. 12, pp. 89098–89112, 2024, doi: 10.1109/ACCESS.2024.3418974.

[13] M. A. Hall, "Correlation-based feature selection for machine learning," 1999. Accessed: Oct. 10, 2025. [Online]. Available: https://hdl.handle.net/10289/15043

[14] A. Sharma and S. Dey, "A comparative study of selection and machine learning techniques for sentiment analysis," *Proceeding 2012 ACM Res. Appl. Comput. Symp. RACS 2012*, pp. 1–7, 2012, doi: 10.1145/2401603.2401605.

[15] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2018-November, pp. 160–163, Jul. 2018, doi: 10.1109/ICSESS.2018.8663882.

[16] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-Square and PCA Based Feature Selection for Diabetes Detection with Ensemble Classifier," *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1931–1949, Jan. 2023, doi: 10.32604/IASC.2023.028257.

[17] T. Almutiri and F. Saeed, "Chi Square and Support Vector Machine with Recursive Feature Elimination for Gene Expression Data Classification," *2019 1st Int. Conf. Intell. Comput. Eng. Towar. Intell. Solut. Dev. Empower. our Soc. ICOICE 2019*, Dec. 2019, doi: 10.1109/ICOICE48418.2019.9035165.

[18] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," *J. Sens. Actuator Networks 2023, Vol. 12, Page 67*, vol. 12, no. 5, p. 67, Sep. 2023, doi: 10.3390/JSAN12050067.

[19] X. -w. C. and J. C. Jeong, "Enhanced recursive feature elimination," *Sixth Int. Conf. Mach. Learn. Appl. (ICMLA 2007), Cincinnati, OH, USA*, pp. 429–435, 2007, doi: 10.1109/ICMLA.2007.35.

[20] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Inf. 2024, Vol. 15, Page 517*, vol. 15, no. 9, p. 517, Aug. 2024, doi: 10.3390/INFO15090517.

[21] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent Advances in Recurrent Neural Networks," Dec. 2017, Accessed: Oct. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1801.01078

[22] E. Alonso, B. Moysset, and R. Messina, "Adversarial generation of handwritten text images conditioned on sequences," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 481–486, Sep. 2019, doi: 10.1109/ICDAR.2019.00083.

[23] A. Beygelzimer, E. Hazan, S. Kale, and H. Luo, "Online Gradient Boosting," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[24] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/S10462-020-09896-5/TABLES/12.

[25] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," *J. Pet. Sci. Eng.*, vol. 208, p. 109244, Jan. 2022, doi: 10.1016/J.PETROL.2021.109244.

[26] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. DEC, p. 63623, Dec. 2013, doi: 10.3389/FNBOT.2013.00021/BIBTEX.

[27] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv.*

Neural Inf. Process. Syst., vol. 30, 2017, Accessed: Oct. 10, 2025. [Online]. Available: https://github.com/Microsoft/LightGBM.

[28] E. Wilson and D. W. Tufts, "Multilayer perceptron design algorithm," *Neural Networks Signal Process. - Proc. IEEE Work.*, pp. 61–68, 1994, doi: 10.1109/NNSP.1994.366063.

[29] V. A. Golovko, "Deep learning: an overview and main paradigms," *Opt. Mem. Neural Networks (Information Opt.*, vol. 26, no. 1, pp. 1–17, Jan. 2017, doi: 10.3103/S1060992X16040081/METRICS.

[30] M. Alnaggar, M. Handosa, T. Medhat, and M. Z. Rashad, "Thyroid Disease Multi-class Classification based on Optimized Gradient Boosting Model," *Egypt. J. Artif. Intell.*, vol. 2, no. 1, pp. 1–14, Apr. 2023, doi: 10.21608/EJAI.2023.205554.1008.