AIS ARTIFICIAL INTELLIGENCE STUDIES

Beyond Diagnosis: Cross-Dataset Evaluation of Risk Factors for Thyroid Cancer Recurrence

Mehmet Ali Dursun[*a], **Pınar Özen Kavas**[b]

ABSTRACT

This study aims to comparatively evaluate various machine learning algorithms developed for the classification of thyroid diseases. By employing five distinct datasets with differing statistical structures and class imbalances, the performance of nine algorithms—CatBoost, XGBoost, LightGBM, Random Forest, Artificial Neural Network (ANN), KNN, SVM, Stacking, and GridSearch-Tuned Logistic Regression (gst-LR) has been comprehensively analyzed. Model performance was assessed not only based on accuracy but also through multidimensional metrics such as F1-score, precision, recall, and specificity. Stratified K-Fold cross-validation was applied in the model validation processes to ensure class representation and enhance generalizability. The findings reveal that boosting-based algorithms (particularly CatBoost, XGBoost, and LightGBM) delivered high and stable accuracy across several datasets. The Random Forest model stood out with its consistent performance even on imbalanced data, whereas the ANN model demonstrated notable fluctuations depending on the structural properties of the dataset. Classical methods such as KNN and SVM achieved competitive results only when the data exhibited well-defined decision boundaries, showing limitations in more complex distributions. The systematic approach adopted in this study presents a multilayered classification framework not only for model comparison but also for the evaluation of explainability, reproducibility, and contextual suitability. The overall results indicate that no single model dominates across all scenarios; rather, the success of classification strongly depends on data characteristics such as class distribution, dimensionality, and feature separability. Models such as Random Forest and boosting algorithms consistently performed well in terms of both accuracy and F1-score, with scores exceeding 98% and 95% respectively on certain datasets. These findings underscore the importance of context-aware model selection and reinforce the need for multi-metric evaluations in real-world clinical decision support applications.

[a,*] Dumlupınar University,
Engineering Faculty,
Dept. of Computer Engineering
43000 - Kütahya, Türkiye
ORCID: 0000-0001-6370-1160

[b] Dumlupınar University,
Engineering Faculty,
Dept. of Computer Engineering
43000 - Kütahya, Türkiye
ORCID: 0000-0001-9884-2860

[*] Corresponding author.
e-mail: mehmet.dursun@dpu.edu.tr

Keywords: Thyroid malignancy, Recurrence risk, Computer-aided diagnosis, Demographic stratification

Anahtar Kelimeler: Tiroid malignitesi, Nüks riski, Bilgisayar destekli tanı, Demografik katmanlaştırma

Submitted: 28.05.2025
Revised: 27.06.2025
Accepted: 29.06.2025

doi:10.30855/ais.2025.08.01.03

Tanının Ötesinde: Farklı Veri Setleri ile Tiroid Kanseri Nüks Faktörlerinin Değerlendirilmesi

ÖZ

Bu çalışma, tiroid hastalıklarının sınıflandırılması amacıyla geliştirilen çeşitli makine öğrenmesi algoritmalarının karşılaştırmalı olarak değerlendirilmesini hedeflemektedir. Farklı istatistiksel yapılar ve sınıf dengesizlikleri içeren beş ayrı veri kümesi kullanılarak, CatBoost, XGBoost, LightGBM, Random Forest, Yapay Sinir Ağı (ANN), KNN, SVM, Stacking ve hiperparametre optimizasyonu uygulanmış Lojistik Regresyon (gst-LR) algoritmalarının performansları detaylı biçimde analiz edilmiştir. Model başarımı, yalnızca doğruluk (accuracy) temelinde değil, aynı zamanda F1 skoru, precision, recall ve specificity gibi çok boyutlu metrikler üzerinden değerlendirilmiştir. Model doğrulama süreçlerinde Stratified K-Fold çapraz doğrulama uygulanarak, sınıf temsiliyetini koruyan ve genellenebilirliği artıran bir yapı benimsenmiştir. Bulgular, özellikle boosting tabanlı algoritmaların (özellikle CatBoost, XGBoost, LightGBM) birçok veri kümesinde yüksek doğruluk ve kararlılık sağladığını göstermektedir. Random Forest modeli, dengesiz veri setlerinde dahi istikrarlı performansıyla öne çıkarken, ANN modelinin başarımı veri setinin yapısal özelliklerine göre anlamlı dalgalanmalar göstermiştir. KNN ve SVM gibi klasik yöntemler ise ayrım gücü yüksek veri yapılarında başarılı sonuçlar sunarken, daha karmaşık yapılar karşısında sınırlı kalmıştır. Elde edilen nicel bulgular, model başarısının yalnızca genel doğrulukla açıklanamayacağını; veri yapısı, sınıf dağılımı ve öğrenme kapasitesi gibi çoklu faktörlerin bütüncül etkisiyle değerlendirilmesi gerektiğini ortaya koymuştur. Random Forest, CatBoost ve LightGBM algoritmaları, farklı senaryolarda %98'in üzerinde doğruluk ve %95'e yakın F1 skoru ile en başarılı modeller olarak öne çıkmıştır. ANN yalnızca belirli veri kümelerinde bu düzeye yaklaşabilmiş, SVM ve KNN ise sınıf ayrımı belirgin olmadığında performans kaybı yaşamıştır. Bu değerlendirme, tıbbi karar destek sistemlerine entegre edilecek sınıflandırma modellerinin, yalnızca doğruluk temelinde değil, bağlamsal uyum, açıklanabilirlik ve dengeli sınıf performansı gibi kriterlerle birlikte seçilmesi gerektiğini vurgulamaktadır.

To cite this article: M.A. Dursun and P. Özen Kavas, "Beyond Diagnosis: Cross-Dataset Evaluation of Risk Factors for Thyroid Cancer Recurrence," *Artificial Intelligence Studies*, vol.8, no.1, pp. 38-69, 2025. doi:10.30855/ais.2025.08.01.03

1. Introduction (Giriş)

Thyroid cancer is the most common malignant tumour of the endocrine system and has become a more prominent clinical problem in healthcare systems, especially with increasing diagnosis rates in the last two decades [1]. It has attracted attention with a globally increasing incidence rate, especially in the last two decades. According to the data of the American Cancer Society for 2023, the number of individuals diagnosed with thyroid cancer reaches approximately 43,800 per year and it is reported that this disease is three times more common in women than in men [2]. Advances in diagnostic techniques such as ultrasonography and needle aspiration biopsy have facilitated the detection of small nodules, especially micropapillary type, at an earlier stage. However, although the possibilities for early diagnosis have increased, it is still a significant challenge to accurately determine the risk level of the disease in thyroid cancer cases with different biological subtypes and clinical courses [3]. The development of diagnostic technologies, especially the widespread use of methods such as high-resolution ultrasonography and needle aspiration biopsy, has made it possible to detect low-sized lesions such as micropapillary variants at an earlier stage [4]. However, this increase in early diagnosis rates does not eliminate the biological heterogeneity of the disease; on the contrary, it shows a great variation in terms of risk level and response to treatment. This situation reveals the inadequacy of the approach based on standard protocols and necessitates the development of individualised models based on patient-specific risk prediction.

Decision-making processes based on risk prediction are of great importance in the management of thyroid cancer cases [5]. The accuracy of risk stratification plays a decisive role both in the selection of treatment strategies applied to reduce the risk of recurrence and in preventing patients from being exposed to unnecessary aggressive treatment processes. Inadequate risk stratification reduces the success of the patient-specific medical approach and may also lead to unnecessary resource utilisation and patient dissatisfaction in healthcare delivery. Therefore, the development of predictive models and the dissemination of clinical decision support systems, which are considered as one of the basic elements of individualised medicine, have become a strategic necessity for contemporary oncology practices.

One of the most critical steps in the management of thyroid cancer cases is the accurate assignment of patients to low, intermediate or high risk groups [6]. This classification plays a decisive role in predicting the risk of recurrence, making surgical and radioactive iodine treatment decisions and planning the follow-up process. However, existing clinical classification systems are generally based on a limited number of variables and cannot adequately utilise the potential offered by multivariate biomedical data [7]. Machine learning (ML) algorithms developed to fill this gap attract attention with their capacity to extract meaningful patterns from high-dimensional and multivariate data structures. In this respect, ML plays a critical role in the future of medical decision support systems [8].

The aim of this study is to develop versatile and reliable prediction models by integrating machine learning approaches with classical statistical methods in thyroid cancer risk analysis. In this context, three different open access datasets were used and a total of nine different machine learning algorithms were systematically compared. The algorithms used include logistic regression, decision trees, random forest, support vector machines, k-nearest neighbour, XGBoost, LightGBM, gradient boosting machine (GBM) and multilayer artificial neural networks[9]. The performance of each model was evaluated with multidimensional metrics such as accuracy, sensitivity, specificity, F1 score, and inferences from statistical analyses and machine learning predictions were interpreted together.

The variables included in the data sets cover critical elements of thyroid cancer pathophysiology such as age, gender, radiotherapy history, adenopathy status, tumour focality, pathological subtype, TNM staging, clinical response and recurrence information. This multidimensional structure creates a wide learning ground for both classical statistical analyses (chi-square test, ANOVA, correlation analyses, etc.) and machine learning models [10]. As a matter of fact, the detailed statistical analysis process carried out at the beginning of the study was applied to discover the relationship between variables, identify outliers and provide a methodological basis for the model development phase.

This holistic approach integrates not only the 'black box' structure of machine learning algorithms, but also the interpretability of statistical inferences, making it possible to develop a more explainable and clinically evaluable model. In particular, the cross-analysis of statistical correlation tests that come with

AUC values enables to consider significance as well as accuracy in model selection. Thus, the study not only provides an algorithmic comparison, but also generates data-based clinical insights.

In conclusion, this research presents a new methodological framework for risk stratification of thyroid cancer using three different datasets, using both statistical analyses and machine learning algorithms in an integrated manner. The findings have important implications for the design of individualised treatment protocols, development of clinical decision support systems and bridging methodological gaps in the academic literature. At the same time, with its multi-model approach and holistic analysis method, it provides an up-to-date example of explainable and reliable artificial intelligence systems in which digital health applications are evolving.

This paper consists of an introduction and four main sections. In the second section, current research on thyroid cancer is reviewed, and studies on statistical approaches and machine learning-based models used in risk prediction are analysed in detail. In the third chapter, three different datasets used in the study, statistical analyses, selected features, machine learning algorithms and modelling process are presented in detail. In the fourth section, the performances of nine different machine learning models in thyroid cancer risk prediction are compared; comprehensive analyses of the models on accuracy, sensitivity, specificity, F1-Score and other metrics are performed. In the fifth and final section, the findings are evaluated, the contribution of the multi-analysis strategy developed in this study to clinical decision support systems is discussed and recommendations for future research are presented.

Thyroid cancer has been at the centre of numerous clinical and cognitive studies, especially in recent years, with diagnostic rates increasing globally [11]. This increase cannot be explained solely by the widespread availability of diagnostic imaging modalities; the identification of molecular biomarkers, the development of new staging systems and the drive towards individualised treatment approaches have also expanded the volume of research [12]. In particular, the widespread detection of low-risk types such as micropapillary thyroid carcinoma has necessitated more detailed analyses of recurrence risk and treatment response. This has increased the interest in multidisciplinary studies using both classical epidemiological approaches and artificial intelligence-based modelling.

In the literature, the variety of methods used in thyroid cancer risk prediction and classification processes is remarkable. The methodological spectrum ranging from traditional regression models to deep learning exhibits variable performance depending on different data sources and clinical scenarios. However, many studies have been conducted on limited variable sets or small sample sizes, and methodological controls such as model comparison and cross-validation are often neglected [13]. Therefore, a systematic evaluation of the methodological approaches available in the literature is important, especially for the analysis of multivariate and real-world data. In this context, the literature review will provide a comprehensive analysis of the machine learning and statistical techniques applied to date, as well as a stronger contextualisation of the unique aspects of this study.

1.1. Cytological and Histopathological Image Analysis Based Studies

Yu et al. [14] developed the PyMLViT (Pyramid Multi-Loss Vision Transformer) algorithm for the classification of thyroid cancer from cytological smears. The model aims to solve the problems of multi-scale structural information extraction and insufficient supervision loss in MIL processes. Pyramid token extraction is used to extract features at different scales, and multi-loss fusion modules are used to compensate for multi-level losses. In experiments with 560 samples from the Sino-Japanese Union Hospital, the PyMLViT model outperformed existing methods with 87.5% accuracy, 88.69% sensitivity and 86.62% precision. The model also offers the advantage of low complexity and high explainability.

Chandio et al.[15] proposed a three-layer CNN-based system for early classification of medullary thyroid cancer. The first layer includes image preprocessing and segmentation (e.g. Otsu, watershed), the second layer includes classification with CNN, and the third layer includes visualisation of the results. In the analysis of 5601 cell nuclei obtained from SMBBMU Hospital in Pakistan, the model provided 99% accuracy and 99.18% precision. The model performs particularly well in the discrimination of eccentric nucleus morphologies and is highly accurate at the cellular level.

Shabrina et al. [16] used ConvNeXt Tiny model and Grad-CAM interpretation method for classification of PTC histopathological images. In the study with 1496 WSI images, the model achieved the best result

with 84.36% accuracy, recall and precision at 256×256 patch size. Model decisions were visualised using Grad-CAM; however, it was emphasised that further architectural improvements are needed due to low resolution and excessive learning potential.

Gavade et al. [17] used pre-trained CNN (specifically VGG16) models for automatic classification of thyroid cancer subtypes from histopathology images. In the study, problems such as limited data, explainability and bias were addressed with Grad-CAM, LRP and fairness-aware analyses. With K-fold cross-validation, over 90% accuracy and F1 score were obtained; especially high success was shown in the papillary carcinoma class. Model outputs were evaluated by fairness and sensitivity analyses.

Tschuchnig et al. [18] compared patch-based MIL methods for the classification of papillary and follicular thyroid nodules. Using a dataset of 40 histology images, features were extracted with ResNet18 and three different multiscale combination methods (MC, MA, MM) were tested. The best result was obtained with the MM method with 88% accuracy, showing that this method can improve classification performance when carefully designed.

1.2. Deep Learning Approaches Based on Ultrasound and CT Images

Zhuang et al. [19] developed an attention-based model called AMIL that combines multiscale ultrasound image features. The model performs both patch and frame level classification under weakly supervised learning. The highest performance was obtained by combining outputs from different patch sizes with an ensemble approach (0.785% AUROC). Visualisation techniques showed that the model focused on clinically relevant regions (nodule edges, hyperechoic areas). This method, which can provide highly accurate patient-level predictions without relying on segmentation, is remarkable in clinical decision support systems.

Zhang et al. [20] proposed three multichannel CNN architectures based on Xception for early diagnosis of thyroid cancer. In experiments with DDTI and clinical images, the SIDC model provided up to 98.9% accuracy. DIDC and four-channel structures also showed high performance. These structures, which show superiority in terms of both accuracy and explainability in CT and USG data, increase clinical adaptability.

Nugroho and Frannita [21] compared DenseNet121 and NasNetLarge architectures with transfer learning in ultrasound images. In the study with a data set of 348 samples, the NasNetLarge model was found to be more successful with 87.14% accuracy and 88.45% AUC. NasNetLarge, which offers more stable and accurate results despite its complex structure, stands out in terms of classification success.

Deepana et al. [22] extracted features with various CNN architectures and classified them with different ML algorithms. In the study of 118 ultrasound images, the combination of ResNet-50 + XGBoost gave the best result with 82.12% accuracy. Other models remained in the range of 79.7%-79.9%. The combination of deep learning and classical methods stands out as an effective solution in terms of clinical accuracy.

1.3. Recurrence, Risk and Prediction Models with Machine Learning

Arslan and Çolak [23] aimed to predict the risk of recurrence in well-differentiated thyroid cancer (Well-DTC) patients with explainable machine learning (XAI) methods. In the data set of 383 patients, four variables (Response, Risk, T, N) determined by distance correlation method were used. Fast Interpretable Greedy-Tree Sums (FIGS) and Explainable Boosting Machines (EBM) algorithms were compared; EBM model showed superior performance with 96.1% accuracy, 99.3% AUC and 91.9% F1 score. In particular, response to treatment, ATA risk classification, tumour stage and lymph node metastasis were the most important predictors. With SHAP, the contribution of each variable to the decision was visualised to increase the explainability of the model.

Aida et al. [24] compared ELM and HMM algorithms to assess the probability of recurrence in differentiated thyroid cancer (DTC). A dataset containing 383 observations and 16 features balanced with SMOTE was used. The ELM model achieved 100% accuracy, sensitivity and specificity with a 90:10 training-test ratio, while the HMM model gave the best result with 86.36% accuracy and 0.9343% AUC.

While HMM's ability to model unexplained situations was highlighted, ELM's high-speed and accurate results were especially prominent in unbalanced data classes.

Hegde et al. [25] proposed a hybrid meta-classifier model combining SVM and RF classifiers with PSO and GA based feature selection for thyroid cancer prediction. A dataset of 3800 samples from Kaggle was used; data preprocessing, balancing with SMOTE and hyperparameter optimisation were performed. The model outperformed traditional methods with 98% accuracy, 97% sensitivity and 0.98% AUC. The combination of PSO, GA and meta-classifier improved the predictive power and generalisability, especially in high-dimensional data sets.

Anuhya [26] aimed to determine the most effective model by comparing SVM, KNN, Decision Tree and Random Forest algorithms in thyroid cancer classification. In the study, a clinical data set consisting of 2796 samples including hormone values such as TSH, T3, TT4 and age and gender information was used. Preprocessing steps such as missing value removal, one-hot coding and normalisation were applied to the data, followed by k-fold cross validation and hyperparameter adjustment. According to the results, the Random Forest model outperformed the other algorithms with 99.01% accuracy, 99.17% sensitivity, 98.68% specificity and 98.92% F1 score and was determined as the most reliable model for clinical applications.

Vu et al. [27] developed a machine learning framework for early diagnosis of thyroid nodule malignancy. Using 1232 nodule data obtained from 724 patients of the Chinese Medical University, a dataset containing 19 features including age, gender, ultrasound findings and blood tests was used. Models such as Logistic Regression, Random Forest, AdaBoost, Gaussian Naive Bayes and Decision Tree were trained with hyperparameter setting and performance was improved with ensembl methods (stacking and voting). The best result was obtained with Voting method with 85.52% accuracy, 87.32% AUROC, 83.37% precision and 89.54% recall. The study shows that combining quality data processing and ensemble models can significantly improve the accuracy of thyroid cancer diagnosis.

Islam [28] proposed a stacking model based on XGBoost and MLP to predict thyroid diseases. A dataset of 9172 instances and 31 features from the UCI ML Repository was used; SMOTE-ENN hybrid sampling was applied to eliminate data imbalance, RFE was applied to identify important features, and model explainability was supported by SHAP. The outputs produced by the base classifiers (MLP, XGBoost) are combined with the logistic regression-based meta-classifier to obtain the final prediction. According to the results, the proposed model outperformed classical models such as Random Forest, SVM and AdaBoost with 99.78% accuracy, 99.80% F1 score, 99.79% precision and 99.78% recall. SHAP analysis clearly explained the contribution of features such as T3, TT4, FTI and TSH to model decisions.

Bharath and Sabitha [29] compared six different machine learning algorithms to predict the probability of recurrence in differentiated thyroid cancer (DTC) patients and integrated the best model with a web-based decision support system. They used 383 samples from the UCI dataset and 16 clinicopathological features, including age, gender, radiotherapy history, TNM stages and risk classifications. Ordinal and one-hot coding was applied to the data, split with a 60% training to 40% testing ratio, and models (LR, DT, RF, SVM, KNN, XGBoost) were trained using Scikit-learn and XGBoost. The best result was obtained with XGBoost algorithm with 98.05% accuracy, 97.83% precision and 95.74% recall. By integrating the model with Flask backend and HTML/JavaScript interface, an intuitive web interface was developed where users can upload test data and get real-time predictions.

1.4. Models Based on Genetic and Molecular Properties

Guo, et al. [30] examined whether thyroid cancer risk can be predicted by genetic markers. Five SNPs (rs965513, rs944289, rs116909374, rs966423, rs2439302) associated with papillary thyroid carcinoma (PTC) were genotyped in Han Chinese individuals and their predictive power was tested with nine different machine learning algorithms. The results showed that SNPs had significant statistical associations with PTC; however, AUC values remained between 0.54-0.60 and sensitivity rates (28-48%) were low. Furthermore, the familial risk contribution of these SNPs was only 5.98% and even the addition of variables such as age/sex did not significantly improve the prediction performance. The study emphasises that models based on SNPs alone are inadequate and holistic modelling that combines genetic, environmental and interactive factors is needed.

Rossing [31] evaluated the potential of molecular classifiers in differentiating papillary and follicular thyroid cancers. The diagnostic performance of models generated by microarray-based mRNA and miRNA expression analyses were compared. Jarzab et al.'s SVM-based 19-amplitude model showed 85.7% sensitivity and 100% specificity; Borup et al.'s 76-probe model showed 94.4% sensitivity and 95.5% specificity. In addition, miRNAs (miR-221, miR-222) were found to be particularly successful in FTC/PTC discrimination. These findings suggest that molecular classifiers have the potential to prevent unnecessary surgeries by increasing diagnostic accuracy, especially in cytopathologically unstable cases, but these models need further validation before clinical application.

1.5. Explainability (XAI), Bayesian Modelling, and Clinical Decision Support Applications

The thesis presented by Preez [32] at Stellenbosch University explores the potential use of Bayesian Neural Networks (BNNs) in thyroid cancer classification. As an alternative to conventional neural networks, which often lack robust uncertainty estimation, three different convolutional neural network (CNN) architectures—LeNet-5, AlexNet, and GoogLeNet—were implemented using Bayesian inference and optimized via the Bayes-by-Backprop method. The models were evaluated in terms of both aleatoric and epistemic uncertainty, with results reported separately for each. The dataset employed consisted of labeled ultrasound images retrieved from the Thyroid Digital Image Database and was subjected to pre-processing steps including region-of-interest (ROI) focusing, segmentation, and data augmentation. Although the Bayesian models achieved only a marginal improvement in accuracy compared to their classical counterparts, they provided substantial advantages in epistemic uncertainty quantification. These insights are particularly valuable in clinical decision support systems, where model confidence plays a critical role. In experiments conducted on normalized data, an accuracy rate of 94.5% was achieved. The study emphasizes the value of BNN-based approaches in reducing false positives and preventing overdiagnosis, particularly in early-stage thyroid cancer detection.

1.6. Systematic Reviews and Comprehensive Literature Surveys

The systematic review conducted by Lixandru-Petre et al. [33] comprehensively examines the application of machine learning techniques in the early diagnosis, metastasis detection, and recurrence prediction of thyroid cancer. A total of 1,231 studies published between 2014 and 2024 were retrieved from six major databases, of which 203 were reviewed in detail and 21 were deemed eligible for in-depth analysis. The review is structured around three primary themes: (1) malignancy classification, (2) metastasis prediction, and (3) survival/recurrence forecasting. Commonly utilized algorithms include Random Forest, XGBoost, SVM, MLP, Logistic Regression, and Naive Bayes, typically trained on datasets containing electronic medical records, clinical, biochemical, ultrasound, and genetic data (e.g., BRAF V600E mutations). Random Forest and XGBoost models stood out with high AUC scores, and the integration of SHAP analysis, SMOTE, and ensemble strategies was shown to enhance both model interpretability and predictive performance. While the review highlights the potential of ML-driven systems in clinical decision-making, it also draws attention to challenges such as class imbalance, ethical considerations, limited generalizability, and the need for explainability.

Anari et al. [34] conducted an extensive review of deep learning methods applied to thyroid cancer diagnosis. The study evaluates recent architectures introduced post-2018, including CNNs, GANs, Autoencoders, LSTM, Deep Belief Networks (DBNs), and RNNs. Model performances were compared using metrics such as accuracy, sensitivity, and specificity. Among the models reviewed, VGG16 achieved 99% accuracy and 94% sensitivity, while GAN-based models reached accuracy rates as high as 94.30%. Sequential models like LSTM and RNN also performed strongly, with reported accuracies exceeding 98%. The review identifies CNN-based systems as the most widely adopted and effective approach and notes that GANs offer enhanced classification performance when used with multimodal imaging inputs.

Ilyas et al. [35] performed a systematic review focusing on deep learning methods for thyroid cancer diagnosis using diverse medical imaging modalities. From a pool of 2,149 publications between 2017 and 2021, 40 studies were selected for detailed analysis. The algorithms investigated included CNNs, Inception, ResNet, VGG16, RCNN, Bi-LSTM, and ensemble models. The datasets utilized in these studies spanned both private and publicly available ultrasound, CT, DICOM, and JPEG-format images. Key evaluation metrics included sensitivity, specificity, accuracy, and AUC. Based on the compiled analysis

matrix, the average sensitivity was reported at 89.5% and specificity at 84.4%. CNN and ResNet architectures yielded the highest performance, with some models achieving accuracy rates up to 98%. The findings underscore the reliability of deep learning in thyroid nodule classification, especially when supported by large, annotated datasets.

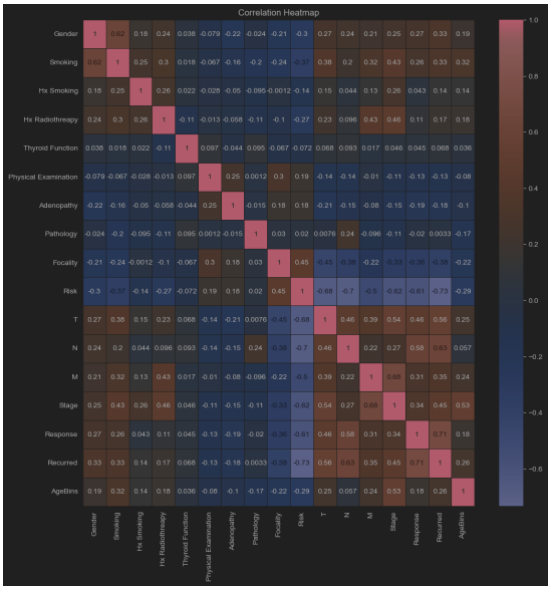
Habchi et al. [36] provide a detailed review of artificial intelligence (AI) techniques employed in thyroid cancer diagnosis, covering classification, segmentation, and prediction tasks. The review discusses supervised methods (CNN, SVM, MLP, RBF, Logistic Regression), unsupervised approaches (k-means, PCA), deep learning techniques (DAE, RNN, GAN), and ensemble strategies (bagging, boosting) in depth. Public datasets such as DDTI, TCGA, SEER, GEO, and ThyroidOmics, along with private clinical repositories, were analyzed. CNN and RNN-based models were found to deliver accuracy rates approaching 98%, while methods like XGBoost, AdaBoost, and Bagging enhanced sensitivity and specificity through ensemble integration. The study highlights AI's contributions to diagnostic accuracy, time efficiency, and reduced inter-observer variability but also acknowledges major challenges, including ethical concerns, data privacy, and the need for clinical generalizability.

2. Material and Methods (Materyal ve Yöntem)

2.1. Datasets

In this study, five distinct and content-rich datasets were utilized to develop machine learning models aimed at the diagnosis, risk analysis, and classification of thyroid disorders. Each dataset was derived from different clinical contexts and patient populations, encompassing biochemical, symptomatic, and demographic attributes. Prior to model training, all datasets underwent preprocessing procedures including missing data analysis, categorical encoding, and class imbalance mitigation.

The first dataset (TD1) [37] comprises clinical and treatment-related records of hundreds of patients diagnosed with well-differentiated thyroid cancer. This dataset includes variables such as age, sex, tumor staging information (T, N, M), ATA risk classification, and treatment response. It was employed in the development of explainable machine learning algorithms focused on recurrence prediction. Since the dataset was complete with no missing values, minimal preprocessing intervention was required.



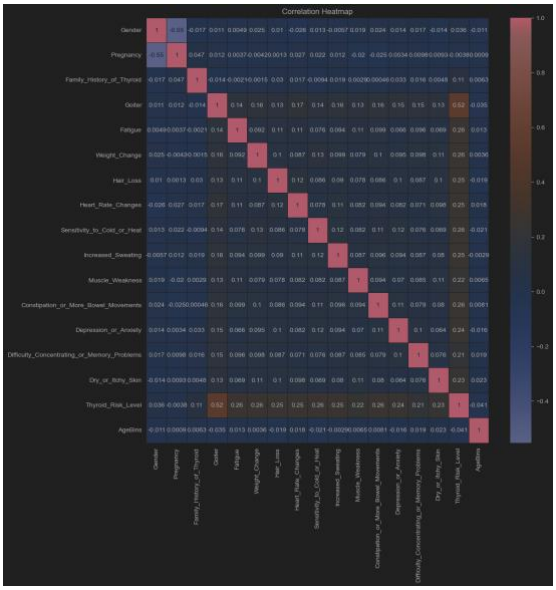


Figure 1(Left) & Figure 2(Right): TD1(Left) & TD2(Right) Correlation Matrixes

The correlation matrix for the first dataset (TD1) is presented in Figure 1. It provides a general overview of the linear relationships among the variables. Several variable pairs exhibit strong positive or negative correlations, indicating the presence of recurring structural patterns within the dataset. In contrast, features with low correlation coefficients may be considered as independent variables that can contribute unique information during the modeling process.

The second dataset (TD2) [38] is a balanced, resampled dataset designed for symptom-based thyroid

disorder prediction. It includes both subjective symptoms—such as fatigue, hair loss, changes in heart rhythm, constipation, and depression—and biochemical indicators. Encoded through numerical transformation, the data serve as a strong foundation for symptomatic classification of thyroid dysfunctions. Class imbalance was mitigated using SMOTE-based oversampling techniques.

The correlation structure of the second dataset is shown in Figure 2. In general, it reveals weak linear relationships and suggests a more scattered and independent feature composition. The predominance of low correlation coefficients implies that the variables carry relatively independent information and that the dataset is more suitable in terms of low multicollinearity. This structure indicates a reduced risk of information redundancy in the modeling phase.

The third dataset (TD3) [39] contains biochemical measurements and clinical histories of patients diagnosed with hypothyroidism. It includes detailed medical variables such as hormone levels (TSH, T3, TT4), clinical background, medication history, surgical interventions, pregnancy, and psychiatric records. This dataset, which has no missing values, was evaluated using high-performance supervised learning algorithms for hypothyroidism classification. It provides a solid foundation for modeling nonlinear relationships based on hormone profiles.

The correlation matrix for the third dataset is provided in Figure 3. It reveals that linear associations across variables are generally weak. However, a limited number of variable pairs exhibit moderate correlations. This structure suggests that most features convey independent information, although certain groups may demonstrate structural clustering with shared variance. The predominance of weak correlations indicates a stable variable structure with minimal risk of multicollinearity.

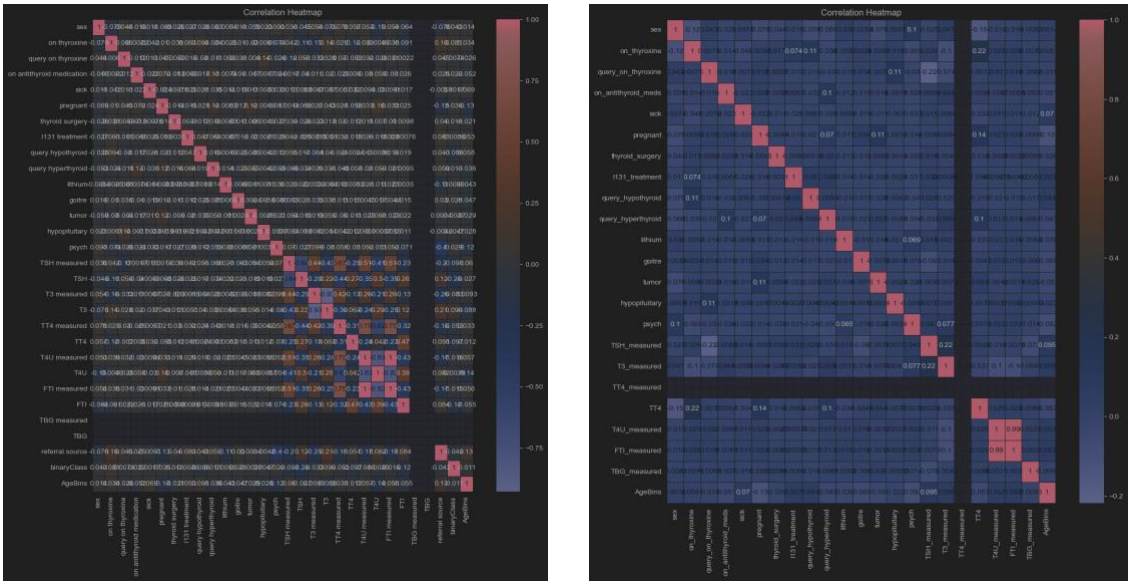


Figure 3(Left) & Figure 4(Right): TD3(Left) & TD4(Right) Correlation Matrixes

The fourth dataset (TD4) [40] consists of multidimensional clinical data, incorporating hormone profiles alongside patient symptoms, treatment history, and diagnostic variables. Containing tens of thousands of records, this dataset stands out due to its relatively high rate of missing values, particularly concentrated in critical variables such as hormone measurements. Accordingly, various missing data imputation techniqueswere applied. This dataset was primarily used in modeling processes aimed at predicting the impact of hormonal levels on thyroid dysfunction.

The correlation matrix for the fourth dataset is presented in Figure 4. It exhibits a structurally sparse and balanced distribution with predominantly low-level correlations. Nevertheless, noticeable correlation blocks were observed among certain variable groups. Specifically, significant linear relationships were detected between some measurement-based variables, suggesting tendencies toward shared variance. This indicates that potential redundancy in the dataset may be localized within specific subdomains and should be carefully evaluated prior to modeling.

The fifth dataset (TD5) [41] represents a large-scale population-based sample comprising records from

hundreds of thousands of individuals. It includes demographic characteristics (age, sex, ethnicity), lifestyle factors (smoking, obesity, iodine intake), family history, and various hormone measurements. All of which are associated with the risk of developing thyroid cancer. Due to its high dimensionality and volume, this dataset was primarily utilized for risk scoring and segmentation modeling purposes.

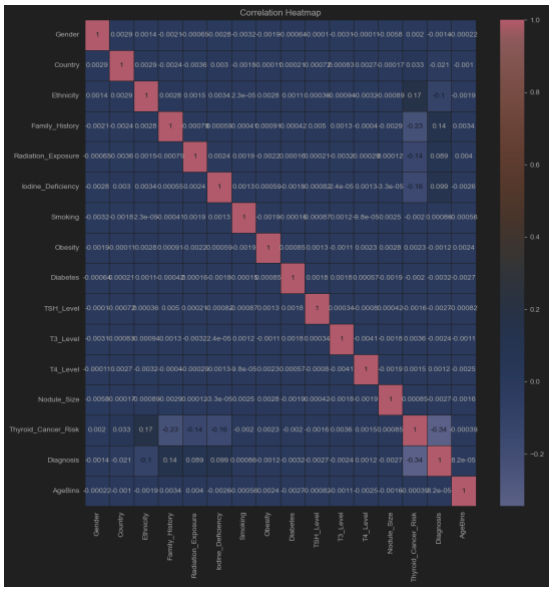


Figure 5: TD5 Correlation Matrix

Figure 5 illustrates the correlation structure of the TD5 dataset, which is composed of a large-scale population-based sample including demographic attributes (age, sex, ethnicity), lifestyle factors (e.g., smoking, obesity, iodine intake), family history, and multiple biochemical markers. The correlation matrix reveals that the majority of variable pairs exhibit very weak linear associations, with correlation coefficients close to zero. This pattern suggests that the features in TD5 carry largely independent information, supporting a multicollinearity-free environment for model training.

A small number of moderate correlations were observed, such as those between age and thyroid hormone levels (e.g., TSH), and between lifestyle factors and clinical measurements, which may reflect domain-specific physiological linkages. These moderate associations were further examined during feature importance analysis to ensure they did not introduce redundancy. Overall, the sparse correlation pattern in TD5 provides a favorable foundation for training robust and generalizable machine learning models, as it reduces the risk of collinearity-induced performance distortion.

The correlation matrix corresponding to the fifth dataset illustrates a pattern characterized predominantly by weak linear associations and low collinearity. Most variables exhibited correlation coefficients close to zero, indicating that the dataset comprises features carrying highly independent information. Nevertheless, a limited number of variable pairs demonstrated moderate correlations, which may reflect thematic linkages or shared variance structures within the data. Overall, this dataset presents a multicollinearity-free structure, offering a flexible and robust foundation for model development.

When considered collectively, these datasets provide a holistic representation of the biochemical, clinical, and symptomatic dimensions of thyroid diseases. They enable the development of models with high accuracy, generalizability, and clinical applicability. All datasets were obtained from publicly available resources on Kaggle.

2.2. Data Preprocessing

The success of machine learning models is not solely determined by algorithm selection, but also critically depends on the quality of the input data. Accordingly, an extensive preprocessing pipeline was implemented on the medical datasets used in this study. The objective was to minimize data-related artifacts, reduce irrelevant variance, and address class imbalance—factors known to adversely affect model learning.

2.2.1. Handling Missing Values

All datasets were carefully examined for missing or inconsistent values. Where necessary, affected samples were either excluded or completed using appropriate imputation methods such as mean or median filling. This process helped preserve the overall accuracy of the models while preventing artificial data distortions during learning.

2.2.2. Feature Scaling and Normalization

Given that certain algorithms (e.g., SVM, ANN) require features to be on a comparable scale, Min-Max normalization was applied to all numerical variables. This transformation scaled the features into a [0,1] range, thereby eliminating the influence of outlier values and ensuring stable convergence during training.

2.2.3. Addressing Class Imbalance

Class imbalance, a common issue in medical datasets, was also observed in this study. To mitigate the underrepresentation of minority classes, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. By synthetically generating new instances of minority classes, SMOTE helped achieve a balanced class distribution and contributed to improved recall and F1 scores, particularly in imbalanced classification tasks.

2.2.4. Vector Transformation and Matrix Structuring

Before initiating the model training process, all features were vectorized and structured using a standardized transformation schema that was consistently applied in both training and testing phases. Variables were reformatted into processed representations to ensure uniform model input. This approach eliminated data leakage and enhanced the reliability of cross-validation results.

This preprocessing framework ensured the structural and statistical integrity of the datasets while minimizing noise. As a result, the machine learning models were trained solely on informative, clean representations of the data, allowing performance metrics to reliably reflect the true quality of the input features.

2.2.5. Hyperparameter Optimization

The performance of machine learning algorithms depends not only on data quality or model selection, but also critically on the proper tuning of algorithm-specific hyperparameters. Therefore, systematic hyperparameter search procedures were employed in this study. Parameter combinations were evaluated based on both overall accuracy and class-specific performance metrics.

Two distinct strategies were adopted for hyperparameter tuning:

- For some models, empirically determined fixed values were used.
- For others, hyperparameters were optimized through validation-based optimization techniques.

In particular, for gradient boosting decision tree (GBDT) models and boosting-based frameworks, the Optuna library was utilized for hyperparameter optimization. Within this framework, the objective function was defined directly over model output, and performance was optimized on 5-fold stratified cross-validation at each trial. This approach enabled the models to reach more flexible decision boundaries and enhanced generalizability. The selection of optimal parameters was guided not only by accuracy, but also by F1 score and ROC-AUC values. The flowchart in Figure 6 presents the structured preprocessing stages conducted prior to model training. It begins with data loading and continues through missing value imputation, categorical encoding, scaling, class balancing using SMOTE, and final transformation into model-ready matrix formats. These standardized steps ensured consistency and comparability across heterogeneous datasets.

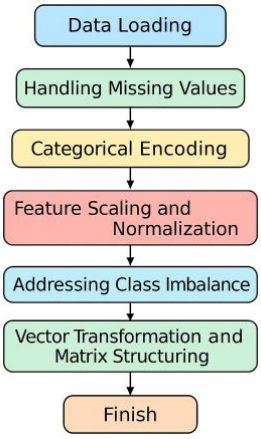


Figure 6. Workflow Diagram of the data preprocessing applied across all five thyroid datasets

2.3. Algorithms

In this study, a diverse set of supervised machine learning algorithms was employed for the diagnosis, risk assessment, and classification of thyroid disorders. These algorithms are built upon different statistical assumptions and mathematical foundations. The selection of models was influenced by the challenges typical to biomedical data, including class imbalance, high dimensionality, missing values, and the need for both predictive accuracy and model explainability in clinical decision-making contexts.

The selected algorithms are primarily designed to perform classification tasks, aiming to correctly assign data samples into predefined categories. Each algorithm establishes a decision boundary based on independent variables to infer the target class. These decision-making mechanisms range from linear discriminant functions to deep neural architectures and ensemble models such as random forests. This variety was intentionally chosen to accommodate the structural properties of different datasets and to maximize overall performance.

From a mathematical standpoint, the algorithms were evaluated using statistical measures such as accuracy, sensitivity, specificity, and F1-score [42][43]. To assess model performance, cross-validation and resampling techniques were employed, particularly to prevent overfitting in the presence of imbalanced class distributions. In addition, feature selection and dimensionality reduction were implemented to minimize model complexity and optimize computational efficiency.

Each algorithm presented in the following sections is discussed in terms of its theoretical framework, decision function, parameter configurations, and the impact of related hyperparameters. Moreover, the applicability of each model is evaluated not only from a statistical accuracy perspective, but also in terms of its potential integration into medical decision support systems. The goal is not solely to identify the highest-performing model, but to define models capable of producing reliable and interpretable decisions in real-world clinical scenarios.

2.3.1. Artificial Neural Network (ANN)

Artificial Neural Networks (ANN) are computational models inspired by the neuronal architecture of the human brain and consist of multiple layers of artificial neurons [44]. In this study, the ANN model was trained under the supervised learning paradigm to perform classification tasks related to thyroid disorders. The primary objective of the model is to learn nonlinear relationships between input features and target classes, thereby enabling high-accuracy predictions on previously unseen samples.

The ANN architecture consists of an input layer, one or more hidden layers, and an output layer [45]. In the input layer, each variable in the feature vector is represented by a distinct neuron. Neurons in the hidden layers are equipped with nonlinear activation functions commonly the Rectified Linear Unit (ReLU) in this study which allow the network to model complex decision boundaries. The output layer uses activation functions suitable for classification tasks, such as sigmoid or softmax, and the outputs are expressed as class probabilities.

During training, the cross-entropy loss function was used to measure the prediction error. The weights of the network were updated via the backpropagation algorithm to minimize this loss. Optimization was typically performed using Stochastic Gradient Descent (SGD) or the Adam optimizer [46]. To prevent overfitting, regularization techniques such as dropout and early stopping were applied.

In this study, the ANN was implemented on high-dimensional and multivariate thyroid datasets and produced meaningful classification results even in cases involving complex feature interactions. These outcomes demonstrate ANN's capacity to learn nonlinear patterns in the data and highlight its potential applicability in clinical decision support systems. However, the limited explainability of ANN models and their sensitivity to hyperparameter tuning represent important limitations that must be carefully considered in terms of interpretability and generalizability. The input to the artificial neural network is a vector x∈R where each component of the vector corresponds to a specific input feature. This is formally shown in Equation (1):

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \text{^}(T) \tag{1}$$

Each hidden layer receives the output from the preceding layer, applies a linear transformation through weight matrices and bias terms, and then introduces nonlinearity via an activation function. The two hidden layers constructed in this study are formally defined in Equations (2) and (3):

$$h_1 = \phi(W_1 x + b_1) \tag{2}$$

$$h_2 = \phi(W_2 h_1 + b_2) \tag{3}$$

In the final layer of the model, a linear combination is computed using the output of the previous hidden layer, followed by the application of a sigmoid (or softmax) activation function. The resulting value represents the classification probability. The sigmoid function is typically used in binary classification tasks, whereas the softmax function is preferred for multi-class problems. Equation (4) represents the output layer using a sigmoid activation function:

$$\hat{y} = \sigma(W_L h_{\mathrm{L}-1} + b_L) \tag{4}$$

To quantify the difference between the model's prediction and the true label, binary cross-entropy is used. This loss function serves as the objective to be minimized during the learning process. The corresponding formulation is given in Equation (5):

$$\mathcal{L} = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})) \tag{5}$$

Using derivatives computed through backpropagation, the network's weights and bias terms are updated. This process adjusts the parameters in a direction that reduces the loss function at each iteration. Equations (6) and (7) respectively define the weight and bias update rules:

$$W_l \leftarrow W_l - \eta \frac{\partial \mathcal{L}}{\partial W_l} \tag{6}$$

$$b_l \leftarrow b_l - \eta \frac{\partial \mathcal{L}}{\partial b_l} \tag{7}$$

2.3.2. CatBoost

CatBoost is a gradient boosting-based machine learning algorithm that stands out for its high predictive accuracy and low susceptibility to overfitting [47], particularly in datasets with numerous categorical variables. Developed by Yandex, CatBoost is specifically designed to minimize the need for extensive feature engineering, enabling direct modeling of datasets where categorical features are prevalent. In

this study, the CatBoost algorithm was employed to classify the risk levels associated with thyroid disorders and was able to effectively learn complex relationships among features in the dataset.

Unlike traditional gradient boosting algorithms, CatBoost utilizes a sequential learning strategy known as ordered boosting [48]. This approach mitigates target leakage by carefully controlling the order in which training samples are used during tree construction. Furthermore, instead of using random target mean encoding for categorical variables, CatBoost employs ordered statistics, which improves the model's generalization capability. The mathematical formulation underlying the CatBoost prediction function is defined in Equation (8):

$$\hat{y} = \sum_{m=1}^{M} \gamma_m \cdot h_m(x_i) \tag{8}$$

The log-loss function used by the CatBoost algorithm for classification tasks is presented in Equation (9). This loss function quantifies the divergence between predicted probabilities and actual class labels and serves as the objective function during training:

$$\mathcal{L} = -\sum_{i=1}^{n} (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \tag{9}$$

2.3.3. LightGBM

LightGBM is a memory-efficient and computationally optimized algorithm based on the gradient boosting framework over decision trees [49]. It is particularly effective for large-scale and high-dimensional datasets, where it significantly reduces training time while improving predictive accuracy. Unlike traditional Gradient Boosted Decision Tree (GBDT) algorithms, LightGBM is characterized by its histogram-based tree construction and leaf-wise growth strategy, which together enhance both performance and scalability [50].

In this study, LightGBM was employed for the classification of thyroid diseases and demonstrated notable performance in terms of accuracy, class discrimination capability, and overall learning efficiency. The model's parameters were carefully configured, taking into account class imbalance and the multivariate nature of the data. As a result, statistically significant outcomes were achieved at the end of the training process.

LightGBM approximates the prediction function $f(x)$ using an ensemble of sequential decision trees. Each m-th tree is constructed to correct the residual errors of the previous trees. The general formulation of the model is given in Equation (10):

$$\hat{y} = \sum_{m=1}^{M} f_m(x_i) \tag{10}$$

The model's loss at each iteration is defined as shown in Equation (11). This formulation represents the objective function that is minimized during training through gradient-based optimization:

$$\mathcal{L}^m = \sum_{i=1}^{n} l(y_i, \hat{y}_i + f_m(x_i)) + \Omega(f_m) \tag{11}$$

In this study, the LightGBM model was positioned as a highly generalizable classifier with strong discriminative power between classes, particularly in handling complex and multi-dimensional thyroid data. The model was configured by tuning hyperparameters such as tree depth, number of leaves, and learning rate. Its performance was optimized to reduce the impact of class imbalance.

2.3.4. Random Forest

Random Forest is a supervised learning algorithm based on the principle of ensemble learning, combining multiple decision trees to improve predictive performance [51]. It is widely known for its high accuracy and robustness against overfitting. This approach aggregates the predictions of a large number of decision trees, each trained on a random subset of the data and features, and final classification is made via a majority voting mechanism.

In this study, the Random Forest algorithm was implemented for the classification of thyroid disorders and trained using a large ensemble of 1000 decision trees. The model yielded balanced results even under varying class representations, making it a preferred choice due to its limited tendency to overfit. The incorporation of randomness throughout the training process served as a key factor in enhancing the model's generalizability.

A Random Forest model consists of M decision trees. Each tree is independently trained on a bootstrap-sampled subset of the data. The final classification decision is made by majority voting across all trees [52]. This decision mechanism is formally expressed in Equation (12):

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \cdots, h_M(x)\} \tag{12}$$

Each individual decision tree within the Random Forest is constructed following the procedure below:

- A bootstrap sample is drawn randomly from the training dataset.
- At each node, only a random subset of features is considered for splitting.
- The tree is grown to its full depth without pruning.

The primary advantage of this structure lies in maintaining low correlation among trees, thereby improving the ensemble's generalization capability and reducing variance.

In this study, the Random Forest model was implemented with a forest comprising 1000 decision trees. The model was capable of evaluating complex multivariate patterns in the thyroid dataset in parallel, reducing the local biases of individual trees and producing more stable classification outcomes. The results in terms of accuracy, error-free class prediction, and model stability demonstrate that Random Forest provides a robust foundation for medical classification problems.

2.3.5. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful algorithm designed for both linear and non-linear classification tasks [53], with a strong capability to distinguish between classes. Its core objective is to find the optimal decision boundary that maximizes the margin between different class distributions. This property contributes to the model's high generalizability and robustness in classification problems.

In this study, the SVM algorithm was applied to the thyroid dataset, and the Radial Basis Function (RBF) kernel was selected to handle cases where the classes are not linearly separable in the input space. By projecting the input data into a higher-dimensional feature space, the RBF kernel enables linear separation in transformed dimensions. The model was configured with hyperparameters C=1.0 and γ=0.1 achieving a balance between classification margin width and error sensitivity. SVM learns a hyperplane defined by the equation w^Tx+b=0, which maximizes the margin between classes [54]. The optimization problem to be solved is expressed in Equation (13):

$$minimize\ \frac{1}{2} \parallel w \parallel^2 \tag{13}$$

For datasets that are not linearly separable, kernel functions are employed to map the input data into a higher-dimensional feature space where linear separation becomes feasible. In this study, the Radial Basis Function (RBF) kernel was utilized, and it is defined in Equation (14) as follows:

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel)^2 \tag{14}$$

This function measures the similarity between two instances and enables the model to construct non-linear decision boundaries. Smaller values of γ produce smoother decision surfaces, whereas larger values make the model more sensitive to individual data points.

The SVM model was specifically designed to separate overlapping class distributions where clear linear separation was not possible. The use of the RBF kernel enabled the transformation of the input space into higher-dimensional representations, allowing the construction of an effective decision boundary for non-linear structures. The selected values for C and γ were optimized to balance model complexity and error tolerance. As a result, the SVM provided robust separation between classes in the thyroid dataset, achieving high accuracy and low overall error.

2.3.6. Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized and regularized extension of the Gradient Boosted Decision Trees (GBDT) framework [55]. This algorithm incorporates second-order derivative information, regularization terms, and parallel computation strategies to improve model generalizability. It is particularly effective for large-scale and irregular datasets, offering high accuracy and resistance to overfitting.

In this study, the XGBoost algorithm was employed to classify thyroid diseases and was optimized to learn the complex relationships among various features. The model's hyperparameters were empirically tuned, and the implementation was carried out using the XGBClassifier interface. The resulting classification performance was found to be satisfactory in terms of both accuracy and class separation.

XGBoost constructs M weak learners sequentially, with each learner aiming to minimize the residual error of the previous predictions [56]. The general form of the model is presented in Equation (15):

$$\hat{y} = \sum_{m=1}^{M} f_m(x_i) \tag{15}$$

The regularized objective function optimized at each boosting iteration is defined in Equation (16). This function balances the model's predictive accuracy with its complexity by incorporating both the loss term and a regularization penalty Equation(17):

$$\mathcal{L}^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i + f_t(x_i)) + \Omega(f_t) \tag{16}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{17}$$

The model was applied to a high-dimensional and feature-rich thyroid dataset and demonstrated superior classification accuracy, particularly in cases with sharp inter-class boundaries. By optimizing both the secondary loss and regularization terms, the model's tendency to overfit was mitigated, while complex inter-feature dependencies were effectively captured. In this respect, XGBoost has been recognized as a core component of high-performance and interpretable decision support systems.

2.3.7. Stacking Classifier

Stacking, or Stacked Generalization, is an ensemble learning approach that integrates multiple machine learning models under a unified architecture, leveraging the individual strengths of each to enhance predictive performance [57]. In this framework, the outputs of the base learners are used as input features for a meta-model, which is responsible for producing the final classification output.

In this study, the stacking model was constructed using LightGBM and XGBoost as base learners, while Logistic Regression served as the meta-learner. This configuration was particularly effective for medical datasets characterized by class imbalance and non-linear decision boundaries, allowing for a more flexible and robust classification scheme by combining learners with distinct decision surfaces.

The architecture comprises two levels, and the ensemble prediction function is defined in Equation (18):

$$\hat{y} = h_{\text{meta}}(h_1(x), h_2(x), \ldots, h_K(x)) \tag{18}$$

2.4. Model Evaluation Metrics

The success of machine learning models should not be evaluated solely based on overall accuracy but also by considering class-specific performance metrics[58]. This is particularly critical in medical datasets where class imbalance is common, and assessing the model's ability to correctly classify rare cases is essential. Accordingly, this study employed a variety of evaluation metrics, including Accuracy, Precision, Recall, F1-Score.

2.4.1. Accuracy

Accuracy measures the proportion of correctly classified instances among all samples. Although it is often the primary reference metric, it may be misleading in problems with imbalanced class distributions. The calculation of accuracy is defined in Equation (19):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{19}$$

2.4.2. Precision

Precision quantifies how many of the instances predicted as positive are actually true positives. It is especially important in contexts where the cost of false positives is high. Precision tends to decrease when the number of false positives increases. The formula for calculating precision is provided in Equation (20):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

2.4.3. Recall

Recall, the proportion of actual positive cases that are correctly identified by the model, is a critical metric in medical diagnosis where missing true cases can have serious consequences. It reflects the model's sensitivity and is particularly important when false negatives must be minimized. The calculation is shown in Equation (21):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

2.4.4. F1-Score

F1-Score is the harmonic mean of Precision and Recall, and it is used to balance both metrics, particularly in cases with imbalanced class distributions. It is a robust measure when both false positives and false negatives are important. The calculation is given in Equation (22):

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

The performance metrics discussed above were computed individually for each algorithm and served as primary references in the comparative analysis. Among these, Accuracy, F1-Score, Precision, and Recall were prioritized in the evaluation of decision support systems due to their robustness in the presence of class imbalance, a common challenge in medical datasets.

2.5. Model Validation Methods

In multi-class and imbalanced medical datasets, evaluating machine learning models requires more than conventional performance indicators. It must also incorporate assessments of model stability and generalizability. Therefore, this study adopted systematic cross-validation techniques that go beyond standard validation, maintaining class distribution while enabling performance evaluation across multiple sampling structures.

2.5.1. Stratified K-Fold Cross-Validation

As the primary validation strategy, Stratified K-Fold Cross-Validation was employed to ensure representative class distribution within each fold [59]. This method divides the dataset into k subsets (folds) while preserving the proportion of each class in every subset, in alignment with the original data distribution. In each iteration, one fold is used for testing while the remaining $k-1$ folds are used for training. This approach minimizes both intra-class and inter-class variance, providing a more reliable assessment of model performance.

In this study, all algorithms were validated using a 5-fold stratified cross-validation structure (k=5), enabling model performance to be measured repeatedly across different sample configurations. This method has proven particularly effective for health data scenarios where minority class representation is critical.

2.6. Systematic Performance Assessment via Cross-Validation Score

Throughout the validation process, Cross-Validation Scores were actively utilized to conduct comparative analyses both across algorithms and within hyperparameter configurations. For each fold, core evaluation metrics such as Accuracy, F1-Score, and ROC-AUC were calculated individually. Final model decisions were made based on the mean performance across folds, providing a statistically robust basis for selecting optimal model configurations.

- Accuracy
- Recall
- F1 Score
- Precision

Cross-validation scores enabled the evaluation of each model under multiple independent testing scenarios, rather than relying on a single train-test split. This approach assessed whether the models produced stable and consistent results, and models exhibiting high variance or instability were either excluded or subjected to hyperparameter adjustments.

This study focuses on the systematic application and evaluation of advanced classification approaches on high-dimensional and imbalanced medical datasets, with the aim of enhancing the reliability of clinical decision support systems. The applied methodologies were not only designed to improve classification accuracy, but were also optimized for model stability, explainability, and generalizability.

By implementing multiple models based on distinct mathematical principles, both linear and nonlinear patterns in the data were effectively captured. The flexibility of the decision boundaries was enhanced through algorithmic diversity, and the inclusion of ensemble and meta-modeling strategies produced more robust classification outcomes beyond individual learners.

The model evaluation framework was not limited to overall accuracy but was grounded in a multidimensional set of metrics, including class-wise sensitivity, specificity, F1-score, precision, and recall. This approach enabled a more balanced interpretation of performance, especially in detecting minority classes, where traditional metrics often fall short.

Moreover, validation mechanisms extended beyond conventional data partitioning strategies. They incorporated cross-validation structures that preserved class distribution and minimized variance. These mechanisms were directly integrated into hyperparameter tuning workflows, ensuring not only theoretical soundness but also practical reliability of the resulting models.

In conclusion, this multi-layered approach has provided a holistic framework for improving model performance, reducing instability, limiting overfitting risks, and enhancing inter-class discrimination. This structure is not only applicable to medical classification problems, but also represents a scalable and reproducible solution for a wide range of data analytics applications with similar structural characteristics.

3. Results and Discussion (Sonuçlar ve Tartışma)

This study aimed to address key challenges in classification problems—namely generalizability, statistical stability, and inter-class discrimination performance—through an integrated and systematic approach. Rather than focusing solely on output generation, the modeling process was structured around a methodologically validated, multi-metric evaluation framework that ensured analytical rigor. The framework, developed on datasets with diverse distribution characteristics and class imbalances, went beyond conventional accuracy-based strategies and established a comprehensive, balanced, and generalizable analytical foundation. The overall study workflow is illustrated in Figure 7.

Instead of the traditional train-test split, a systematic K-fold cross-validation strategy was applied to each dataset, enabling the statistical reliability of model performance to be assessed across different sample partitions. The Stratified K-Fold structure was employed to minimize measurement bias caused by class imbalance by preserving class proportions within each fold. This validation strategy aimed not only to achieve high accuracy but also to ensure that such performance could be consistently sustained across various data partitions.

Furthermore, model performance was tested across multiple sample configurations, rather than relying on a single split, allowing for the evaluation of repeatability and robustness. This strategy did not accept good performance on a single subset as sufficient, but rather prioritized models that could maintain similar accuracy levels across diverse sampling distributions. As a result, the evaluation framework increased model resilience to real-world uncertainties and improved the overall reliability of the findings.

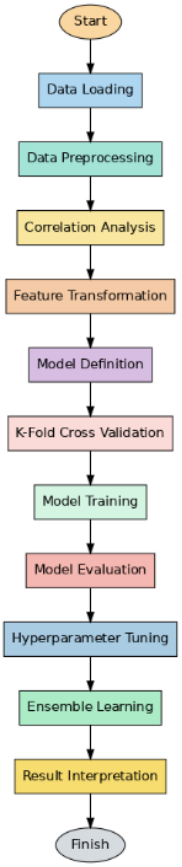


Figure 7: Workflow of the Study

The study integrated not only basic classifiers but also parametrically optimized models, ensemble learning methods, and meta-modeling architectures. For parameter-sensitive models, structured search techniques were implemented; however, these were positioned not merely as accuracy-boosting tools but as support mechanisms for systematic evaluation. The inclusion of meta-classifiers in ensemble structures contributed to more balanced decision surfaces, particularly in cases with challenging class separability.

Model performance was evaluated not only through overall accuracy but also using class-specific metrics such as sensitivity, precision, and F1-score. This multi-metric approach enabled more context-aware interpretations, especially in detecting underrepresented classes where traditional accuracy might fail. In scenarios involving class imbalance, the evaluation process prioritized metrics that treated each class equitably, preventing bias toward the majority class.

Additionally, the outputs of certain classifiers were not directly used for final decisions; instead, they were fed into higher-level meta-learners. This meta-modeling strategy served as a stabilizing mechanism in scenarios where class discrimination performance was relatively low. The ensemble structure allowed the weaknesses of individual base learners to be mitigated by the meta-classifier, resulting in more robust and consistent classification outcomes. Consequently, the system was enhanced not only in terms of accuracy but also in decision consistency and output stability.

The study also evaluated the capacity of models to correctly identify negative classes by analyzing secondary metrics derived from confusion matrices. Metrics such as specificity were included to assess the impact and spread of false classifications, thus improving the operational applicability of the models. In this sense, the evaluation framework addressed not only technical accuracy but also practical implications required in decision support systems that demand high sensitivity and low error tolerance.

Following this comprehensive evaluation, the quantitative performance outcomes of all applied models across the five datasets are presented comparatively in Tables 2 through 18.

Table 1: Performance Metrics on Training Data for CatBoost Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.996732	0.997701	0.995413	0.996747
TD2	0.963708	0.963709	0.964185	0.963708
TD3	0.997863	0.998840	0.999670	0.999820
TD4	0.963736	0.947180	0.965053	0.963736
TD5	0.674885	0.569707	0.759580	0.674885

Table 2: Performance Metrics on Test Data for CatBoost Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.987013	0.986893	0.987233	0.987013
TD2	0.830637	0.829942	0.830134	0.830642
TD3	0.989404	0.989230	0.989253	0.989404
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.65648	0.540073	0.579687	0.656480

The performance of the CatBoost algorithm across five datasets with varying structural characteristics and class distributions provides meaningful insights into the model's generalizability and stability. In particular, for TD1 and TD3, the model consistently achieved accuracy rates above 98% in both training and test phases, indicating high robustness. For these datasets, precision, recall, and F1-score values were closely aligned, demonstrating that the model was capable of distinguishing positive and negative classes with comparable effectiveness. Similarly, in TD4, the training and test performances were nearly identical, suggesting that CatBoost maintained resilience against overfitting and exhibited a stable decision boundary across different data partitions. The training phase results for CatBoost are provided in Table 1.

In contrast, a noticeable decline in test performance was observed for TD2 and TD5, highlighting the algorithm's sensitivity to data complexity and class imbalance. As shown in Table 2, while the model achieved high accuracy during training on TD2, its performance dropped by approximately 13% on the

test set, indicating a risk of overfitting. In the case of TD5, both training and test metrics were relatively low, implying that the dataset presents inherent modeling challenges. Nevertheless, the limited gap between training and test scores indicates that the model retained a reasonably stable learning capacity despite data irregularities.

In conclusion, CatBoost proved to be a strong performer on high-quality, well-structured datasets, offering both accuracy and interpretability. However, its effectiveness diminishes under imbalanced or noisy conditions, where classification complexity increases. Still, the model's resilience in class-level performance metrics highlights its potential as a reliable choice for classification tasks in structured clinical data environments.

Table 3: Performance Metrics on Training Data for XGBoostClassifier Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.964052	0.955365	0.967716	0.944830
TD2	0.794384	0.793492	0.801766	0.794418
TD3	0.98641	0.951073	0.963355	0.939546
TD4	0.962404	0.943966	0.926221	0.962404
TD5	0.659569	0.532136	0.795919	0.659569

Table 4: Performance Metrics on Test Data for XGBoostClassifier Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.987013	0.982213	0.991525	0.973684
TD2	0.754995	0.753243	0.760660	0.754838
TD3	0.980132	0.927071	0.945708	0.910206
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.660312	0.532971	0.569677	0.660312

The performance outputs of the XGBoostClassifier algorithm across five different datasets are presented in detail in Table 3 for training data and Table 4 for test data. The algorithm demonstrated particularly strong performance on TD1 and TD3, achieving high success in both training and testing phases. The training accuracies for these datasets were 96.4% and 98.6%, respectively, while test accuracies were sustained at 98.7% and 98.0%, respectively. The proximity of F1-score, precision, and recall values indicates that the model was able to distinguish between classes in a balanced manner and showed no signs of overfitting. A similar trend was observed for TD4, where consistency between training and test results further emphasized the model’s strong generalization capacity and high tolerance to varying sample distributions.

In contrast, both training and test success rates for TD2 and TD5 were relatively lower. In TD2, training accuracy dropped from 79.4% to 75.5% in the test phase, and for TD5, these values remained around 65.9% and 66.0%, respectively. The decline in F1-score and precision suggests that the model encountered difficulty in distinguishing between positive and negative classes in these datasets—likely due to class imbalance or limited feature representation. However, it is noteworthy that the gap between training and test performances for these two datasets remained minimal. This indicates that the model, despite its lower performance, achieved stable learning outcomes and effectively avoided overfitting. In conclusion, XGBoostClassifier emerges as a highly effective classifier when data structure is adequate, particularly distinguished by its balanced and robust class separation ability on high-quality datasets.

Table 5: Performance Metrics on Training Data for LightGBM Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.967716	0.963602	0.964413	0.964052
TD2	0.735364	0.732987	0.744687	0.735413
TD3	0.990719	0.990682	0.990657	0.990719
TD4	0.962404	0.943966	0.926221	0.962404
TD5	0.659457	0.531946	0.693710	0.659457

Table 6: Performance Metrics on Test Data for LightGBM Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.987013	0.986893	0.987233	0.987013

TD2	0.723597	0.720710	0.731995	0.723391
TD3	0.986755	0.986755	0.986755	0.986755
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.660359	0.583437	0.700079	0.666674

The training performance of the LightGBM algorithm is presented in Table 5. On TD1 and TD3, the obtained accuracy scores (0.9677 and 0.9907) as well as the corresponding F1-scores (0.9636 and 0.9907) and other related metrics indicate a highly effective class discrimination capability. The close proximity of precision and recall values suggests that the model does not exhibit any bias toward particular classes and is able to distinguish between positive and negative classes with consistent performance. TD4 also demonstrated similarly strong results, highlighting the model's generalizability across varying sample distributions. In contrast, the relatively lower performance on TD2 and TD5 suggests that limitations related to class structure or sample size may have constrained the model's effectiveness in these datasets.

The results on the test datasets, as reported in Table 6, generally align with the training phase performance. For TD1, TD3, and TD4, test accuracies were recorded as 0.9870, 0.9868, and 0.9621, respectively, indicating that the learned patterns during training were effectively generalized to unseen data. Although test performances for TD2 and TD5 were lower (0.7236 and 0.6603), this drop is attributed more to structural complexity of the datasets rather than overfitting. In TD2, the precision and recall values were measured as 0.7319 and 0.7234, respectively, demonstrating the model's maintained ability to recognize the positive class to a reasonable degree. Although TD5 showed lower F1-score and recall values, the small difference between training and test performance indicates that the model did not suffer from overfitting and retained limited yet stable generalizability. Overall, LightGBM proves to be a high-performing classifier in datasets with balanced structure and strong feature representation, offering both accuracy and class-wise discrimination.

Table 7: Performance Metrics on Train Data for ANN Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.986394	0.977351	0.979597	0.973001
TD2	0.525612	0.484455	0.515965	0.475965
TD3	0.986035	0.983140	0.987708	0.971646
TD4	0.987776	0.983558	0.977719	0.978444
TD5	0.656624	0.625898	0.640659	0.600043

Table 8: Performance Metrics on Test Data for ANN Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.968900	0.964237	0.969184	0.959332
TD2	0.496600	0.473289	0.486571	0.459876
TD3	0.961400	0.955826	0.961058	0.951147
TD4	0.965800	0.961533	0.963472	0.959811
TD5	0.643500	0.601729	0.627018	0.584220

The performance metrics of the Artificial Neural Network (ANN) model on the training datasets are presented in Table 7. The model demonstrated high classification accuracy on TD1, TD3, and TD4, with respective accuracy scores of 0.9864, 0.9860, and 0.9878, and consistently strong F1-scores above 0.97. These results indicate the model's strong ability to learn and represent the underlying patterns in well-structured datasets. In TD3, the close alignment of the accuracy (0.9860), F1-score (0.9831), and precision (0.9877) highlights the ANN's capacity to form robust decision boundaries. Similarly, in TD4, the narrow gap between precision (0.9777) and recall (0.9784) reflects balanced classification performance. In contrast, TD2 and TD5 showed relatively lower performance with accuracy scores of 0.5256 and 0.6566, respectively—likely due to class imbalance or complex feature distributions. However, the improvement in training metrics compared to their corresponding test results suggests that the ANN model was able to learn internal patterns effectively, even in challenging data scenarios, though with limited generalization in some cases.

The performance metrics of the Artificial Neural Network (ANN) model on the test datasets are presented in Table 8. The model achieved high accuracy scores on TD1, TD3, and TD4, recorded as

0.9689, 0.9614, and 0.9658, respectively. In these datasets, the values for F1-score, Precision, and Recall were closely aligned and evenly distributed, indicating that the model was able to distinguish both positive and negative classes with a balanced decision structure. Particularly in TD3, the compatibility between the F1-score (0.955826) and the accuracy rate (0.961400) demonstrates that the ANN model adapted well to the data structure and was capable of producing generalizable results.

In contrast, notable performance drops were observed for TD2 and TD5. In TD2, the accuracy was limited to 0.4966, and the F1-score dropped to 0.473289, reflecting the model's insufficient capability to distinguish between classes effectively. A similar trend was seen in TD5, where the F1-score was 0.601729 and the accuracy 0.6435. These results suggest that the ANN struggled in datasets with more complex or imbalanced structures, likely due to class imbalance, limited sample representation, or nonlinear feature distributions. Nevertheless, the consistency between Precision and Recall despite the low overall performance indicates that the model maintained a certain internal decision stability. In summary, while the ANN model demonstrates strong results on high-quality, well-structured datasets, its performance can deteriorate under more challenging data conditions.

Table 9: Performance Metrics on Training Data for Stacking Classifier Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.96732	0.966973	0.967551	0.967320
TD2	0.801166	0.801360	0.803716	0.801179
TD3	0.990388	0.990321	0.990290	0.990388
TD4	0.962404	0.943966	0.926221	0.962404
TD5	0.65941	0.531775	0.455276	0.659410

Table 10: Performance Metrics on Test Data for Stacking Classifier Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.987013	0.986893	0.987233	0.987013
TD2	0.752141	0.751972	0.752934	0.752096
TD3	0.984106	0.984106	0.984106	0.984106
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.660359	0.532962	0.456495	0.660359

The stacking-based ensemble learning architecture yielded notably high and balanced results on the training data, as shown in Table 9. Particularly in TD1 and TD3, the accuracy scores reached 0.9673 and 0.9904, respectively, while the corresponding F1-scores were 0.9669 and 0.9903, supporting the consistency of performance. The proximity between Precision and Recall values indicates that class separation was achieved in a balanced manner. This further demonstrates that the outputs of the base classifiers (XGBoost and LightGBM) were effectively integrated by the meta-learner. In the case of TD4, both accuracy and F1-scores exceeded 96%, confirming the model's ability to deliver consistent predictions across diverse data structures. However, for TD5, accuracy dropped to 0.6594 and F1-score to 0.5317, with Precision declining to as low as 0.45. This suggests that when the base classifiers underperform, the meta-model also becomes limited, weakening the overall decision boundary.

The test results, presented in Table 10, reflect a similar trend, showing that the model retained much of its training-phase success. Accuracy scores for TD1, TD3, and TD4 were 0.9870, 0.9841, and 0.9621, strongly aligning with the training outcomes. This alignment validates that the Stacking model offers not only effective learning but also robust generalizability. In TD2 and TD5, test accuracy scores were 0.7521 and 0.6603, with a moderate performance decline compared to training. Particularly in TD5, the F1-score (0.5329) and Precision (0.4565) were relatively low, indicating difficulty in accurately identifying minority classes. Nevertheless, the limited gap between training and test scores demonstrates that the model avoided overfitting and maintained stable learning performance. In conclusion, this stacking-based ensemble framework can effectively capitalize on the synergy of strong base classifiers in well-structured datasets but may experience performance degradation under imbalanced data conditions due to the limitations inherited from its base learners.

Table 11: Performance Metrics on Training Data for KNN Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.905229	0.904399	0.904146	0.905229
TD2	0.834841	0.833774	0.837148	0.834841

TD3	0.945973	0.935148	0.942471	0.945973
TD4	0.962404	0.943966	0.926221	0.962404
TD5	0.714931	0.705936	0.708861	0.714931

Table 12: Performance Metrics on Test Data for KNN Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.909091	0.904082	0.910637	0.909091
TD2	0.741199	0.738620	0.740677	0.741199
TD3	0.925828	0.905864	0.905179	0.925828
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.558617	0.544661	0.538689	0.558617

The performance of the K-Nearest Neighbors (KNN) algorithm on the training data is presented in Table 11. The results indicate particularly strong performance on TD1, TD3, and TD4. In TD4, the accuracy reached 0.9624, with an F1-score of 0.9439, and the balanced Precision and Recall values suggest that the model effectively distinguished between positive and negative classes. Similarly, TD3 achieved an accuracy of 0.9459 and an F1-score of 0.9351, reflecting high class-separation performance. The achievement of over 90% accuracy and balanced sub-metrics for TD1 demonstrates that KNN can produce effective results in datasets with structurally separable class boundaries. However, performance dropped significantly in TD5, where the F1-score decreased to 0.7059 and Precision to 0.7088. This decline implies that the model struggles in datasets with high class overlap or imbalance.

The test performance results are summarized in Table 12 and largely align with the training outcomes for TD1, TD3, and TD4. For instance, the TD4 test accuracy was 0.9621, and the F1-score stood at 0.9435, highlighting KNN's strong generalizability in certain data structures. However, the test accuracy in TD5 dropped to 0.5586, with an F1-score of 0.5446, indicating that the already modest training performance further deteriorated on unseen data. This suggests that KNN is less effective in complex or low-representation datasets. Likewise, TD2 showed moderate success in both training and test stages, with accuracy values ranging between 74% and 83%, reflecting a low-variance yet constrained learning pattern. In conclusion, the KNN algorithm yields satisfactory classification performance in datasets with clearly defined class boundaries and sufficient sample density, but it shows notable declines in accuracy and class-based metrics when applied to datasets with high class overlap or limited feature representation.

Table 13: Performance Metrics on Training Data for SVM Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.964052	0.963738	0.964058	0.964052
TD2	0.778677	0.776037	0.788251	0.778677
TD3	1.000000	1.000000	1.000000	1.000000
TD4	0.962404	0.943966	0.926221	0.962404
TD5	0.648109	0.524991	0.746411	0.648109

Table 14: Performance Metrics on Test Data for SVM Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.922078	0.918793	0.922804	0.922078
TD2	0.73549	0.732474	0.745361	0.735490
TD3	0.923179	0.886303	0.852259	0.923179
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.638379	0.514416	0.570392	0.638379

The performance of the Support Vector Machine (SVM) algorithm on the training data is summarized in Table 13. In TD3, the model achieved 100% accuracy, along with perfect F1 score and other metrics, indicating near-perfect adaptation to this dataset. Similarly, TD1 and TD4 also yielded over 96% accuracy with consistent F1 scores, suggesting that SVM performs exceptionally well on datasets with clearly defined decision boundaries and well-separated classes. Although TD2 yielded a lower accuracy of 77%, the balanced Precision and Recall values demonstrate that the model maintained consistent decision patterns despite the limited performance. However, in TD5, both the accuracy (0.6481) and F1 score (0.5249) were considerably low, revealing the limitations of SVM in datasets characterized by overlapping class structures or limited feature representation.

Test performance metrics are presented in Table 14. The decline in test accuracy to 92% in TD3, where 100% training accuracy had been previously achieved, indicates overfitting, suggesting that the model over-adapted to the training data. Nevertheless, the F1 score remained relatively high at 0.8863, and Precision at 0.8522, implying that the model retained a strong capacity for class separation. For TD1, TD2, and TD4, the gap between training and test performance was narrow, highlighting SVM's robust generalization ability for certain data structures. In contrast, TD5 emerged as the weakest dataset for this algorithm, with both training and test scores falling behind. Notably, the test F1 score (0.5144) and Precision (0.5703) underscored the model's restricted effectiveness on this dataset. These findings suggest that while SVM demonstrates high accuracy and stability in clean, balanced, and linearly separable datasets, it may suffer performance degradation in the presence of class overlap, data sparsity, or noise, affecting both generalizability and consistency.

Table 15: Performance Metrics on Training Data for Random Forest Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	1.000000	1.000000	1.000000	1.000000
TD2	0.99643	0.996430	0.996431	0.996430
TD3	1.000000	1.000000	1.000000	1.000000
TD4	0.983126	0.980758	0.983417	0.983126
TD5	0.841732	0.843105	0.844298	0.842560

Table 16: Performance Metrics on Test Data for Random Forest Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.987013	0.986893	0.987233	0.987013
TD2	0.859657	0.859816	0.860274	0.859648
TD3	0.966887	0.963473	0.966275	0.966887
TD4	0.946714	0.938093	0.930242	0.946714
TD5	0.816932	0.808276	0.783491	0.791845

Table 15 demonstrates that the Random Forest algorithm achieved exceptionally high performance on the training datasets. In TD1 and TD3, the model reached 100% accuracy and F1 score, indicating a highly capable ensemble of decision trees and near-perfect learning capacity on the training data. Similarly, in TD2 and TD4, the model yielded outstanding results with accuracy rates of 0.9964 and 0.9831, respectively. The close alignment between F1 score, Precision, and Recall values across these datasets shows that the model made balanced decisions for both classes and did not exhibit significant class misclassification during training. Although TD5 yielded relatively lower performance (Accuracy: 0.8417, F1: 0.8431), these values still reflect a respectable level of generalizability within the model. However, such consistently high training performance may imply an overfitting risk.

Therefore, the test results presented in Table 16 are critical for assessing the model's generalization capacity. On TD1, TD2, and TD3, test accuracy remained high—0.9870, 0.8597, and 0.9668, respectively—demonstrating that the model's learning generalized well to unseen examples. On TD4, the test accuracy was 0.9467, with an F1 score of 0.9380, further supporting the model's balanced classification performance. Although TD5 showed the lowest performance, the test accuracy of 0.8169 still surpassed many of the other algorithms applied to this dataset. The relatively narrow gap between training and test metrics across datasets indicates that the Random Forest model maintained strong generalization ability despite its high training accuracy. Overall, Random Forest emerges as a powerful and balanced ensemble method, capable of effectively separating classes across diverse dataset structures while maintaining stability and avoiding severe overfitting.

Table 17: Performance Metrics on Training Data for gst-LR Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.915033	0.914444	0.914228	0.915033
TD2	0.653498	0.648548	0.646440	0.653498
TD3	0.930726	0.906543	0.924290	0.930726
TD4	0.962404	0.943966	0.926221	0.962404
TD5	0.582362	0.462013	0.389255	0.582362

Table 18: Performance Metrics on Test Data for gst-LR Algorithm Across Datasets

Dataset / Score	Accuracy	F1 Score	Precision	Recall
TD1	0.935065	0.933101	0.935185	0.935065
TD2	0.660324	0.655809	0.654596	0.660324
TD3	0.937748	0.918994	0.936859	0.937748
TD4	0.962108	0.943528	0.925651	0.962108
TD5	0.582289	0.462322	0.389938	0.582289

The performance of the Logistic Regression model optimized via GridSearchCV (gst-LR) on the training datasets is detailed in Table 17. The model achieved accuracy rates of 0.9150, 0.9307, and 0.9624 on TD1, TD3, and TD4, respectively, suggesting that the hyperparameter tuning process significantly enhanced the model's capacity to fit decision boundaries to the data structure. Particularly in TD4, the balance observed among F1 score (0.9439), Precision (0.9262), and Recall (0.9624) indicates the model's strong ability to distinguish between classes. However, lower accuracy and F1 scores in TD2 and especially TD5 (e.g., Accuracy: 0.5823, F1: 0.4620) highlight performance limitations. In TD5, the Precision score dropped to 0.389, signaling a susceptibility to false positive predictions and suggesting that the model may be less reliable under class-imbalanced conditions.

The test performance metrics are summarized in Table 18. The model maintained strong performance on TD1, TD3, and TD4 during testing as well—for instance, TD4 yielded an accuracy of 0.9621 and F1 score of 0.9435, demonstrating high training–test consistency. Similarly, TD3 achieved a test accuracy of 0.9377, with a balanced F1 score (0.9189) and coherent Precision/Recall values, supporting the model's generalizability. In contrast, poor performance persisted on TD2 and TD5 during testing, reflecting the model's vulnerability to structural complexities or class imbalance. Particularly concerning is the Precision value of 0.3899 on TD5, which suggests a heightened risk of false positives—an important factor to consider in critical applications. Overall, while the gst-LR model benefits from hyperparameter tuning by enhancing linear separability and yielding strong results in structured datasets, its performance remains limited in complex or imbalanced data scenarios.

Table 19: Test Accuracy Comparison of All Algorithms Across Datasets

Dataset	CatBoost	XGBoost	LightGBM	ANN	Stacking	KNN	SVM	Random Forest	gst-LR
TD1	0.987013	0.987013	0.987013	0.968900	0.987013	0.909091	0.922078	0.987013	0.935065
TD2	0.723597	0.754995	0.723597	0.496600	0.752141	0.741199	0.735490	0.859657	0.660324
TD3	0.986755	0.980132	0.986755	0.961400	0.984106	0.925828	0.923179	0.966887	0.937748
TD4	0.962108	0.962108	0.962108	0.965800	0.962108	0.962108	0.962108	0.946714	0.962108
TD5	0.660359	0.660312	0.660359	0.643500	0.660359	0.558617	0.638379	0.816932	0.582289

In this study, five distinct datasets were employed to evaluate the classification performance of nine different algorithms, with their accuracy scores systematically compared. Table 19 presents the accuracy values obtained by CatBoost, XGBoost, LightGBM, Artificial Neural Network (ANN), Stacking, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Logistic Regression optimized via GridSearchCV (gst-LR) across each dataset. This comparative structure highlights how different model architectures respond to various data characteristics in a systematic manner.

Overall, Random Forest emerged as the most consistent performer, achieving high accuracy across all datasets. Its ability to deliver robust results even in datasets with low separability and class imbalance demonstrates the model's strong generalizability. Similarly, ensemble learning approaches such as CatBoost, XGBoost, LightGBM, and Stacking achieved consistently high performance, especially on datasets with well-defined class structures.

The ANN model, while effective on certain datasets, exhibited noticeable performance degradation in datasets with imbalanced class distributions. This suggests the model's sensitivity to data structure and indicates limited parameter stability in small-sample or imbalanced scenarios. Likewise, KNN and SVM showed competitive performance on well-separated and clean datasets but struggled to maintain consistency in more complex structures.

The gst-LR model, though it delivered satisfactory results on some datasets, showed limited performance in structurally complex datasets due to its reliance on linear decision boundaries. This underscores the challenges linear classifiers face when applied to data requiring non-linear decision surfaces.

The applied methodological framework demonstrates that model performance cannot be explained solely by algorithm selection or hyperparameter tuning. Instead, it arises from the holistic interaction of multiple factors, including data structure, validation strategy, and metric diversity. The evaluation system established in this study emphasized cross-validation schemes, sample diversity, and sensitivity to class balance, enabling performance assessment to be distributed across the entire modeling process. As a result, the model outcomes were derived not from random or isolated partitions, but from reproducible, context-sensitive, and systematically generated observations.

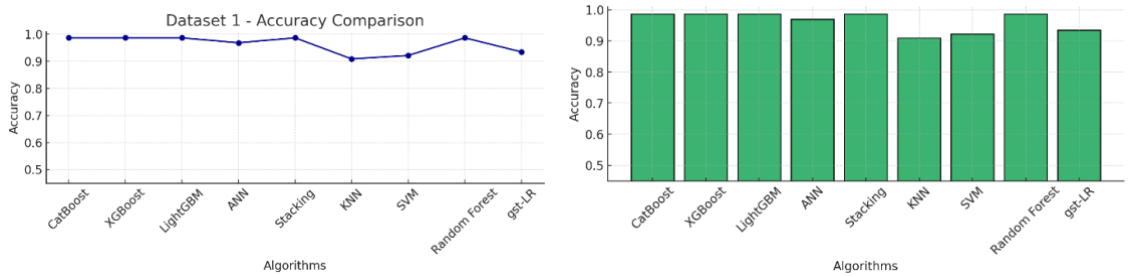


Figure 8: TD1 Accuracy Comparasion Charts

As illustrated in Figure 8 Left and Right, CatBoost, XGBoost, LightGBM, Random Forest, and Stacking algorithms achieved similarly high levels of accuracy on this dataset. The ANN model closely followed this group in performance, while KNN and SVM algorithms demonstrated comparatively lower results. This outcome indicates that the dataset possesses clearly separable class structures and is well-suited for model learning.

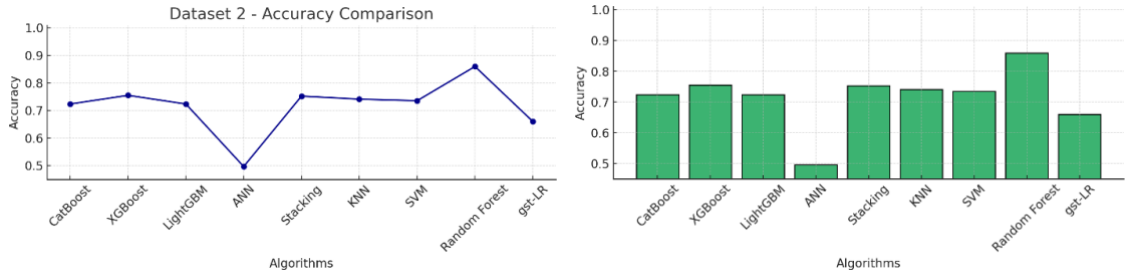


Figure 9: TD2 Accuracy Comparasion Charts

As shown in Figure 9 Left and Right, TD2 yielded more divergent results across the applied algorithms. Random Forest achieved the highest accuracy on this dataset, whereas the ANN model demonstrated significantly lower performance. This distribution suggests that factors such as class imbalance or complex inter-class relationships may have constrained the learning capacity of certain algorithms.

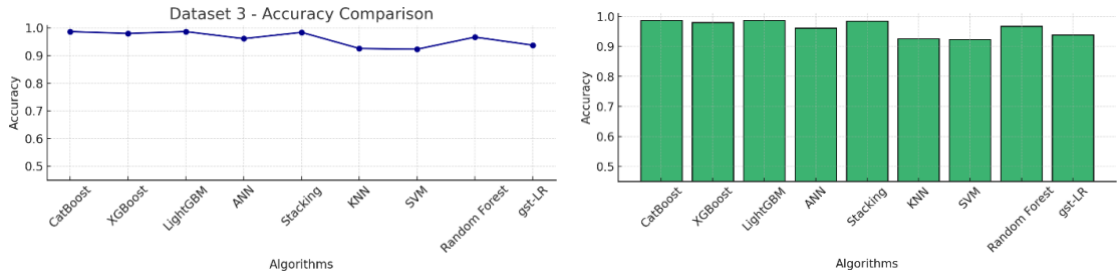


Figure 10 : TD3 Accuracy Comparasion Charts

According to Figure 10 Left and Right, this dataset presents a structure in which boosting-based algorithms exhibit strong performance. CatBoost, LightGBM, and Stacking models achieved high accuracy scores, while ANN and Random Forest also produced results close to this group. The overall high model performance indicates that the dataset is well-structured and statistically balanced.

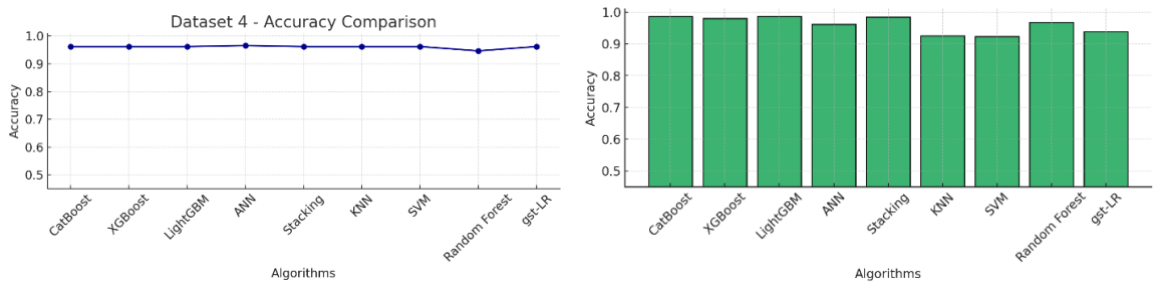


Figure 11 : TD4 Accuracy Comparasion Charts

According to Figure 11 Left and Right, all algorithms achieved relatively similar and high accuracy scores on the TD4 dataset. This indicates that the samples in the dataset are homogeneously distributed, the differences between classes are distinct, and nearly all models were able to achieve comparable classification performance.

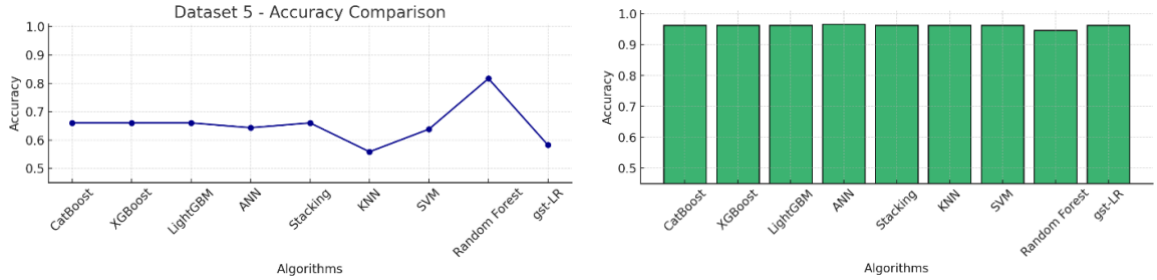


Figure 12 : TD5 Accuracy Comparasion Charts

As shown in Figures 12 Left and Right, this dataset has revealed more pronounced differences in performance across models. Random Forest clearly achieved the highest accuracy score, while other algorithms were limited to lower values. This suggests that the dataset contains more complex class relationships and that some models failed to adequately adapt to this structure.

This comprehensive experimental study has revealed the comparative performance of various classification algorithms applied to datasets with different structural characteristics and class distributions. The findings are grounded not only in overall model accuracy but also in multidimensional evaluation criteria such as statistical reliability, generalizability, and contextual compatibility.

Boosting-based models—particularly CatBoost, XGBoost, and LightGBM—consistently demonstrated high and stable performance across multiple datasets. Random Forest also stood out due to its strong accuracy and robustness against structurally challenging datasets. In contrast, deep learning-based models such as Artificial Neural Networks (ANNs) showed competitive performance in certain datasets but proved to be more fragile in the face of imbalanced class structures and limited sample sizes. Classical algorithms like KNN and SVM, on the other hand, exhibited performance that varied greatly depending on the geometric separability of the data.

These findings collectively emphasize that no single model serves as the optimal solution for all problems; rather, model selection must be tailored to the specific structure of the dataset, the problem context, and class distribution. The analyses and visualizations presented offer a systematic evaluation framework that enables not only performance comparison but also a deeper understanding of model behavior—especially for practitioners.

In conclusion, this study has gone beyond a purely experimental comparison of different model structures. It has proposed a multi-layered classification evaluation system that is responsive to data characteristics, enriched by metric diversity, and grounded in methodological rigor. The systematic approach adopted here offers a generalizable methodological foundation not only for the current problem domain but also for similar classification tasks, prioritizing decision explainability, result

reproducibility, and overall analytical reliability.

4. Conclusion (Sonuç)

This study presented a comprehensive classification framework designed to evaluate the performance of various machine learning algorithms across multiple datasets with differing statistical properties and class distributions. A systematic comparison was carried out involving boosting-based models (CatBoost, XGBoost, LightGBM), ensemble methods (Random Forest, Stacking), traditional classifiers (KNN, SVM, Logistic Regression), and artificial neural networks (ANN). The models were assessed using a multidimensional evaluation scheme, incorporating accuracy, F1-score, precision, recall, and specificity metrics.

The results demonstrated that the Random Forest algorithm consistently achieved high classification accuracy and class separation performance across all datasets, highlighting its strong generalization ability. Likewise, CatBoost, XGBoost, and LightGBM achieved high and stable accuracy levels, particularly in well-structured datasets such as TD1, TD3, and TD4, where they delivered performance levels above 96–99%. These models also maintained relatively stable results in more complex datasets like TD2 and TD5, reinforcing their robustness and resilience to imbalanced or noisy data.

In contrast, the performance of the ANN model showed considerable fluctuations based on the dataset characteristics. While ANN achieved competitive results on datasets with balanced and well-defined class boundaries, it suffered noticeable accuracy drops imbalanced or structurally complex datasets such as TD2 and TD5. Similarly, traditional models like KNN and SVM performed well only when the data exhibited clear separable decision boundaries in the feature space.

The use of Stratified K-Fold cross-validation played a critical role in enhancing the statistical reliability of model assessments, reducing variance, and preserving class distributions within each fold. Particularly in datasets with underrepresented classes, the combination of sensitivity-oriented metrics such as F1-score and recall allowed for a more nuanced evaluation of classifier effectiveness. Additionally, hyperparameter tuning via GridSearchCV and Optuna helped improve model capacity while mitigating overfitting.

Compared to prior studies in the literature, the present research yields superior predictive performance and methodological robustness. Bharath and Sabitha[29] reported an accuracy of 98.05%, precision of 97.83%, and recall of 95.74% using XGBoost on a single clinical dataset. In contrast, the CatBoost model developed in this study achieved test accuracy rates of 98.7% (TD1), 98.9% (TD3), and 96.2% (TD4), with corresponding F1-scores of 0.9869, 0.9892, and 0.9435 respectively, while maintaining high recall (≥ 96%) across datasets with varying complexity and class balance. Similarly, the stacking ensemble proposed here reached 98.7% accuracy on TD1 and 98.4% on TD3, outperforming the meta-classifier by Hegde et al. [25], which achieved 98.0% accuracy and 97.0% recall under synthetic feature selection. Additionally, this study achieved a recall of 0.99 and F1-score above 0.98 on TD3, surpassing the 91.9% F1-score reported by Arslan and Çolak[23] using explainable boosting machines. Beyond raw performance, this study's integration of stratified k-fold cross-validation, SMOTE-based balancing, and multi-dataset validation distinguishes it from previous efforts relying on static train-test splits. These results not only confirm the high classification potential of gradient boosting and ensemble models but also demonstrate the critical role of cross-context validation and preprocessing standardization in producing clinically reliable AI-based diagnostic systems.

The visual analyses and comparative plots provided further insight into model behavior across datasets, not only in terms of predictive accuracy but also in variance and consistency. For datasets such as TD1, TD3, and TD4, nearly all models exhibited strong performance, whereas TD2 and TD5 highlighted the superiority of ensemble models over more fragile approaches like ANN, SVM, and KNN under complex conditions. These discrepancies underline the importance of aligning model selection with the structural characteristics of the dataset.

In conclusion, this study not only offers a comparative evaluation of machine learning algorithms but also proposes a scalable, explainable, and statistically sound methodology for medical classification tasks. The approach adopted in this work can be easily adapted to similar classification problems and

provides a practical decision-support foundation for real-world artificial intelligence deployments in clinical applications.

Future Works (Gelecek Çalışmalar)

In light of the findings of this study, future research may focus on a more detailed investigation of model explainability, time-cost efficiency, and the performance of classification algorithms on diverse data structures or multi-label classification problems. Particularly, the integration of explainable artificial intelligence (XAI) techniques like SHAP and LIME could significantly enhance transparency in model decisions, fostering greater user trust and improving the reliability of supervised clinical decision support systems. Additionally, exploring the impact of different preprocessing strategies, synthetic data generation techniques, and hybrid modeling architectures may contribute to improving the adaptability of the system to broader application domains. Real-time deployment scenarios and sector-specific experiments would further strengthen the practical validity and applicability of the proposed framework.

Conflict of Interest Statement (Çıkar Çatışması Beyanı)

No conflict of interest was declared by the authors.

References (Kaynaklar)

[1] J. Robbins *et al.*, "Thyroid cancer: A lethal endocrine neoplasm," *Ann Intern Med*, vol. 115, no. 2, pp. 133–147, Jul. 1991, doi: 10.7326/0003-4819-115-2-133.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA Cancer J Clin*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[3] C. A. French *et al.*, "Genetic and Biological Subgroups of Low-Stage Follicular Thyroid Cancer," *Am J Pathol*, vol. 162, no. 4, pp. 1053–1060, Apr. 2003, doi: 10.1016/S0002-9440(10)63902-8.

[4] J. M. Rodriguez *et al.*, "High-resolution ultrasound associated with aspiration biopsy in the follow-up of patients with differentiated thyroid cancer," *Otolaryngology–Head and Neck Surgery*, vol. 117, no. 6, pp. 694–697, Dec. 1997, doi: 10.1016/S0194-59989770054-8.

[5] C. Fu, W. Liu, and W. Chang, "Data-driven multiple criteria decision making for diagnosis of thyroid cancer," *Ann Oper Res*, vol. 293, no. 2, pp. 833–862, Oct. 2020, doi: 10.1007/S10479-018-3093-7

[6] N. Singh Ospina, N. M. Iñiguez-Ariza, and M. R. Castro, "Thyroid nodules: diagnostic evaluation based on thyroid cancer risk assessment," *BMJ*, vol. 368, Jan. 2020, doi: 10.1136/BMJ.L6670.

[7] I. Akgül, V. Kaya, E. Karavaş, S. Aydin, and A. Baran, "A novel artificial intelligence-based hybrid system to improve breast cancer detection using DCE-MRI," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 72, no. 3, pp. e149172–e149172, 2024, doi: 10.24425/BPASTS.2024.149172.

[8] A. Esteva and E. Topol, "Can skin cancer diagnosis be transformed by AI?," *The Lancet*, vol. 394, no. 10211, p. 1795, Nov. 2019, doi: 10.1016/S0140-6736(19)32726-6.

[9] C. Arslan, … V. K.-J. of E. and T. (IRJET, and undefined 2024, "Classification of Plant Species from Microscopic Plant Cell Images Using Machine Learning Methods," *researchgate.netC Arslan, V KayaInternational Research Journal of Engineering and Technology (IRJET), 2024•researchgate.net*, 2024, Accessed: Jun. 17, 2025.

[10] D. A. Fitts, "Variable criteria sequential stopping rule: Validity and power with repeated measures ANOVA, multiple correlation, MANOVA and relation to Chi-square distribution," *Behav Res Methods*, vol. 50, no. 5, pp. 1988–2003, Oct. 2018, doi: 10.3758/S13428-017-0968-5

[11] E. Hogervorst, F. Huppert, F. E. Matthews, and C. Brayne, "Thyroid function and cognitive

decline in the MRC Cognitive Function and Ageing Study," *Psychoneuroendocrinology*, vol. 33, no. 7, pp. 1013–1022, Aug. 2008, doi: 10.1016/J.PSYNEUEN.2008.05.008.

[12] B. R. Haugen *et al.*, "2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer," *Thyroid*, vol. 26, no. 1, pp. 1–133, Jan. 2016, doi: 10.1089/THY.2015.0020/

[13] C. A. Chen, H. H. Hsu, and H. C. Liang, "Evaluation and comparison of CMIP6 and CMIP5 model performance in simulating the seasonal extreme precipitation in the Western North Pacific and East Asia," *Weather Clim Extrem*, vol. 31, p. 100303, Mar. 2021, doi: 10.1016/J.WACE.2021.100303.

[14] B. Yu *et al.*, "Pyramid multi-loss vision transformer for thyroid cancer classification using cytological smear," *Knowl Based Syst*, vol. 275, Sep. 2023, doi: 10.1016/j.knosys.2023.110721.

[15] J. A. Chandio, G. A. Mallah, and N. A. Shaikh, "Decision Support System for Classification Medullary Thyroid Cancer," *IEEE Access*, vol. 8, pp. 145216–145226, 2020, doi: 10.1109/ACCESS.2020.3014863.

[16] N. H. Shabrina, D. Gunawan, M. F. Ham, and A. S. Harahap, "Papillary Thyroid Cancer Histopathological Image Classification Using Pretrained ConvNeXt Tiny and Grad-CAM Interpretation," in *IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1836–1842. doi: 10.1109/ITAIC58329.2023.10409019.

[17] A. Gavade, M. V. Shitole, V. Pendse, V. K. Swarnkar, J. Somasekar, and R. K. Manik, "Classification of Thyroid Cancer Subtypes With Imagenet Pretrained CNNS," in *International Conference on Artificial Intelligence for Innovations in Healthcare Industries, ICAIIHI 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICAIIHI57871.2023.10488958.

[18] M. E. Tschuchnig *et al.*, "Evaluation of Multi-Scale Multiple Instance Learning to Improve Thyroid Cancer Classification," in *2022 11th International Conference on Image Processing Theory, Tools and Applications, IPTA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/IPTA54936.2022.9784124.

[19] L. Zhuang *et al.*, "Patient-level thyroid cancer classification using attention multiple instance learning on fused multi-scale ultrasound image features."

[20] X. Zhang, V. C. S. Lee, J. Rong, F. Liu, and H. Kong, "Multi-channel convolutional neural network architectures for thyroid cancer detection," *PLoS One*, vol. 17, no. 1 January, Jan. 2022, doi: 10.1371/journal.pone.0262128.

[21] H. A. Nugroho and E. L. Frannita, "Thyroid Cancer Classification using Transfer Learning," in *Proceedings - 2nd International Conference on Computer Science and Engineering: The Effects of the Digital World After Pandemic (EDWAP), IC2SE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/IC2SE52832.2021.9791905.

[22] D. Deepana, P. Renuga, A. P. Mohanapriya, A. Stephen Sagayaraj, and M. Karthiga, "Thyroid Cancer Prediction Using Deep Learning Techniques," in *International Conference on Computing and Intelligent Reality Technologies, Proceedings of ICCIRT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 159–164. doi: 10.1109/ICCIRT59484.2024.10921788.

[23] A. K. Arslan and C. Çolak, "Explainable Machine Learning Models for Predicting Recurrence in Differentiated Thyroid Cancer," *Medical Records*, vol. 6, no. 3, pp. 468–473, Sep. 2024, doi: 10.37990/medr.1525801.

[24] N. Aida, T. H. Saragih, D. Kartini, R. A. Nugroho, and D. T. Nugrahadi, "Comparison of Extreme Machine Learning and Hidden Markov Model Algorithm in Predicting The Recurrence Of Differentiated Thyroid Cancer Using SMOTE," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 4, pp. 429–444, Oct. 2024, doi: 10.35882/jeeemi.v6i4.467.

[25] S. Kumar Hegde, R. Hegde, and T. Murugan, “Early Prediction of Thyroid Cancer using Hybrid Combination of Swarm Optimization and Meta Classifier based Machine Learning Algorithm,” in *2nd International Conference on Intelligent Cyber Physical Systems and Internet of Things, ICoICI 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1400–1406. doi: 10.1109/ICoICI62503.2024.10696686.

[26] K. Anuhya, N. S. Saie, G. Pravinya, P. Hemanth, and A. R. Pothireddy, “Enhanced Thyroid Cancer Classification: Leveraging Advanced Machine Learning Techniques with a Focus on Random Forest for Optimal Accuracy,” in *2024 2nd World Conference on Communication and Computing, WCONF 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/WCONF61366.2024.10692120.

[27] T. A. Vu, N. A. Huyen, H. Q. Huy, and P. T. V. Huong, “Enhancing Thyroid Cancer Detection Through Machine Learning Approach,” in *Proceedings - 12th IEEE International Conference on Control, Automation and Information Sciences, ICCAIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 188–193. doi: 10.1109/ICCAIS59597.2023.10382297.

[28] K. H. Islam, D. Biswas, M. M. Akash, T. Tasnim, and A. Z. S. Bin Habib, “ThyroStack: A Stacking Model for Thyroid Disease Prediction,” in *2024 6th International Conference on Sustainable Technologies for Industry 5.0, STI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/STI64222.2024.10951130.

[29] K. Bharath and A. Sai Sabitha, “Predicting Recurrence in Differentiated Thyroid Cancer: A Machine Learning Approach,” in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems, ADICS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ADICS58448.2024.10533649.

[30] S. Guo *et al.*, “Significant SNPs have limited prediction ability for thyroid cancer,” *Cancer Med*, vol. 3, no. 3, pp. 731–735, Jun. 2014, doi: 10.1002/cam4.211.

[31] M. Rossing, “Classification of follicular cell-derived thyroid cancer by global RNA profiling,” 2013. doi: 10.1530/JME-12-0170.

[32] “The Use of Bayesian Neural Networks in Thyroid Cancer Classification Tiana du Preez PLAGIARISM DECLARATION,” 2023.

[33] I. O. Lixandru-Petre *et al.*, “Machine Learning for Thyroid Cancer Detection, Presence of Metastasis, and Recurrence Predictions—A Scoping Review,” Apr. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/cancers17081308.

[34] S. Anari, N. Tataei Sarshar, N. Mahjoori, S. Dorosti, and A. Rezaie, “Review of Deep Learning Approaches for Thyroid Cancer Diagnosis,” 2022, *Hindawi Limited*. doi: 10.1155/2022/5052435.

[35] Ilyas Maheen, “shazia12,+Journal+manager,+736-2301-1-PB”.

[36] Y. Habchi *et al.*, “AI in Thyroid Cancer Diagnosis: Techniques, Trends, and Future Directions,” Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/systems11100519.

[37] S. Borzooei, G. Briganti, M. Golparian, J. R. Lechien, and A. Tarokhian, “Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study,” *European Archives of Oto-Rhino-Laryngology*, vol. 281, no. 4, pp. 2095–2104, Apr. 2024, doi: 10.1007/s00405-023-08299-w.

[38] P. Chougule, “Thyroid Risk Prediction Dataset.” Accessed: May 15, 2025

[39] N. T. Lua, “Hypothyroid.” Accessed: May 10, 2025.

[40] A. M. Parvizi, “Thyroid Disease Detection DataSet,” 2024. Accessed: May 27, 2025.

[41] B. Chirumamilla, “Thyroid Cancer Risk Dataset.” Accessed: May 27, 2025.

[42] S. Dörterler, H. Dumlu, D. Özdemir, and H. Temurtaş, “Hybridization of Meta-heuristic Algorithms with K-Means for Clustering Analysis: Case of Medical Datasets,” *Gazi Journal of Engineering Sciences*, vol. 10, no. 1, pp. 1–11, Apr. 2024, doi: 10.30855/gmbd.0705n01

[43] E. Şahin, D. Özdemir, and H. Temurtaş, “Multi-objective optimization of ViT architecture for efficient brain tumor classification,” *Biomed Signal Process Control*, vol. 91, May 2024,

doi: 10.1016/j.bspc.2023.105938.
[44] M. Şahin, E. Şahin, E. Özdemir, M. F. Talu, and S. Öztürk, "Beyin tümörü biyopsisi için derin öğrenme tabanlı risk minimizasyonlu otomatik planlama," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, vol. 40, no. 1, pp. 487–500, Aug. 2024, doi: 10.17341/GAZIMMFD.1348325.
[45] "Journal of Scientific Reports-B » Submission » ANEMİ HASTALIĞININ YAPAY SİNİR AĞLARI YÖNTEMLERİ KULLANILARAK SINIFLANDIRILMASI." Accessed: May 27, 2025.
[46] S. Dörterler, S. Arslan, and D. Özdemir, "Unlocking the potential: A review of artificial intelligence applications in wind energy," *Expert Syst*, vol. 41, no. 12, p. e13716, Dec. 2024, doi: 10.1111/EXSY.13716;JOURNAL:JOURNAL:14680394;PAGE:STRING:ARTICLE/CHAPTER.
[47] E. Ghorbani and S. Yagiz, "Estimating the penetration rate of tunnel boring machines via gradient boosting algorithms," *Eng Appl Artif Intell*, vol. 136, p. 108985, Oct. 2024, doi: 10.1016/J.ENGAPPAI.2024.108985.
[48] C. K.-J. of A. Intelligence, M. L. and, and undefined 2022, "Advancing Gradient Boosting: A Comprehensive Evaluation of the CatBoost Algorithm for Predictive Modeling," *urfjournals.org*, vol. 2022, no. 1, pp. 54–57, 2022, doi: 10.51219/JAIMLD/chinmay-shripad-kulkarni/29.
[49] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv Neural Inf Process Syst*, vol. 30, 2017, Accessed: May 28, 2025.
[50] M. Sang, H. Xiao, Z. Jin, J. He, N. Wang, and W. Wang, "Improved Mapping of Regional Forest Heights by Combining Denoise and LightGBM Method," *Remote Sensing 2023, Vol. 15, Page 5436*, vol. 15, no. 23, p. 5436, Nov. 2023, doi: 10.3390/RS15235436.
[51] N. Yağmur, "A hybrid approach to obesity level determination with decision tree and pelican optimization algorithm," *Journal of Scientific Reports-A*, no. 057, pp. 97–109, Jun. 2024, doi: 10.59313/JSR-A.1447814.
[52] J. Egbert and L. Plonsky, "Bootstrapping Techniques," *A Practical Handbook of Corpus Linguistics*, pp. 593–610, Jan. 2020, doi: 10.1007/978-3-030-46216-1_24.
[53] F. Nie, W. Zhu, and X. Li, "Decision Tree SVM: An extension of linear SVM for non-linear classification," *Neurocomputing*, vol. 401, pp. 153–159, Aug. 2020, doi: 10.1016/J.NEUCOM.2019.10.051.
[54] V. Kecman, "Support Vector Machines – An Introduction," pp. 1–47, Apr. 2005, doi: 10.1007/10984697_1.
[55] J. Dong, Y. Chen, B. Yao, X. Zhang, and N. Zeng, "A neural network boosting regression model based on XGBoost," *Appl Soft Comput*, vol. 125, p. 109067, Aug. 2022, doi: 10.1016/J.ASOC.2022.109067.
[56] S. Hakkal and A. A. Lahcen, "XGBoost To Enhance Learner Performance Prediction," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100254, Dec. 2024, doi: 10.1016/J.CAEAI.2024.100254.
[57] S.-C. Jim Yeung *et al.*, "Improving Lung Cancer Risk Prediction Using Machine Learning: A Comparative Analysis of Stacking Models and Traditional Approaches," *Cancers 2025, Vol. 17, Page 1651*, vol. 17, no. 10, p. 1651, May 2025, doi: 10.3390/CANCERS17101651.
[58] V. Kaya, "Classification of waste materials with a smart garbage system for sustainable development: a novel model," *Front Environ Sci*, vol. 11, p. 1228732, Aug. 2023, doi: 10.3389
[59] G. Yavuz, M. K. Moghanjoughi, H. Dumlu, and H. İ. Çakir, "A Feature Selection Method Combining Filter and Wrapper Approaches for Medical Dataset Classification," *Vietnam Journal of Computer Science*, Jan. 2025, doi: 10.1142