Eurasian Academy of Sciences
Eurasian Econometrics, Statistics & Empirical Economics Journal
2026 Volume:27 S: 1- 20
Published Online May 2026 (http://econstat.eurasianacademy.org)
http://dx.doi.org/10.17740/eas.stat.2026-V27-01

# DIAGNOSIS OF THYROID CANCER WITH MACHINE LEARNING ALGORTIHMS[1]

**İrem TURAN*** **Çiğdem ARICIGİL ÇİLAN****

*Independent Researcher, ireemturan@hotmail.com, ORCID ID: 0000-0002-7794-147X
**Prof. Dr. Istanbul University, Faculty of Business Administration, ccilan@istanbul.edu.tr, ORCID ID: 0000-0002-7862-7028

Received Date:05.02.2026 Accepted Date:04.05.2026

Copyright © 2026 İrem TURAN, Çiğdem ARICIGİL ÇİLAN. This is an open access article distributed under the Eurasian Academy of Sciences License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## ABSTRACT

In recent years, thyroid cancer has become the second most common type of cancer worldwide after breast cancer. Today, with the widespread use of fast computers and the development of Artificial Intelligence algorithms, early diagnosis and accurate treatment have improved the quality of life for patients and increased survival rates. Traditional methods for diagnosing thyroid cancer can have limited accuracy, leading to misdiagnosis. In medical diagnostics, Machine Learning, a subfield of Artificial Intelligence, is frequently used. This method is widely applied in the diagnosis of diseases. Machine Learning algorithms help analyze large amounts of data to provide accurate and precise diagnoses. In this study, a dataset related to thyroid disease was used to predict whether thyroid nodules are malignant or benign as a classification problem in Machine Learning, without the need for biopsies of the patients. The available dataset was trained using classification algorithms under Supervised Learning, and each algorithm was used to predict models with Stratified Train Test Split and Stratified K-Fold Cross Validation methods. These models were later used to make predictions on the current test dataset. The classification algorithms employed in the study included Linear Discriminant Analysis, Naive Bayes, K-Nearest Neighbors, Random Forest, Support Vector Machines, Decision Trees, Artificial Neural Networks, Logistic Regression, Adaboost, XGboost, Gradient Boost, LightGBM Boost, and Catboost, which were implemented using Python. When comparing the performance metrics obtained from the results, it was found that Naive Bayes algorithm achieved the highest accuracy value with 92.31%, while XGboost algorithm had the highest ROC/AUC value with 92%.

**Keywords:** Machine Learning, Thyroid Disease, Supervised Learning Algorithms, Naive Bayes
**JEL-Classification:** C45, C53, I10

## 1. Introduction

Technological advancements have played a significant role in the healthcare sector in recent years. Rapid improvements in computational power, the growth of large-scale datasets, and the development of algorithms capable of learning from data have enabled artificial intelligence (AI) systems to match and, in some cases, exceed human performance in specific medical tasks. However, the success of AI applications in healthcare depends not only on the algorithms employed but also on the proper preparation, integration, and analysis of clinical data. Maranghi et al. (2022) emphasized that data preparation and data analytics processes play a critical role in supporting clinical decision-making mechanisms in AI applications based on real-world patient data. Furthermore, recent studies have highlighted the growing role of

[1]This article is derived from İrem Turan's master's thesis, supervised by Çiğdem Arıcıgil Çilan, titled "Thyroid Cancer Diagnosis Using Machine Learning Algorithms".

artificial intelligence in supporting diagnostic and clinical decision-making processes across a wide range of medical specialties (Zhao et al., 2026).

In the field of thyroid cancer diagnosis, artificial intelligence-based systems have demonstrated promising results in the assessment of thyroid nodules, cervical lymph nodes, and cytological specimens, with several ultrasound-based platforms already receiving regulatory approval for clinical use (Johnson & Tessler, 2026).

According to recent epidemiological reports, thyroid cancer is among the most frequently diagnosed endocrine malignancies worldwide and represents a growing public health concern due to its increasing incidence rates (Pizzato et al., 2022). Early diagnosis is particularly important because timely detection significantly improves treatment outcomes and patient survival. Thyroid cancer diagnosis is of particular importance because accurate and timely diagnosis directly influences treatment planning and prognosis.

Conventional diagnostic approaches include ultrasonography, fine-needle aspiration biopsy, blood tests, and imaging techniques (Chakrabarty et al., 2024; Fei et al., 2024; Osseis et al., 2023). Although these methods are widely used in clinical practice, they have certain limitations and may not always provide sufficiently accurate diagnostic results. Recent advances in imaging-based diagnostic systems and machine learning-assisted ultrasound interpretation have further improved the diagnostic evaluation of thyroid nodules (Chakrabarty et al., 2024; Gatta et al., 2025).

Machine learning (ML), a subfield of artificial intelligence, has become an important tool in medical diagnosis by enabling the analysis of large volumes of data and generating accurate diagnostic predictions. In recent years, machine learning algorithms have increasingly been applied to distinguish between benign and malignant thyroid nodules, with the ultimate goal of improving the early detection of thyroid cancer. Previous studies have demonstrated that machine learning techniques can provide high diagnostic performance and support clinical decision-making in thyroid disease diagnosis (Lee et al., 2022; Xi et al., 2022; Kumar et al., 2023). Recent systematic reviews and meta-analyses have further confirmed the effectiveness of artificial intelligence and machine learning models for thyroid nodule classification, reporting high diagnostic accuracy, sensitivity, and specificity across multiple clinical settings (Gatta et al., 2025; Jassal et al., 2025). In addition, growing evidence suggests that AI-assisted diagnostic systems can improve ultrasound-based risk stratification and support clinicians in reducing unnecessary invasive procedures while maintaining diagnostic reliability (Edström et al., 2025; Ni et al., 2025).

By analyzing data derived from imaging studies, ultrasonographic examinations, laboratory tests, and biopsy findings, machine learning models can assist in identifying malignant lesions and potentially contribute to disease staging. Furthermore, these algorithms can classify different thyroid cancer subtypes based on predefined characteristics and support clinicians in selecting appropriate treatment strategies. The increasing life expectancy of the global population has also led to a growing burden of chronic disease management on healthcare systems (Pizzato et al., 2022). Consequently, the identification of high-risk patients, timely diagnostic evaluation, and effective treatment planning have become essential for maintaining cost-effective healthcare services. In this context, machine learning-based decision-support systems offer promising opportunities to improve diagnostic efficiency, optimize healthcare resource utilization, and support personalized medicine approaches.

Overall, machine learning algorithms represent a valuable tool for enhancing the accuracy of thyroid cancer diagnosis and supporting improved patient outcomes. Therefore, this study aims to investigate the application of supervised machine learning algorithms in

thyroid cancer diagnosis, with particular emphasis on their predictive performance, effectiveness, and practical applicability in a clinical setting.

## 2. Literature Review

Machine learning (ML) techniques have been increasingly applied to the diagnosis and prediction of thyroid diseases. Previous studies have demonstrated that supervised learning algorithms can achieve high diagnostic performance when combined with appropriate feature selection and data preprocessing techniques.

Riajulislam et al. (2021) investigated the early detection of hypothyroidism using machine learning and feature selection methods, including Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Support Vector Machines (SVM), Decision Trees, Random Forest, Logistic Regression, and Naïve Bayes. Their results indicated that SVM combined with RFE achieved the highest accuracy of 99.35%. Similarly, Li et al. (2012) proposed a computer-aided diagnosis system based on PCA and Extreme Learning Machines (ELM), reporting an accuracy of 98.1% for thyroid disease classification. Ma et al. (2019) utilized convolutional neural networks on SPECT images and demonstrated that machine learning-based approaches outperformed traditional diagnostic methods. More recently, Zhao et al. (2026) highlighted the growing role of artificial intelligence-based diagnostic systems in thyroid disease classification and clinical decision support. Yao et al. (2025) developed a multimodal generative artificial intelligence model (ThyGPT) for thyroid nodule risk assessment and management, demonstrating the growing potential of transparent and interpretable AI systems in clinical decision support.

Several studies have focused specifically on thyroid cancer prediction. Xi et al. (2022) evaluated six machine learning models, including Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Random Forest, and Gradient Boosting techniques, for malignancy prediction in thyroid nodules. Lee et al. (2022) reviewed machine learning applications in thyroid disease diagnosis and reported accuracy values ranging from 64.3% to 99.5%, depending on the dataset and algorithm used. More recently, Kumar et al. (2023) applied machine learning and feature selection techniques to thyroid cancer data obtained from the SEER database, highlighting the importance of data-driven decision support systems in clinical practice. Alsaadawi (2023) investigated thyroid disease classification using traditional machine learning algorithms and Recursive Feature Elimination (RFE), reporting accuracy values above 97% for selected models. The study highlighted the importance of feature selection in improving predictive performance.

Recent systematic reviews and meta-analyses have further confirmed the effectiveness of machine learning algorithms in thyroid cancer diagnosis. Gatta et al. (2025) reported that machine learning-based approaches can achieve high diagnostic performance in differentiating benign and malignant thyroid nodules. Similarly, Jassal et al. (2025) emphasized the growing importance of artificial intelligence-assisted diagnostic systems and explainable artificial intelligence techniques for thyroid nodule evaluation. In addition, Edström et al. (2025) demonstrated that human–AI collaboration improved ultrasound-based thyroid nodule assessment, while Ni et al. (2025) reported that artificial intelligence-assisted management systems may contribute to reducing unnecessary biopsies and surgical procedures.

Although previous studies have reported promising results, direct comparisons of multiple supervised learning algorithms on real-world thyroid nodule datasets remain limited. Furthermore, many studies focus on a small subset of algorithms or rely primarily on imaging data. Therefore, this study aims to compare thirteen supervised machine learning algorithms using clinical, laboratory, ultrasonographic, and biopsy-related variables collected from

patients evaluated at a tertiary healthcare institution. The findings may contribute to the development of reliable decision-support systems for thyroid cancer diagnosis and help reduce unnecessary invasive procedures.

## 3. MATERIAL AND METHODS

### 3.1. Research Objective and Significance

The primary objective of this study is to investigate whether supervised machine learning classification algorithms can effectively classify thyroid cancer cases and support the diagnostic process.

Evaluating the effectiveness of supervised machine learning algorithms in thyroid cancer diagnosis is of considerable importance for improving early detection and treatment outcomes. Thyroid cancer is one of the malignancies with a high treatment success rate when detected at an early stage. However, the interpretation of biopsy findings, ultrasonographic characteristics, and laboratory test results often requires specialized clinical expertise and may lead to variations in diagnostic decisions. By learning patterns from historical patient data, supervised learning algorithms can classify benign and malignant thyroid nodules with high accuracy, thereby supporting clinicians in their decision-making processes. Such systems may contribute to reducing diagnostic errors and minimizing unnecessary surgical interventions.

Another important contribution of this research is its potential to support the development of artificial intelligence-based clinical decision support systems. Comparing the performance of classification algorithms such as Decision Trees, Support Vector Machines, Random Forests, Artificial Neural Networks, and ensemble learning methods enables the identification of the most effective approaches for thyroid cancer diagnosis. The findings of this study may facilitate the development of rapid, reliable, and cost-effective diagnostic systems for clinical practice. Furthermore, the integration of machine learning and big data technologies into healthcare has the potential to advance personalized medicine and improve the overall quality of patient care.

### 3.2. Dataset

This study utilized a dataset obtained from Bakırköy Dr. Sadi Konuk Training and Research Hospital following approval from the institutional ethics committee (Approval No: 2023/189). The initial dataset consisted of 1,994 patient records, including thyroid ultrasonography findings, thyroid-stimulating hormone (TSH) measurements, and fine-needle aspiration biopsy results.

During the data preprocessing stage, clinical reports, laboratory findings, ultrasonographic evaluations, and biopsy records were reviewed with the assistance of an endocrinology specialist. Relevant clinical information was extracted and transformed into numerical and categorical variables suitable for machine learning applications.

After applying the inclusion criteria and performing data preprocessing procedures, a substantial number of records were excluded due to missing laboratory results, incomplete ultrasonographic findings, unavailable biopsy information, or inconsistencies across clinical records. Consequently, 193 complete cases with no missing observations were retained for the final analysis and model development.

Although this process substantially reduced the sample size, it ensured that all observations used for model development were complete and clinically reliable, eliminating the need for missing-data imputation procedures.

The final dataset consisted of 193 patients and nine variables. Eight variables (age, gender, halo presence, border characteristics, number of nodules, micro/macrocalcification, nodule composition, and thyroid-stimulating hormone [TSH] level) were used as predictor variables, while Bethesda classification served as the target variable. For the purposes of binary classification, Bethesda II cases were labeled as benign, whereas Bethesda V and VI cases were labeled as malignant.

The dataset included 163 benign cases (84.5%) and 30 malignant cases (15.5%), indicating a substantial class imbalance. To preserve the original class distribution during model training and evaluation, Stratified Train-Test Split and Stratified K-Fold Cross-Validation techniques were employed throughout the study.

### 3.3. Methods

Thirteen supervised machine learning algorithms were evaluated in this study, including Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Random Forest (RF), AdaBoost, XGBoost, Gradient Boosting (GB), LightGBM, and CatBoost.

Prior to model development, categorical variables were transformed into machine-readable formats, and the dataset was reviewed for consistency and completeness. Since the dataset exhibited class imbalance, stratified sampling techniques were employed to preserve the distribution of benign and malignant cases during model training and evaluation.

Model performance was assessed using two validation strategies: Stratified Train-Test Split (80% training, 20% testing) and Stratified K-Fold Cross-Validation. The Stratified Train-Test Split method was primarily used to evaluate predictive performance on an unseen test dataset, while Stratified K-Fold Cross-Validation was applied to assess model robustness and generalizability.

Performance evaluation was conducted using Accuracy, Precision, Recall (Sensitivity), F1-Score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Accuracy was used to measure overall classification performance, whereas ROC-AUC was used to evaluate the models' ability to discriminate between benign and malignant thyroid nodules across different classification thresholds.

All machine learning models were implemented in Python using commonly adopted machine learning libraries and evaluated under identical experimental conditions to ensure a fair comparison of predictive performance.

## 4. RESULTS

The results obtained using different algorithms are given below.

**Naive Bayes**

Using the Stratified Train-Test Split validation strategy, the Naive Bayes model achieved the highest classification accuracy among all evaluated algorithms (92.31%). In addition, the model produced an ROC-AUC value of 86%, demonstrating strong discriminative performance in distinguishing benign from malignant thyroid nodules. These findings suggest that Naive Bayes represents an effective and computationally efficient approach for thyroid cancer classification within the studied dataset.

**Table 1. Performance Evaluation Metrics for the Naive Bayes Classifier**

| Model | Accuracy | Precision | Recall | F1 Score | ROC-AUC |
|---|---|---|---|---|---|
| Naive Bayes | %92.31 | %66.67 | %80 | %72.73 | %86 |

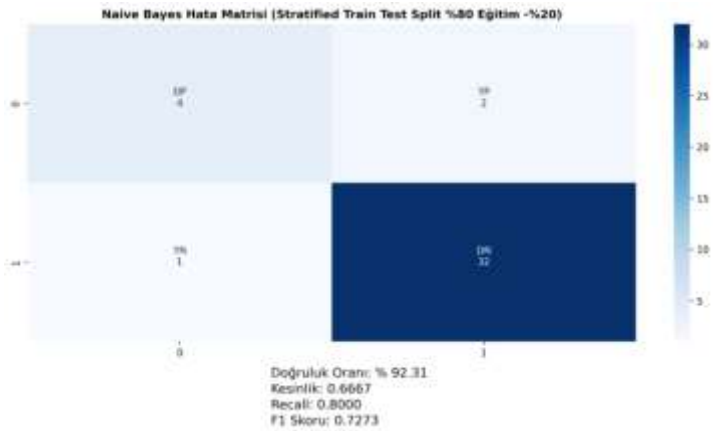


**Figure 1.** Confusion Matrix of the Naïve Bayes Classifier Using the Stratified Train-Test Split Method (80% Training, 20% Testing)

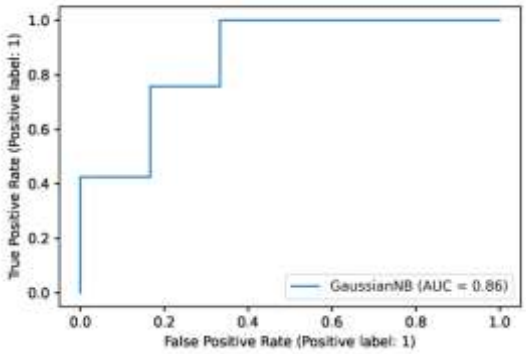


**Figure 2.** ROC Curve of the Naive Bayes Classifier

```
model_name = "Naive Bayes"
start_time = datetime.datetime.now()
(score, accuracy, predictions) = model_building(GaussianNB(), X_train, X_test, y_train, y_test )
end_time = datetime.datetime.now()
time_taken = end_time - start_time
conf_mat = confusion_matrix(y_test, predictions)
Tp = conf_mat[0, 0]
Tn = conf_mat[1, 1]
Fp = conf_mat[0, 1]
Fn = conf_mat[1, 0]
result = {
    "Accuracy": (Tp+Tn)/(Tp+Tn+Fp+Fn)*100,
    "Precision": Tp/(Tp+Fp),
    "Recall": Tp/(Tp+Fn),
    "F1_Score": 2*Tp/(2*Tp+Fp+Fn),
    "Time": time_taken.microseconds/1000000,
}
print("{} - Stratified Train Test Split\n".format(model_name), result)
```

```
✓ 0.0s
Naive Bayes - Stratified Train Test Split
 {'Accuracy': 92.30769230769233, 'Precision': 0.6666666666666666, 'Recall': 0.8, 'F1_Score': 0.7272727272727271, 'Time': 0.004999}
```

**Figure 3.** Python Code Used for the Naive Bayes

### Decision Tree Classifier

The Decision Tree classifier was evaluated using the Stratified Train-Test Split approach, in which 80% of the dataset was used for training and 20% for testing. The model correctly classified 4 patients with Bethesda V–VI findings (suspected thyroid malignancy) and 26 patients with Bethesda II findings (benign thyroid nodules). Misclassifications occurred for 2 malignant cases and 7 benign cases. The model achieved an ROC-AUC value of 71% and an overall classification accuracy of 76.92%, indicating moderate classification performance compared with the other evaluated algorithms.

**Table 2.** Performance Evaluation Metrics for the Decision Tree Classifier

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Decision Tree | 76.92 | 66.67 | 36.36 | 47.06 | 71.00 |

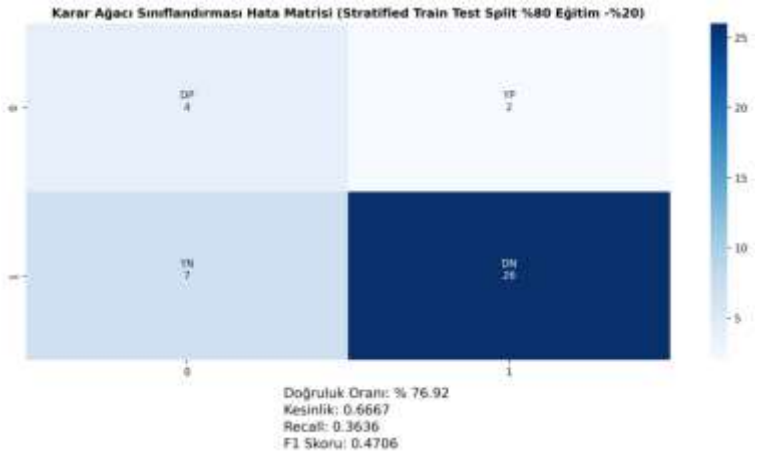


**Figure 4.** Confusion Matrix of the Decision Tree Classifier

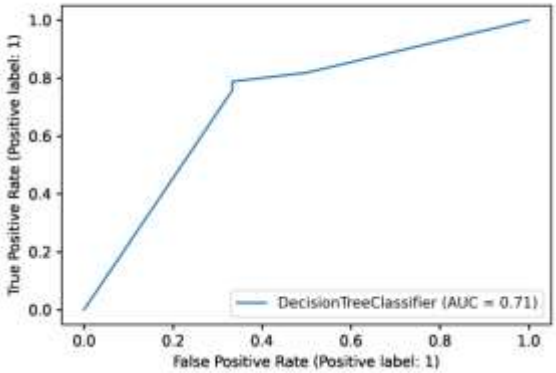


**Figure 5.** ROC Curve of the Decision Tree Classifier


**Figure 6.** Python Code Used for the Decision Tree

**Support Vector Machine Classifier**

The Support Vector Machine (SVM) classifier was evaluated using the Stratified Train-Test Split approach, in which 80% of the dataset was allocated for training and 20% for testing. On the test dataset consisting of 39 patients, the model correctly classified 34 benign cases (Bethesda II). However, none of the malignant cases (Bethesda V–VI) were correctly identified, resulting in five false-negative predictions. No false-positive classifications were observed. The model achieved an overall classification accuracy of 87.18% and an ROC-AUC value of 85%. Although the SVM classifier demonstrated relatively high accuracy, its inability to correctly identify malignant cases suggests limitations in sensitivity for thyroid cancer detection within the studied dataset.

**Table 3.** Performance Evaluation Metrics for the Support Vector Machine Classifier

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Support Vector Machine (SVM) | 87.18 | 0.00 | 0.00 | 0.00 | 85.00 |

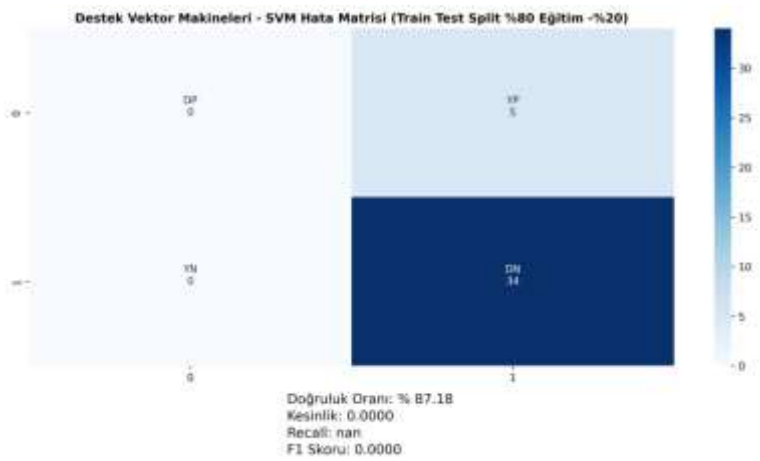


**Figure 7.** Confusion Matrix of the Support Vector Machine Classifier

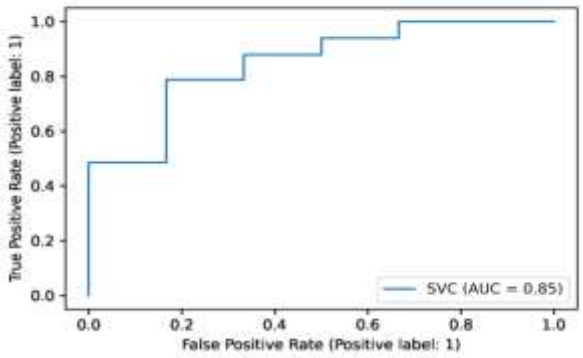


**Figure 8.** ROC Curve of the Support Vector Machine Classifier

```
model_name = "Destek Vektor Makineleri - SVM"
start_time = datetime.datetime.now()
(score, accuracy, predictions) = model_building(SVC(), X_train, X_test, y_train, y_test )
end_time = datetime.datetime.now()
time_taken = end_time - start_time
conf_mat = confusion_matrix(y_test, predictions)
Tp = conf_mat[0, 0]
Tn = conf_mat[1, 1]
Fp = conf_mat[0, 1]
Fn = conf_mat[1, 0]
result = {
    "Accuracy": (Tp+Tn)/(Tp+Tn+Fp+Fn)*100,
    "Precision": Tp/(Tp+Fp),
    "Recall": Tp/(Tp+Fn),
    "F1_Score": 2*Tp/(2*Tp+Fp+Fn),
    "Time": time_taken.microseconds/1000000,
}
print("{} - Stratified Train Test Split\n".format(model_name), result)
```

✓ 0.0s

```
Destek Vektor Makineleri - SVM - Stratified Train Test Split
 {'Accuracy': 84.61538461538461, 'Precision': 0.0, 'Recall': nan, 'F1_Score': 0.0, 'Time': 0.007996}
```

**Figure 9.** Python Code Used for the Support Vector Machine

### Logistic Regression Model

The Logistic Regression model was evaluated using the Stratified Train-Test Split approach, with 80% of the dataset allocated for training and 20% for testing. On the test dataset consisting of 39 patients, the model correctly classified 33 benign cases (Bethesda II) and 1 malignant case (Bethesda V–VI). No benign cases were incorrectly classified as malignant; however, 5 malignant cases were misclassified as benign. The model achieved an overall classification accuracy of 87.18% and an ROC-AUC value of 88%. These results indicate that Logistic Regression provided strong overall classification performance and excellent specificity, although its sensitivity for detecting malignant cases remained limited within the studied dataset.

**Table 4.** Performance Evaluation Metrics for the Logistic Regression Model

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Logistic Regression | 87.18 | 16.67 | 100.00 | 28.57 | 88.00 |

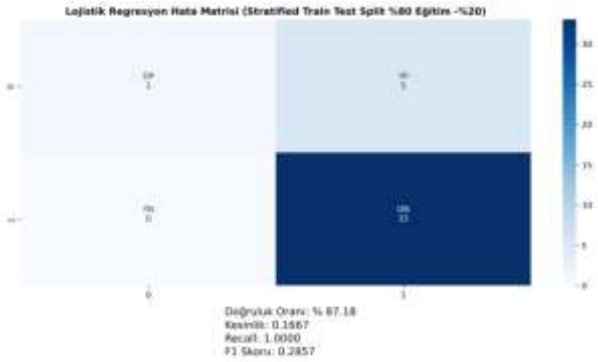


**Figure 10.** Confusion Matrix of the Logistic Regression Model

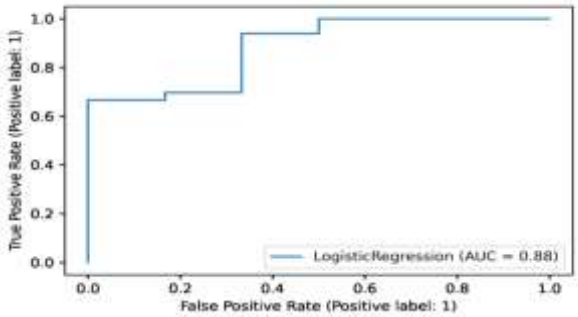


**Figure 11.** ROC Curve of the Logistic Regression Model

```
model_name = "Lojistik Regresyon"
start_time = datetime.datetime.now()
(score, accuracy, predictions) = model_building(LogisticRegression(), X_train, X_test, y_train, y_test )
end_time = datetime.datetime.now()
time_taken = end_time - start_time
conf_mat = confusion_matrix(y_test, predictions)
Tp = conf_mat[0, 0]
Tn = conf_mat[1, 1]
Fp = conf_mat[0, 1]
Fn = conf_mat[1, 0]
result = {
    "Accuracy": (Tp+Tn)/(Tp+Tn+Fp+Fn)*100,
    "Precision": Tp/(Tp+Fp),
    "Recall": Tp/(Tp+Fn),
    "F1_Score": 2*Tp/(2*Tp+Fp+Fn),
    "Time": time_taken.microseconds/1000000,
}
print("{} - Stratified Train Test Split\n".format(model_name), result)
```

```
✓ 0.0s
Lojistik Regresyon - Stratified Train Test Split
{'Accuracy': 87.17948717948718, 'Precision': 0.16666666666666666, 'Recall': 1.0, 'F1_Score': 0.2857142857142857, 'Time': 0.012417}
```

**Figure 12.** Python Code Used for the Logistic Regression Model

### Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis (LDA) classifier was evaluated using the Stratified Train-Test Split approach, in which 80% of the dataset was used for training and 20% for testing. On the test dataset consisting of 39 patients, the model correctly classified 33 benign cases (Bethesda II) and 2 malignant cases (Bethesda V–VI). No benign cases were incorrectly classified as malignant, whereas 4 malignant cases were misclassified as benign. The classifier achieved an overall accuracy of 89.74% and an ROC-AUC value of 89%. These results indicate that LDA provided strong classification performance and excellent specificity while maintaining competitive discriminative capability for thyroid cancer diagnosis.

**Table 5.** Performance Evaluation Metrics for the Linear Discriminant Analysis Classifier

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Linear Discriminant Analysis (LDA) | 89.74 | 33.33 | 100.00 | 50.00 | 89.00 |

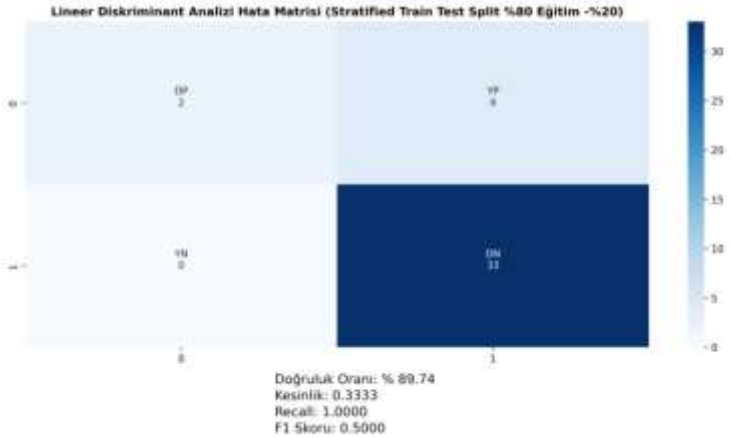


**Figure 13.** Confusion Matrix of the Linear Discriminant Analysis Classifier

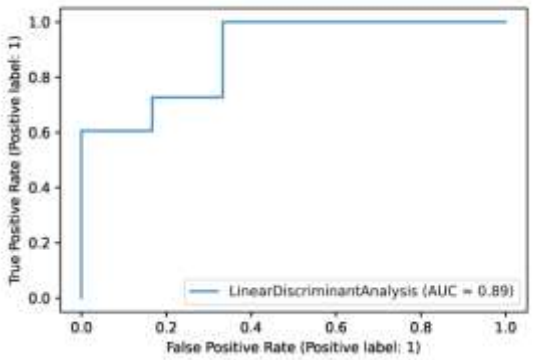


**Figure 14.** ROC Curve of the Linear Discriminant Analysis Classifier

```
model_name = "Lineer Diskriminant Analizi"
start_time = datetime.datetime.now()
(score, accuracy, predictions) = model_building(LinearDiscriminantAnalysis(), X_train, X_test, y_train, y_test )
end_time = datetime.datetime.now()
time_taken = end_time - start_time
conf_mat = confusion_matrix(y_test, predictions)
Tp = conf_mat[0, 0]
Tn = conf_mat[1, 1]
Fp = conf_mat[0, 1]
Fn = conf_mat[1, 0]
result = {
    "Accuracy": (Tp+Tn)/(Tp+Tn+Fp+Fn)*100,
    "Precision": Tp/(Tp+Fp),
    "Recall": Tp/(Tp+Fn),
    "F1_Score": 2*Tp/(2*Tp+Fp+Fn),
    "Time": time_taken.microseconds/1000000,
}
print("{} - Stratified Train Test Split\n".format(model_name), result)
```

✓ 0.0s

```
Lineer Diskriminant Analizi - Stratified Train Test Split
 {'Accuracy': 89.74358974358975, 'Precision': 0.3333333333333333, 'Recall': 1.0, 'F1_Score': 0.5, 'Time': 0.005638}
```

**Figure 15.** Python Code Used for the Linear Discriminant Analysis

### Random Forest Classifier

The Random Forest classifier was evaluated using the Stratified Train-Test Split approach, in which 80% of the dataset was allocated for training and 20% for testing. On the test dataset consisting of 39 patients, the model correctly classified 29 benign cases (Bethesda II) and 2 malignant cases (Bethesda V–VI). Misclassifications occurred for 4 benign cases and 4 malignant cases. The classifier achieved an overall accuracy of 79.49% and an ROC-AUC value of 85%. Although the Random Forest model demonstrated moderate classification performance, its ROC-AUC value suggests a reasonable ability to discriminate between benign and malignant thyroid nodules.

**Table 6.** Performance Evaluation Metrics for the Random Forest Classifier

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| **Random Forest** | **%79.49** | **%33.33** | **%33.33** | **%33.33** | **%85** |

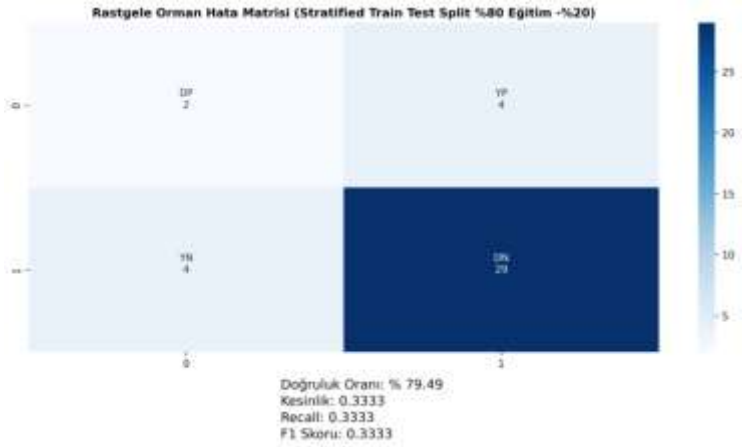


**Figure 16.** Confusion Matrix of the Random Forest Classifier

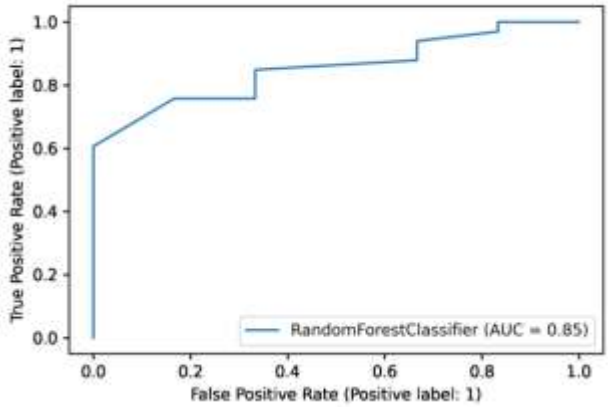


**Figure 17.** ROC Curve of the Random Forest Classifier

```
model_name = "Rastgele Orman"
start_time = datetime.datetime.now()
(score, accuracy, predictions) = model_building(RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=5), X_train, X_test, y_train, y_test )
end_time = datetime.datetime.now()
time_taken = end_time - start_time
conf_mat = confusion_matrix(y_test, predictions)
Tp = conf_mat[0, 0]
Tn = conf_mat[1, 1]
Fp = conf_mat[0, 1]
Fn = conf_mat[1, 0]
result = {
    "Accuracy": (Tp+Tn)/(Tp+Tn+Fp+Fn)*100,
    "Precision": Tp/(Tp+Fp),
    "Recall": Tp/(Tp+Fn),
    "F1_Score": 2*Tp/(2*Tp+Fp+Fn),
    "Time": time_taken.microseconds/1000000,
}
print("{} - Stratified Train Test Split\n".format(model_name), result)
```

```
Rastgele Orman - Stratified Train Test Split
{'Accuracy': 79.48717948717949, 'Precision': 0.3333333333333333, 'Recall': 0.3333333333333333, 'F1_Score': 0.3333333333333333, 'Time': 0.017575}
```

**Figure 18.** Python Code Used for the Random Forest

### K-Nearest Neighbors (KNN) Classifier

The K-Nearest Neighbors (KNN) classifier was evaluated using the Stratified Train-Test Split approach, in which 80% of the dataset was allocated for training and 20% for testing. On the test dataset consisting of 39 patients, the model correctly classified 31 benign cases (Bethesda II) and 1 malignant case (Bethesda V–VI). Misclassifications occurred for 2 benign cases and 5 malignant cases. The classifier achieved an overall accuracy of 82.05% and an ROC-AUC value of 78%. These results indicate that the KNN model demonstrated moderate classification performance however, its relatively low sensitivity in detecting malignant cases may limit its effectiveness for thyroid cancer diagnosis within the studied dataset.

**Table 7.** Performance Evaluation Metrics for the K-Nearest Neighbors Classifier

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| K-Nearest Neighbors (KNN) | 82.05 | 16.67 | 33.33 | 22.22 | 78.00 |

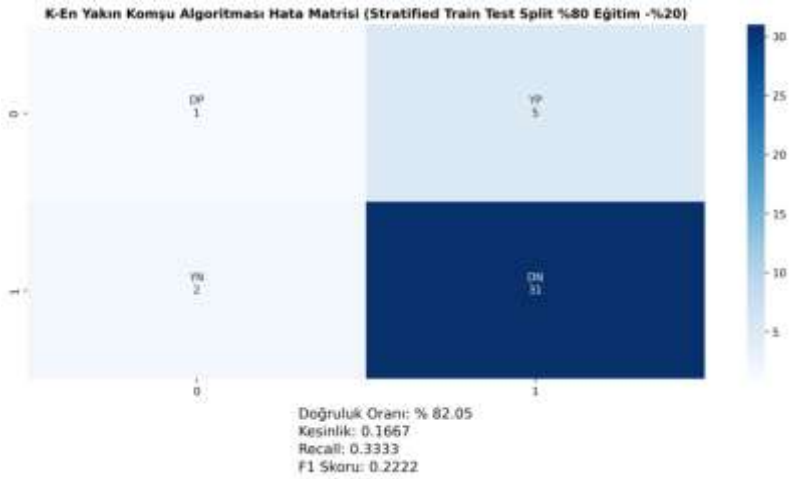


**Figure 19.** Confusion Matrix of the K-Nearest Neighbors Classifier

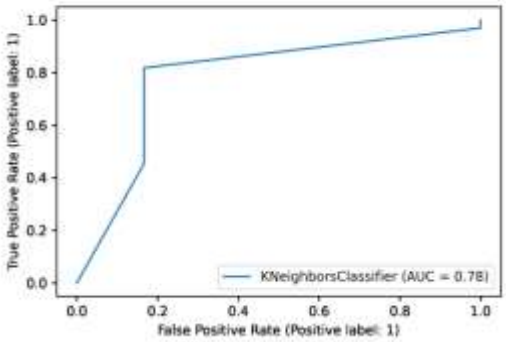


**Figure 20.** ROC Curve of the K-Nearest Neighbors Classifier

```
model_name = "K-En Yakın Komşu Algoritması"
start_time = datetime.datetime.now()
(score, accuracy, predictions) = model_building(KNeighborsClassifier(), X_train, X_test, y_train, y_test )
end_time = datetime.datetime.now()
time_taken = end_time - start_time
conf_mat = confusion_matrix(y_test, predictions)
Tp = conf_mat[0, 0]
Tn = conf_mat[1, 1]
Fp = conf_mat[0, 1]
Fn = conf_mat[1, 0]
result = {
    "Accuracy": (Tp+Tn)/(Tp+Tn+Fp+Fn)*100,
    "Precision": Tp/(Tp+Fp),
    "Recall": Tp/(Tp+Fn),
    "F1_Score": 2*Tp/(2*Tp+Fp+Fn),
    "Time": time_taken.microseconds/1000000,
}
print("{} - Stratified Train Test Split\n".format(model_name), result)
```

✓ 0.0s

```
K-En Yakın Komşu Algoritması - Stratified Train Test Split
 {'Accuracy': 82.05128205128204, 'Precision': 0.16666666666666666, 'Recall': 0.3333333333333333, 'F1_Score': 0.2222222222222222, 'Time': 0.012081}
```

**Figure 21.** Python Code Used for the K-Nearest Neighbors

### Artificial Neural Network (ANN) Classifier

The Artificial Neural Network (ANN) classifier was evaluated using the Stratified Train-Test Split approach, in which 80% of the dataset was allocated for training and 20% for testing. On the test dataset consisting of 39 patients, the model correctly classified 33 benign cases (Bethesda II). However, none of the malignant cases (Bethesda V–VI) were correctly identified, resulting in six false-negative predictions. No false-positive classifications were observed. The classifier achieved an overall accuracy of 84.62% and an ROC-AUC value of 50%. Although the ANN model produced a relatively high accuracy value, its inability to identify malignant cases and its ROC-AUC value of 50% indicate poor discriminative performance for thyroid cancer classification within the studied dataset.

**Table 8.** Performance Evaluation Metrics for the Artificial Neural Network Classifier

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Artificial Neural Network (ANN) | 84.62 | 0.00 | 0.00 | 0.00 | 50.00 |

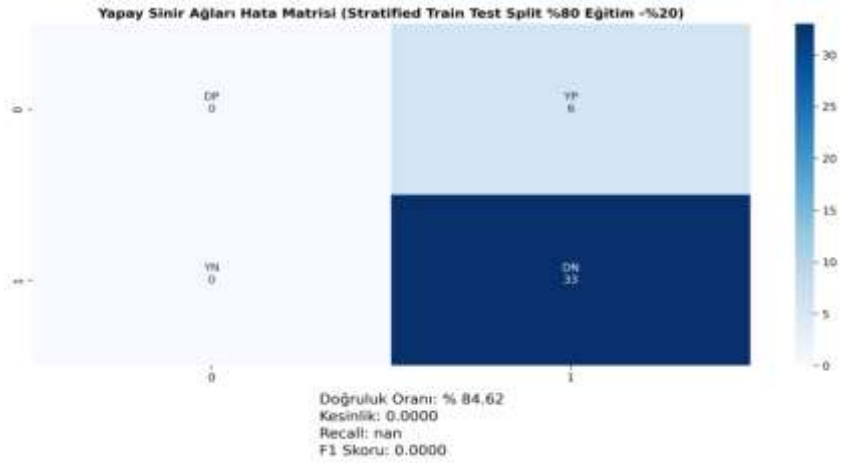


**Figure 22.** Confusion Matrix of the Artificial Neural Network Classifier

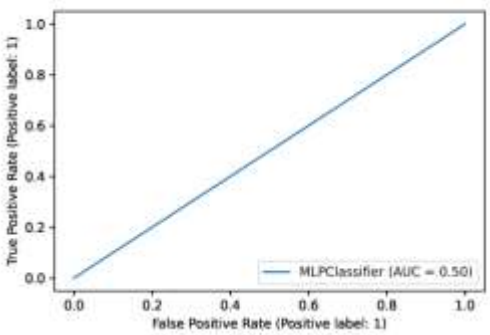


**Figure 23.** ROC Curve of the Artificial Neural Network Classifier


**Figure 24.** Python Code Used for the Artificial Neural Network

## 5. COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS

The predictive performance of thirteen supervised machine learning algorithms was evaluated using Stratified Train-Test Split and Stratified K-Fold Cross-Validation approaches. Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and computational efficiency metrics.

Overall, ensemble learning methods demonstrated the strongest predictive performance. In particular, XGBoost achieved the highest ROC-AUC value (92%), indicating superior discrimination capability between benign and malignant thyroid nodules. CatBoost, LightGBM, AdaBoost, and Gradient Boosting also produced competitive results, with ROC-AUC values exceeding 90% in several evaluation scenarios.

Among all evaluated models, Naïve Bayes achieved the highest classification accuracy (92.31%), correctly classifying 36 of the 39 observations in the test dataset. Despite its relatively simple probabilistic structure, the model demonstrated strong predictive capability and competitive ROC-AUC performance (86%).

Linear Discriminant Analysis (LDA) also produced favorable results, achieving an accuracy of 89.74% and an ROC-AUC value of 89%. Furthermore, LDA required one of the shortest prediction times among the highest-performing models, highlighting its potential suitability for computationally efficient clinical decision-support applications.

In contrast, some algorithms exhibited lower predictive performance. Decision Tree and K-Nearest Neighbors produced ROC-AUC values of 71% and 78%, respectively. Similarly, the Artificial Neural Network model achieved limited discrimination capability, with an ROC-AUC value close to random classification performance. This outcome may be attributed to the relatively small sample size available for model training, which can limit the effectiveness of neural network-based approaches.

The comparative results indicate that while traditional statistical learning approaches remain effective for thyroid cancer classification, ensemble learning methods provide superior discrimination performance and appear better suited for capturing complex nonlinear relationships among clinical variables.

**Table 9.** Comparison of Classification Performance Across Machine Learning Models

| Model | Accuracy (%) | ROC-AUC (%) | Prediction Time (s) |
|---|---|---|---|
| Artificial Neural Network | 84.62 | 50.00 | 0.010003 |
| Logistic Regression | 87.18 | 88.00 | 0.009521 |
| XGBoost | 89.74 | 92.00 | 0.049512 |
| AdaBoost | 87.18 | 91.00 | 0.074549 |
| LightGBM | 87.18 | 91.00 | 0.020999 |
| CatBoost | 87.18 | 91.00 | 0.052616 |
| Gradient Boosting | 84.62 | 81.00 | 0.054029 |
| Random Forest | 79.49 | 85.00 | 0.018518 |
| Decision Tree | 76.92 | 71.00 | 0.004000 |
| Support Vector Machine | 84.62 | 85.00 | 0.005031 |
| Naive Bayes | 92.31 | 86.00 | 0.004003 |
| K-Nearest Neighbors | 82.05 | 78.00 | 0.009998 |
| Linear Discriminant Analysis | 89.74 | 89.00 | 0.003000 |

Table 9 summarizes the predictive performance of the evaluated machine learning algorithms. Overall, ensemble learning methods demonstrated the strongest discrimination capability, whereas probabilistic approaches achieved the highest classification accuracy.

Among all evaluated models, Naïve Bayes achieved the highest accuracy (92.31%). In the test dataset consisting of 39 patients, the model correctly classified 36 cases, including four malignant and thirty-two benign nodules. The model achieved an ROC-AUC value of 86%, indicating strong diagnostic performance despite its relatively simple probabilistic structure.

XGBoost achieved the highest ROC-AUC value (92%), demonstrating superior discrimination capability between benign and malignant thyroid nodules. Other ensemble learning approaches, including LightGBM, CatBoost, AdaBoost, and Gradient Boosting, also produced competitive results and consistently ranked among the highest-performing models.

Linear Discriminant Analysis (LDA) achieved comparable performance, with an accuracy of 89.74% and an ROC-AUC value of 89%, while requiring relatively low computational resources. In contrast, Decision Tree, K-Nearest Neighbors, and Artificial Neural Networks exhibited lower discrimination performance, suggesting that more complex nonlinear models did not necessarily benefit from the relatively limited dataset size.

## 6. DISCUSSION

The findings of this study demonstrate that supervised machine learning algorithms can effectively support thyroid cancer diagnosis using clinical, laboratory, ultrasonographic, and biopsy-related variables. Among the evaluated models, Naïve Bayes achieved the highest classification accuracy (92.31%), while XGBoost obtained the highest ROC-AUC value (92%), indicating superior discrimination capability between benign and malignant thyroid nodules.

Interestingly, although ensemble learning algorithms such as XGBoost, AdaBoost, LightGBM, and CatBoost consistently produced the highest ROC-AUC values, Naïve Bayes outperformed all other models in terms of overall classification accuracy. This finding may suggest that the relationships present within the dataset are sufficiently structured for probabilistic learning approaches to capture effectively, despite the relatively limited sample size. Ensemble learning methods, however, appear to provide stronger discrimination performance by capturing more complex nonlinear interactions among clinical variables.

The results obtained in this study are generally consistent with previous findings reported in the literature. Xi Miles et al. (2022) reported accuracy values of approximately 77% and 78% for Gradient Boosting and Logistic Regression models, respectively. In comparison, the corresponding models in the present study achieved accuracy values of 84.61% and 87.18%. Similarly, Support Vector Machines and Linear Discriminant Analysis demonstrated competitive performance, further supporting the applicability of machine learning techniques for thyroid cancer prediction. Differences across studies may be explained by variations in dataset characteristics, feature selection procedures, preprocessing strategies, and patient populations.

The performance of ensemble learning approaches observed in this study is also consistent with previous research emphasizing the effectiveness of tree-based ensemble models in classification tasks (Tolomei et al., 2017). In particular, XGBoost, AdaBoost, LightGBM, and CatBoost achieved some of the highest ROC-AUC values among all evaluated models, suggesting that ensemble methods may be especially effective in capturing complex relationships associated with thyroid malignancy. These findings support the growing evidence that ensemble learning approaches can provide robust predictive performance in medical decision-support applications.

An important strength of the present study is the rigorous data preparation process. Although the original dataset contained 1,994 patient records, extensive quality-control procedures and strict inclusion criteria were applied to identify complete and clinically reliable observations. As a result, only 193 complete cases were retained for analysis. While this substantially reduced the sample size, it eliminated the need for missing-data imputation procedures and increased the reliability of the final dataset. Previous studies have highlighted the critical role of data preparation and data quality in achieving reliable artificial intelligence-based clinical decision support systems, particularly when using real-world healthcare data (Maranghi et al., 2022).

Another contribution of this study is the comprehensive comparison of thirteen supervised machine learning algorithms within a single experimental framework. Unlike many previous studies that focused on a limited number of models, the present research simultaneously evaluated traditional statistical classifiers, machine learning algorithms, neural networks, and multiple ensemble learning techniques. This broad comparison provides a more complete understanding of the strengths and limitations of different approaches for thyroid cancer diagnosis.

Several limitations should be acknowledged. First, the final dataset consisted of only 193 patients, which may limit the generalizability of the findings. Second, class imbalance was present, with substantially fewer malignant cases than benign cases. Although stratified sampling methods were employed to preserve class distributions during model evaluation, model performance may still have been influenced by this imbalance. Third, data were obtained from a single healthcare institution, limiting external validity and reducing the ability to generalize findings to broader patient populations.

Future research should evaluate these algorithms using larger multicenter datasets and external validation cohorts. Furthermore, integrating explainable artificial intelligence (XAI) techniques may improve the transparency and clinical interpretability of machine learning predictions. The incorporation of multimodal learning approaches that combine imaging data with structured clinical information may further enhance predictive performance. Finally, future studies may benefit from moving beyond purely correlation-based predictive models toward causally informed machine learning frameworks capable of supporting more trustworthy and clinically meaningful decision-making processes.

In conclusion, the findings indicate that supervised machine learning algorithms can provide accurate and clinically relevant predictions for thyroid cancer diagnosis. While Naïve Bayes achieved the highest classification accuracy, XGBoost demonstrated the strongest discrimination capability. Overall, the results support the potential of machine learning techniques as decision-support tools in thyroid cancer diagnosis and provide a foundation for future research on explainable, multimodal, and causally informed artificial intelligence systems in healthcare.

## REFERENCES


- Türkiye Endokrinoloji ve Metabolizma Derneği. (2025). Tiroit hastalıkları tanı ve tedavi kılavuzu [Guidelines for the diagnosis and treatment of thyroid diseases]. Türkiye Endokrinoloji ve Metabolizma Derneği.
- Maranghi, M., Baldassarre, M., Bianchi, F., & Ferri, F. (2022). Data preparation and analytics processes for artificial intelligence-based clinical decision support systems. Healthcare Analytics, 2, 100089.
- Zhao, H., Xue, T., Zhang, Y., Lu, Y., Wang, D., Yin, P., Jiang, L., Liu, D., & Wang, Y. (2026). Artificial intelligence in the diagnosis of thyroid diseases: Applications and challenges. Frontiers in Radiology, 6, 1740915.
- Johnson, T., & Tessler, F. N. (2026). Artificial intelligence applications in thyroid cancer diagnosis: 2026 update. Journal of Diagnostic Medical Sonography, 36(2). https://doi.org/10.1177/10507256251412316
- Pizzato, M., Li, M., Vignat, J., Laversanne, M., Singh, D., La Vecchia, C., & Vaccarella, S. (2022). The epidemiological landscape of thyroid cancer worldwide: GLOBOCAN estimates for incidence and mortality rates in 2020. The Lancet Diabetes & Endocrinology, 10(4), 264–272.
- Chakrabarty, N., Mahajan, A., Basu, S., & D'Cruz, A. K. (2024). Comprehensive review of the imaging recommendations for diagnosis, staging, and management of thyroid carcinoma. Journal of Clinical Medicine, 13(10), 2904.
- Fei, M., Zhang, F., Ding, D., Ouyang, X., Shen, W., Zhang, B., & Qin, L. (2024). Diagnostic significance of ultrasound-guided fine-needle aspiration biopsy and on-site assessment by pathologists for thyroid micronodules. International Journal of General Medicine, 17, 315–321.
- Osseis, M., Jammal, G., Kazan, D., & Noun, R. (2023). Comparison between fine needle aspiration cytology with histopathology in the diagnosis of thyroid nodules. Journal of Personalized Medicine, 13(8), 1197.
- Gatta, G., Trimboli, P., Grani, G., Durante, C., Filetti, S., Campanella, P., et al. (2025). Machine learning for diagnosis of malignant thyroid nodules based on thyroid ultrasound:

Systematic review and meta-analysis of studies with external validation. European Journal of Endocrinology, 192, K1–K13.

- Lee, J. H., Kim, Y. J., & Park, S. H. (2022). Machine learning applications in thyroid disease diagnosis: A systematic review. Diagnostics, 12(8), 1875.
- Xi, N. M., Zhang, H., Wang, J., & Chen, Y. (2022). Machine learning models for thyroid nodule malignancy prediction: A comparative study. BMC Medical Informatics and Decision Making, 22(1), 245.
- Kumar, A., Sharma, R., & Singh, P. (2023). Machine learning approaches for thyroid cancer prediction using SEER data. Healthcare Analytics, 3, 100180.
- Jassal, K., Edwards, M., Koohestani, A., Brown, W., Serpell, J. W., & Lee, J. C. (2025). Beyond genomics: Artificial intelligence-powered diagnostics for indeterminate thyroid nodules—a systematic review and meta-analysis. Frontiers in Endocrinology, 16, 1506729.
- Edström, A. B., et al. (2025). Human–AI collaboration for ultrasound diagnosis of thyroid nodules: A clinical trial. European Archives of Oto-Rhino-Laryngology, 282, 3221–3231.
- Ni, J. H., et al. (2025). Optimizing thyroid nodule management with artificial intelligence in real clinical settings. JMIR Medical Informatics, 13, e71740.
- Riajulislam, M., Rahman, M. M., Hossain, M. S., & Islam, M. R. (2021). Early detection of hypothyroidism using machine learning and feature selection techniques. Applied Sciences, 11(22), 10874.
- Li, X., Wang, L., & Zhang, Y. (2012). Thyroid disease diagnosis based on principal component analysis and extreme learning machine. Journal of Medical Systems, 36(6), 3647–3652.
- Ma, J., Wu, F., Jiang, T., Zhao, Q., & Kong, D. (2019). Ultrasound image-based thyroid nodule classification using convolutional neural networks. IEEE Access, 7, 34201–34208.
- Yao, J., et al. (2025). Multimodal GPT model for assisting thyroid nodule risk assessment and management. NPJ Digital Medicine, 8, Article 152.
- Alsaadawi, M. A. W. (2023). Detection of thyroid disease using machine learning models [Master's thesis, Karabük University].
- Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 465–474.