Cancer Management and Research

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

Improving Non-Invasive Prediction of Thyroid Nodule Malignancy: A Machine Learning-Based Clinical Approach

Maja Reiner, Hanna Drobińska, Michał Miciak, Michał Kisiel, Szymon Biernat, Krzysztof Kaliszewski

Department of General Surgery, University Centre of General and Oncological Surgery, Faculty of Medicine, Wroclaw Medical University, Wroclaw, Poland

Correspondence: Maja Reiner, Department of General Surgery, University Centre of General and Oncological Surgery, Faculty of Medicine, Wroclaw Medical University, Borowska Street 213, Wroclaw, 50-556, Poland, Email maja.reiner@student.umw.edu.pl

Background: Thyroid cancer (TC) is the most commonly diagnosed endocrine malignancy, with rising global incidence. Current diagnostic techniques, including ultrasound and fine-needle aspiration biopsy (FNAB), often yield inconclusive results, leading to unnecessary thyroidectomies for benign nodules. Improving preoperative risk stratification using non-invasive methods remains an important clinical challenge. This study aimed to develop machine learning (ML) models to enhance the classification of thyroid nodules (TNs) as malignant or benign based solely on selected ultrasonographic features.

Patients and methods: Data from 5928 patients who underwent thyroidectomy at Wroclaw Medical University (2008–2023) were retrospectively analyzed. Five ultrasonographic features were included: hypoechogenicity, microcalcifications, shape, irregular margins, and vascularity. Five ML models – Random Forest, Logistic Regression, Multilayer Perceptron (MLP), Gradient Boosting Machines, and Decision Tree – were trained and evaluated. Model performance was assessed using accuracy, precision, recall, F1 score, specificity, and the area under the receiver operating characteristic curve (ROC-AUC). Feature importance was analyzed to determine the contribution of each variable.

Results: Among the evaluated models, Random Forest achieved the highest overall performance, with an accuracy of 0.905, specificity of 0.939, ROC-AUC of 0.843, and recall of 0.616. Nodule vascularity was identified as the most influential predictor, followed by microcalcifications, irregular margins, and hypoechogenicity.

Conclusion: ML models based on a limited set of ultrasonographic features can effectively support the non-invasive identification of benign TNs, potentially reducing unnecessary surgical interventions. However, the modest recall underscores that the current approach is insufficient for reliable standalone malignancy detection. Incorporation of additional imaging parameters and cytological data would be necessary to enhance sensitivity and improve clinical applicability.

Keywords: thyroid cancer, thyroidectomy, artificial intelligence, machine learning, thyroid nodule diagnosis, ultrasonography

Introduction

Thyroid cancer (TC) is the most frequently diagnosed endocrine carcinoma, ranking seventh in global incidence according to 2022 GLOBOCAN statistics.[1] This high prevalence underscores the importance of accurate diagnostic methods that can effectively identify malignancies and minimize unnecessary interventions. Currently, ultrasound (US) combined with fine-needle aspiration biopsy (FNAB) constitutes the gold standard for the diagnostic evaluation of TC.[2] Ultrasonography serves as the first step in TC diagnosis, providing detailed images of nodules and enabling the assessment of critical characteristics such as shape, margins, microcalcifications, echogenicity, and vascularity. However, interpretation becomes difficult when features typically associated with malignancy appear in benign nodules, or when malignant nodules present with benign characteristics.[3] Based on ultrasound characteristics, clinicians determine whether a nodule requires further investigation with FNAB, which enables cytological assessment of the cells within the nodule. The results of FNAB are classified according to the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC), with categories ranging from benign (Bethesda II) to malignant (Bethesda VI).[4] A significant diagnostic challenge emerges with Bethesda categories III (atypia of undetermined significance), IV (follicular neoplasm), and V (suspicious for malignancy), which represent indeterminate or inconclusive

https://doi.org/10.2147/CMAR.S571308
Received: 7 October 2025
Accepted: 1 March 2026
Published: 28 March 2026

cytological findings.[4,5] According to the third edition of TBSRTC, these categories account for approximately 1/3 of all FNAB results and are associated with a malignancy risk of 22% for category III, 30% for category IV, and 74% for category V, posing a substantial clinical dilemma, especially category III, which is the most heterogeneous.[5] Given this diagnostic uncertainty, many patients with indeterminate cytology undergo thyroidectomy to establish a definitive diagnosis.[6]

A recent study by Mavromati et al found that many thyroidectomies performed for nodules with indeterminate cytology ultimately demonstrated benign pathology on final histological examination. Surgery was deemed unnecessary in 56%, 68%, and 21% of patients with Bethesda III, IV, and V nodules, respectively.[7] These unnecessary surgical interventions carry significant consequences for patients, including permanent hormone replacement and potential complications such as hypoparathyroidism or recurrent laryngeal nerve palsy, and are associated with high costs.[8–10] Cancer diagnosis and treatment may also have a negative effect on patients' psychological health and decrease their quality of life.[11] This high rate of unnecessary surgeries highlights a critical clinical need to improve pre-surgical diagnostic precision for thyroid nodules. Improving diagnostic accuracy could reduce the number of surgical interventions and their associated risks and costs.

Recent advances in artificial intelligence (AI), particularly machine learning (ML), have drawn increasing attention to the potential applications of AI and ML in medical diagnostics.[12,13] ML models require training on datasets to correctly identify complex patterns and make predictions, such as classifying thyroid nodules as malignant or benign.[14–17] Research has shown that ML models may outperform human judgment.[15] In this study, we aimed to evaluate the clinical performance and limitations of commonly used, interpretable ML models for pre-surgical malignancy risk assessment of TNs, based exclusively on routinely available, non-invasive ultrasound features in a large real-world cohort. Importantly, the aim of this study was not to develop novel ML algorithms.

Materials and Methods

All procedures were conducted in accordance with the ethical standards of the institutional and/or national research committee and the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Our retrospective study was approved by the Bioethics Committee of the Wroclaw Medical University, Poland. Signature Number: KB-241/2023; 24 February 2023. Every patient participating in the study provided informed consent at admission, stating that the results may be used for research purposes. Data from established medical records were analyzed retrospectively and anonymously. Due to data anonymization, the authors did not have direct access to the study participants.

Data Collection

The present study analyzed the medical records of 5986 patients admitted to the Department of General Surgery at Wroclaw Medical University during the study period (between January 2008 and December 2023). All of these patients underwent thyroid ultrasound examinations and subsequently had total or partial thyroidectomy due to the presence of either single or multiple thyroid nodules (TNs). The ultrasonography equipment used at our center was consistently of the highest clinical standard, with regular updates to maintain state-of-the-art capabilities. It was performed by two experienced ultrasonographers, each with over 20 years of experience in thyroid imaging. Postoperative samples were collected in each case, and histopathology reports were obtained. Only patients with benign TNs or differentiated thyroid cancer (TC) were considered. Differentiated TC included papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC).[18] Due to changes in histopathological classification over the study period, Hürthle cell TC could not be consistently distinguished from FTC in this retrospective cohort and was therefore analyzed collectively within the malignant group. Fifty-five patients with either poorly differentiated TC, anaplastic TC, medullary TC, secondary TC, sarcoma, lymphoma, extramedullary plasmacytoma, or squamous cell carcinoma were excluded from this study. In total, 5928 records were used in this research; thus, 3 patients were excluded from the study due to missing data. The selection process is illustrated in a flow diagram in Figure 1.

Among these individuals, 4954 were women (83.6%), and 974 were men (16.4%). The mean age was 51.7 ± 14.5 years. TN characteristics, such as (1) hypoechogenicity, (2) microcalcifications, (3) shape, (4) margins and (5) vascularity, were gathered from ultrasonography imaging. The selected ultrasonographic features correspond to key risk descriptors incorporated in structured reporting systems such as EU-TIRADS, in which high-risk nodules (with an

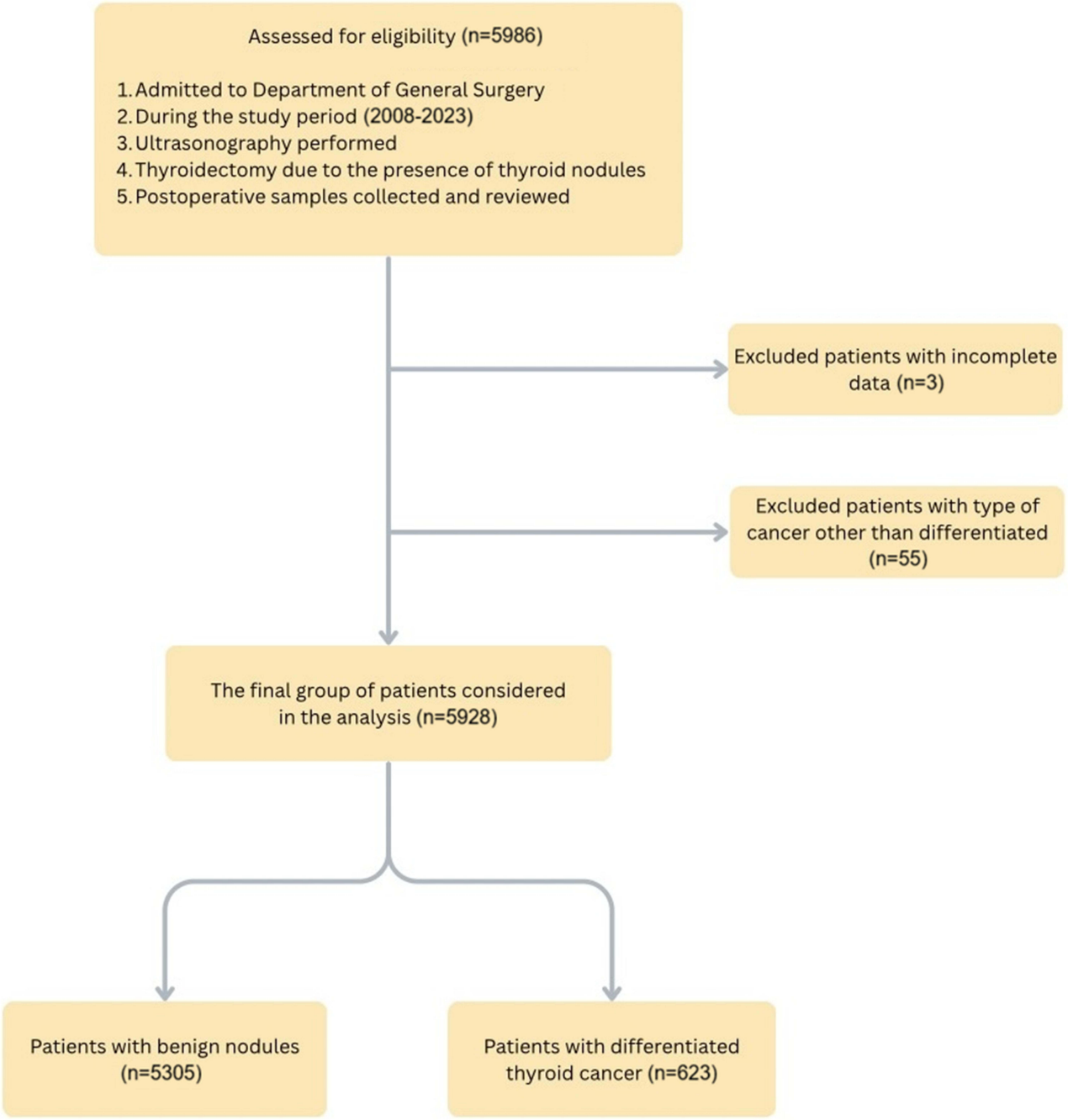

Figure 1 Selection of the study group from the records of 5986 patients admitted to the Department of General Surgery during the study period. All selected patients underwent thyroid ultrasonography prior to thyroidectomy. In each case histopathology results were obtained. Only cases of differentiated thyroid cancer or benign thyroid nodules were analyzed.

estimated malignancy risk of approximately 26–87%) are characterized by hypoechogenicity, microcalcifications, irregular shape, and irregular or ill-defined margins. Vascularity pattern represents an additional relevant factor, as malignant TNs are more likely to exhibit marked intranodular vascular flow.[19] Only categorical ultrasound variables were included, as they are routinely assessed in daily clinical practice, less dependent on equipment-specific settings, and more robust to inter-operator variability, thereby facilitating model interpretability and potential clinical implementation. Table 1 provides an overall summary of the study population.

Table 1 A General Summary of the Overall Study Population, Including Sex Distribution, Mean Age, and Key Ultrasound Features (Hypoechogenicity, Microcalcifications, Shape, Margins and Vascularity)

Variable	Total (n = 5928)
Sex	
Female	4954
Male	974
Mean age	51.7 ± 14.5 years
Ultrasound features	
Hypoechogenicity (+)	2725
Microcalcifications (+)	675
Irregular shape (+)	419
Irregular margins (+)	447
High vascularity (+)	1761

Notes: A (+) sign means the presence of the investigated feature.

Table 2 presents considered ultrasonographic features along with their occurrence typically associated with benign or malignant TNs.[19–21] However, it is important to note that these characteristics often overlap. Features typically associated with malignancy may appear in benign TNs, or conversely, malignant nodules may present with benign characteristics. In Figure 2, there are example ultrasound scans of the thyroid with nodules, illustrating the features discussed above.

All 5 features, presented in Table 3, were further used to build ML models predicting the malignancy of TNs.

Data Preprocessing

Among the 5928 cases eligible for this study, 5305 were classified as benign (89.5%) and 623 as malignant (10.5%) based on histopathology reports. The dataset was then randomly divided into two subsets: a training cohort (80%) and a test cohort (20%). To deal with data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was used in the training dataset. In the test cohort, however, we intentionally preserved the original class distribution (10.5%

Table 2 Features Commonly Associated with Thyroid Nodule Malignancy (Hypoechogenicity, Microcalcifications, Shape, Margins and Vascularity) and Their Encoding for Machine Learning Model Development

Features	Benign Nodule	Malignant Nodule
Hypoechogenicity	No (0)	Yes (1)
Microcalcifications	No (0)	Yes (1)
Shape	Regular (0)	Irregular (1)
Margins	Sharp (0)	Blurred (1)
Vascularity	Low (0)	High (1)

Notes: A value of (0) indicates the absence of the feature in binary coding, whereas a value of (1) indicates its presence.

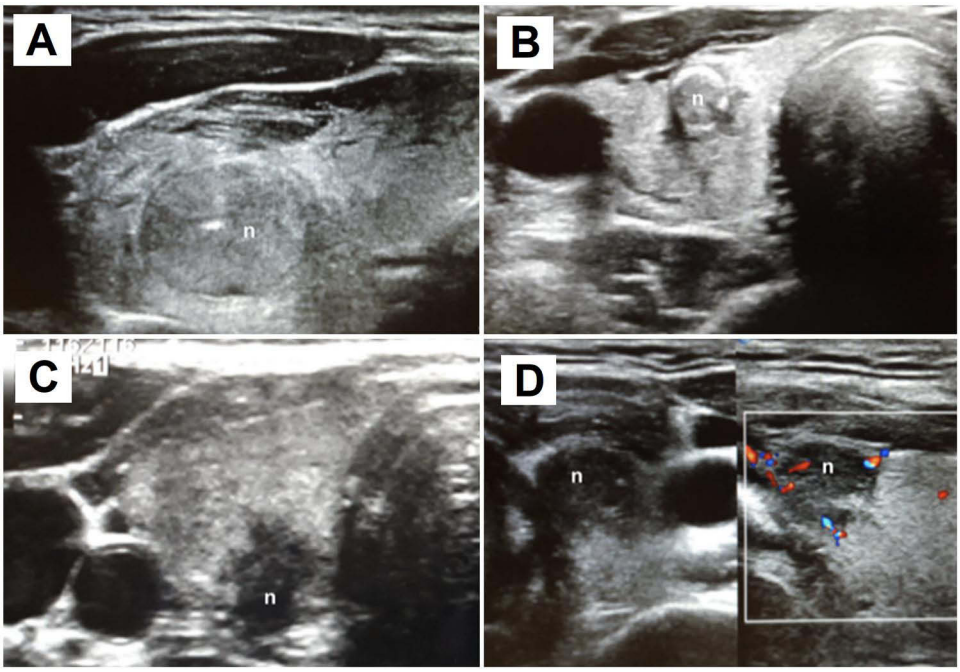

Figure 2 Ultrasound images of a benign and malignant thyroid nodules. (**A**) A benign isoechoic nodule of a regular shape with sharpen margins. (**B**) A malignant nodule with massive peripheral calcifications. (**C**) A malignant hypoechogenic nodule of irregular shape with blurred margins. (**D**) A malignant hypoechogenic nodule of irregular shape with irregular vascularity.
Abbreviation: n, nodule.

malignant cases) to realistically simulate clinical conditions and evaluate the model's true diagnostic performance, as this setup reflects the actual prevalence encountered in routine practice and does not influence model learning, which was conducted on a balanced training set.

Table 3 Distribution of Clinical and Ultrasound Features in Benign and Malignant Thyroid Nodules Among 5928 Patients. Data Include Key Ultrasound Characteristics Analyzed in This Study (Hypoechogenicity, Microcalcifications, Shape, Margins and Vascularity)

	Benign Nodule n = 5305		**Malignant Nodule n = 623**	
Variables	**n**	**[%]**	**n**	**[%]**
(1) Hypoechogenicity				
Yes	2223	41.9%	502	80.6%
No	3082	58.1%	121	19.4%
(2) Microcalcifications				
Yes	348	6.6%	327	52.5%
No	4957	93.4%	296	47.5%

(*Continued*)

https://doi.org/10.2147/CMAR.S571308

Table 3 (Continued).

	Benign Nodule n = 5305		Malignant Nodule n = 623	
Variables	**n**	**[%]**	**n**	**[%]**
(3) Tumor shape				
Regular	5218	98.4%	291	46.7%
Irregular	87	1.6%	332	53.5%
(4) Sharpened margins				
Yes	5193	97.9%	288	46.2%
No	112	2.1%	335	53.8%
(5) Nodule vascularity				
High	1432	27.0%	329	52.8%
Low	3873	73.0%	294	47.2%

Machine Learning Methods

Five machine learning models were taught to classify thyroid tumors as benign (0) or malignant (1) based on data from the training dataset. The program Orange3 provided algorithms for the machine learning models and was used to conduct the training and evaluation process. The models' training and evaluation process is presented in Figure 3.

The models selected for comparison were Random Forest, Logistic Regression, Neural Network Multilayer Perceptron (MLP), Gradient Boosting Machines (GBMs), and Decision Tree. Machine learning models were implemented in Orange3 with fixed hyperparameter settings. Random Forest was trained with 300 trees, maximum tree depth limited to 5, and a minimum node size of 10 instances required for further splitting. Gradient Boosting (scikit-learn implementation) used 300 estimators with a learning rate of 0.05, maximum tree depth of 3, a minimum split size of 10 instances, and subsampling of 0.70 of training instances; replicable training was enabled. Logistic Regression used L2 (ridge) regularization with C = 0.5. A single Decision Tree was induced as a binary tree with maximum depth 5, minimum leaf size of 5 instances, minimum split size of 10 instances, and early stopping when majority reached 95%. The Neural Network classifier used two hidden layers (30 and 10 neurons), ReLU activation, the Adam optimizer, L2 regularization $\alpha = 0.001$, and a maximum of 500 training iterations; replicable training was enabled. Ten-fold cross-validation was applied to obtain reliable performance metrics and enable unbiased comparison between classifiers.[22]

The models' performance was evaluated using five measurements: classification accuracy (CA), precision (Prec), recall (Rec), F1 score, and receiver operating characteristic area under the curve (ROC-AUC). Accuracy measures overall correctness, precision reflects the proportion of correctly identified malignant cases among all cases classified as malignant, recall (sensitivity) indicates the proportion of actual malignant cases correctly detected, specificity measures correct identification of benign cases, F1 score balances precision and recall, and ROC-AUC quantifies discriminatory ability between classes.[23,24] For each model and metric, performance was estimated using ten-fold cross-validation and reported as mean ± standard deviation across folds. Statistical comparisons between models were conducted using paired Wilcoxon signed-rank tests applied to fold-wise results, with Holm-Bonferroni correction for multiple comparisons ($\alpha = 0.05$). Following cross-validation and statistical comparison of model performance, one model was selected for final evaluation on the held-out test dataset.

Results

Random Forest, Gradient Boosting Machines (GBMs), and the Neural Network (MLP) achieved comparable classification accuracy (≈0.79) and ROC-AUC values (≈0.88), indicating similar overall discriminatory performance. The MLP

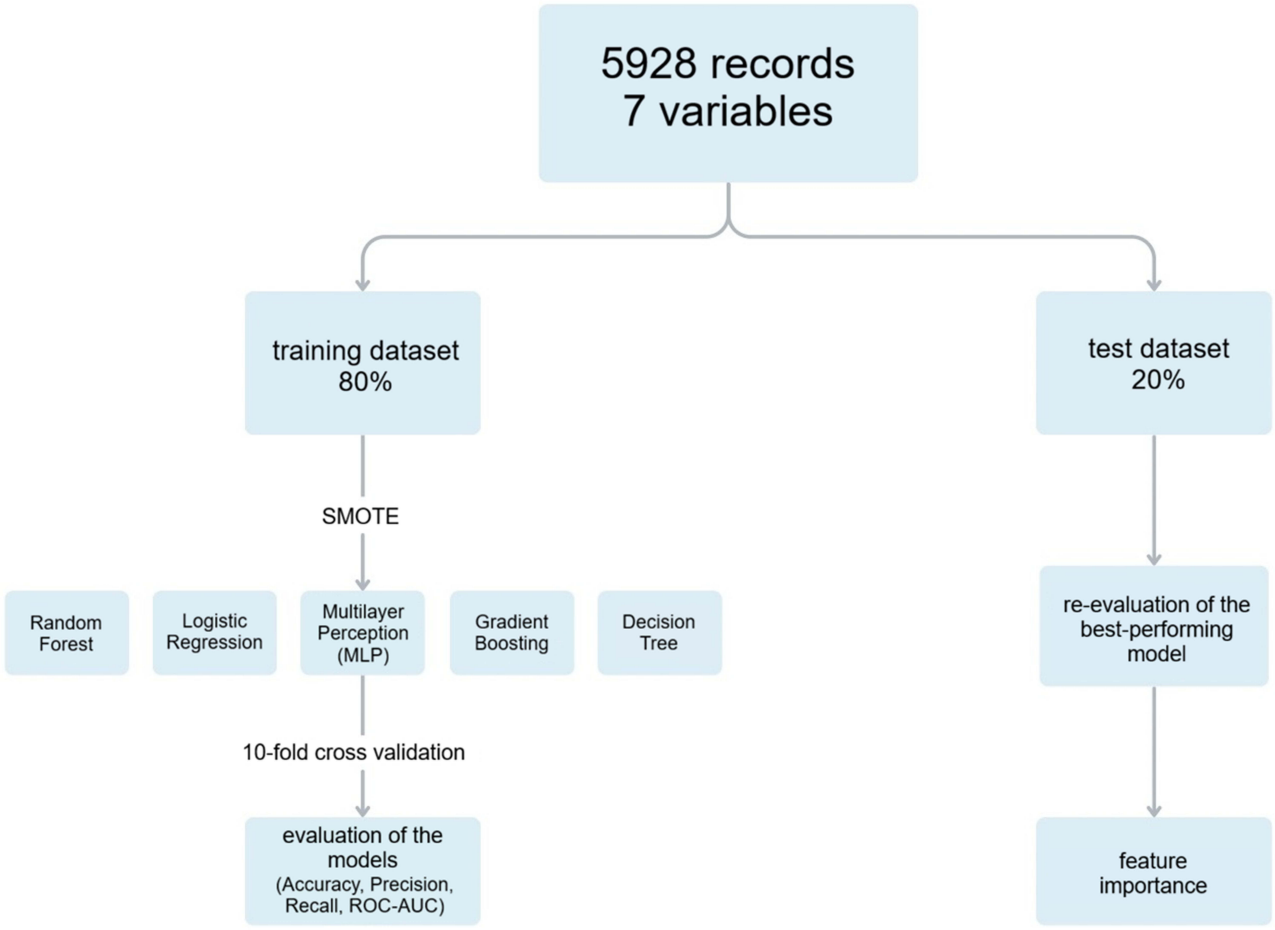

Figure 3 Workflow of the study. The dataset (n = 5928) was split into training (80%) and test (20%) subsets. Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data, and five machine learning models were developed and evaluated. The best-performing model was further evaluated on the test dataset, and feature importance analysis was conducted.
Abbreviation: ROC-AUC, receiver operating characteristic-area under the curve.

obtained the highest precision, whereas Logistic Regression achieved the highest recall, reflecting differences in sensitivity-specificity trade-offs across models. Statistical analysis across cross-validation folds showed no single classifier to be consistently superior across all performance metrics. Given its stable performance, high specificity, and interpretability, Random Forest was selected for further evaluation on the independent test dataset. The detailed results are presented in Table 4.

Table 4 Results Achieved by Machine Learning Models on the Training Dataset, Assessed Using Standard Performance Metrics

Accuracy, CA			
ML model	**Mean**	**± SD**	**p-value vs best (GBMs)**
Random Forest	0.7936	0.0145	1.00
Logistic Regression	0.7871	0.0158	0.02
Neural Network Multilayer Perceptron (MLP)	0.7931	0.0141	1.00
Gradient Boosting Machines (GBMs)	0.7938	0.0144	–
Decision Tree	0.7898	0.0156	0.11

(*Continued*)

https://doi.org/10.2147/CMAR.S571308

Table 4 (Continued).

Precision, Prec			
ML model	**Mean**	**± SD**	**p-value vs best (MLP)**
Random Forest	0.9173	0.0222	0.75
Logistic Regression	0.8888	0.0208	0.01
Neural Network Multilayer Perceptron (MLP)	0.9202	0.0244	–
Gradient Boosting Machines (GBMs)	0.9168	0.0245	0.50
Decision Tree	0.9027	0.0220	0.02
Specificity			
ML model	**Mean**	**± SD**	**p-value vs best (MLP)**
Random Forest	0.9416	0.0164	0.75
Logistic Regression	0.9178	0.0159	0.01
Neural Network Multilayer Perceptron (MLP)	0.9439	0.0181	–
Gradient Boosting Machines (GBMs)	0.9409	0.0184	0.50
Decision Tree	0.9298	0.0163	0.02
Recall, Rec			
ML model	**Mean**	**± SD**	**p-value vs best (Logistic Regression)**
Random Forest	0.6456	0.0215	0.02
Logistic Regression	0.6564	0.0237	–
Neural Network Multilayer Perceptron (MLP)	0.6423	0.0216	0.02
Gradient Boosting Machines (GBMs)	0.6468	0.0244	0.08
Decision Tree	0.6499	0.0228	0.02
F1 score			
ML model	**Mean**	**± SD**	**p-value vs best (GBMs)**
Random Forest	0.7576	0.0182	1.00
Logistic Regression	0.7550	0.0199	1.00
Neural Network Multilayer Perceptron (MLP)	0.7563	0.0177	1.00
Gradient Boosting Machines (GBMs)	0.7581	0.0185	–
Decision Tree	0.7555	0.0196	1.00
ROC-AUC			
ML model	**Mean**	**± SD**	**p-value vs best (GBMs)**
Random Forest	0.8787	0.0141	0.02
Logistic Regression	0.8669	0.0152	0.01
Neural Network Multilayer Perceptron (MLP)	0.8790	0.0143	0.23
Gradient Boosting Machines (GBMs)	0.8791	0.0141	–
Decision Tree	0.8681	0.0144	0.01

Abbreviations: ML, machine learning; SD, standard deviation; ROC-AUC, area under the receiver operating characteristic curve.

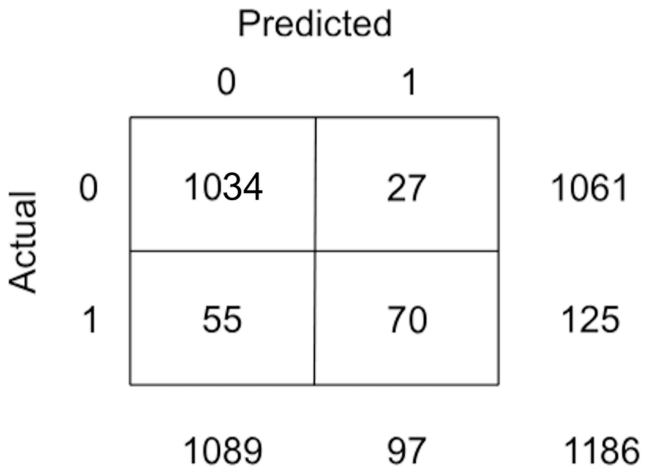

Figure 4 Confusion matrix of the Random Forest model evaluated on the test dataset.
Notes: 0 – benign, 1 – malignant.

The test cohort (20% of the original dataset) consisted of 1061 benign (0) and 125 malignant (1) cases. The confusion matrix revealed that the model correctly detected 77 out of 125 malignant TNs and 996 out of 1061 benign TNs. Figure 4 illustrates the confusion matrix of Random Forest.

Random Forest's performance was then reassessed on the test dataset using the same metrics. As presented in Table 5, the model achieved a classification accuracy of 0.905 (95% CI: 0.887–0.920), precision of 0.577 (95% CI: 0.460–0.622), specificity of 0.939 (95% CI: 0.923–0.952), recall of 0.616 (95% CI: 0.528–0.697), F1 score of 0.577 (95% CI: 0.498–0.644), and ROC-AUC of 0.843 (95% CI: 0.734–0.819). The Random Forest model achieved a Brier score of 0.114, indicating good overall calibration and reliable probability estimates for malignancy prediction. These results confirmed the model's strong performance in correctly identifying benign cases, although its ability to detect malignant nodules was comparatively lower. The observed decrease in precision compared to the training set may be explained by several factors related to dataset composition and model training strategy. First, the test cohort preserved the natural class distribution observed in clinical practice, with malignant cases constituting only 10.5% of the population, whereas the training dataset was balanced using SMOTE. In low-prevalence settings, precision is particularly sensitive to false positive predictions, such that even a modest increase in false positives can lead to a substantial reduction in precision values.[25] Second, SMOTE generates synthetic malignant samples through interpolation between existing minority-class observations. While this approach improves class balance during training, it does not fully reproduce the diversity of real malignant nodules observed in clinical practice.[19] As a result, the model, trained on a dataset containing a large proportion of synthetic cases, was exposed during testing to a broader spectrum of malignant and borderline presentations that had not been fully represented in the training phase. Consequently, this mismatch between the synthetic training distribution and the more heterogeneous, imbalanced, real-world test data led to an increased number of benign nodules being misclassified as malignant, thereby reducing precision.

Feature importance analysis was conducted to evaluate the contribution of each variable to the predictions made by the Random Forest model using the permutation method. The greater the decrease in ROC-AUC after excluding each variable, the larger the impact a feature has on distinguishing between benign and malignant TNs. As shown in Figure 5,

Table 5 Results Achieved by the Random Forest Model on the Test Dataset, Assessed Using Standard Performance Metrics

CA	Prec	Specificity	Rec	F1	ROC-AUC	Brier Score
0.905	0.577	0.939	0.616	0.577	0.843	0.114

Abbreviations: CA, classification accuracy; Prec, precision; Rec, recall; F1, F1 score; ROC-AUC, area under the receiver operating characteristic curve.

https://doi.org/10.2147/CMAR.S571308

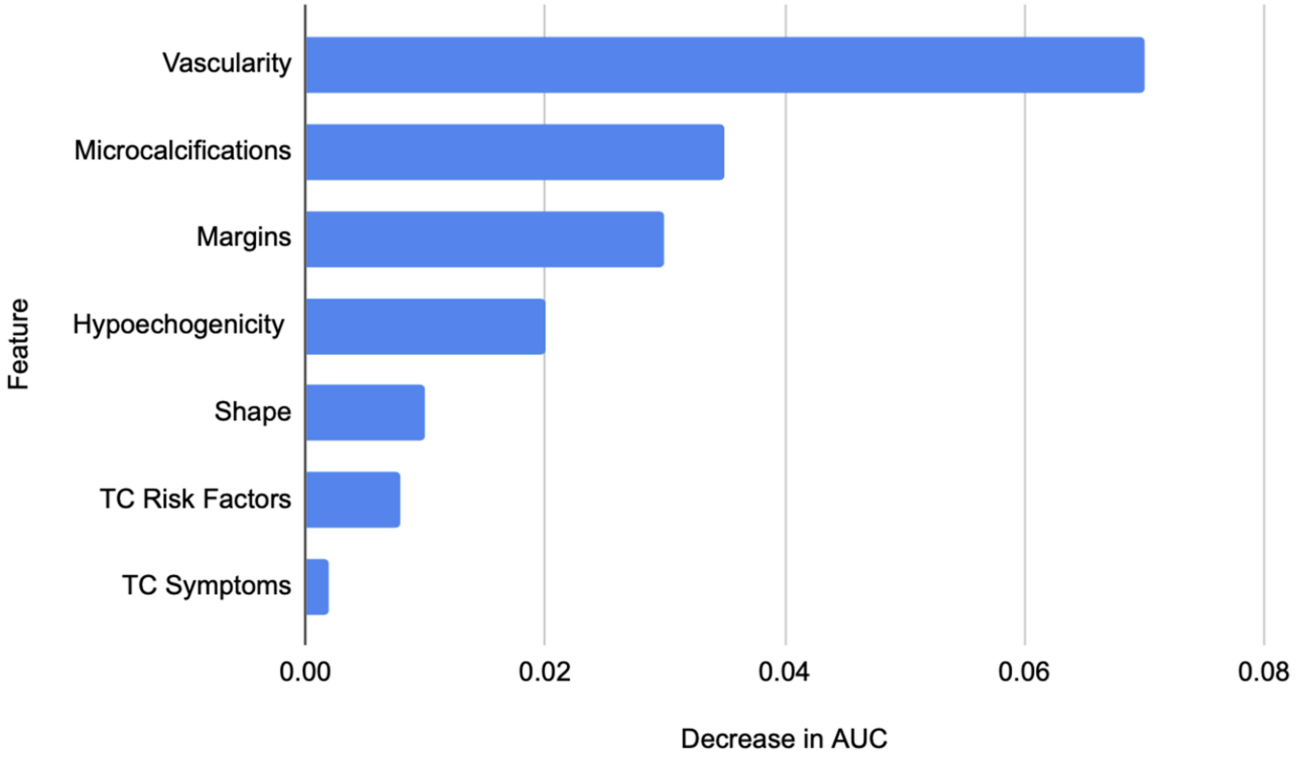

Figure 5 Feature importance analysis identified vascularity, followed by the presence of microcalcifications and the type of nodule margin, as the most decisive factors in determining nodule malignancy.
Abbreviations: TC, thyroid cancer, AUC, area under the curve.

nodule vascularity emerged as the most influential factor in decision-making. Microcalcifications, irregular margins, and hypoechogenicity also played a significant role in determining the malignancy of a thyroid nodule.

Discussion

In this study, we analyzed the effectiveness of various models in diagnosing TC based on data extracted from ultrasonography imaging and patients' medical histories. This study aimed to investigate whether accurate TC diagnosis is possible using only non-invasive diagnostic techniques supported by ML models, which would be less burdensome for the patient than the classic diagnostic approach consisting of ultrasound and FNAB.

To validate the effectiveness of our method, we compared our results with findings from recent publications that explored the diagnostic accuracy of TC using ultrasonography and FNAB. In the study by Alhajlan et al, the authors reported a recall of 83.33%, an ROC-AUC of 87.8%, and a specificity of 79.14%.[26] Another study by Mehanna et al reported a recall of 87% and a notably lower specificity of 67%.[27]

In contrast, our Random Forest model achieved a specificity of 93.9%, an ROC-AUC of 84.3%, and a recall of 61.6%. These results indicate that our approach substantially improves specificity and presents a notably lower recall compared to the cited studies. The higher specificity suggests that our model is more effective in correctly identifying benign cases, which is crucial for reducing unnecessary biopsies. The slightly lower ROC-AUC in our case may be attributed to the sole use of non-invasive input features, which supports the clinical significance of our findings: high diagnostic accuracy can be achieved without relying on invasive procedures. A lower recall score indicates that our model performed worse in terms of correctly identifying malignant nodules, which could be attributed to data class imbalance and the application of the SMOTE.

Malignant cases occur much less frequently than benign ones, with an overall malignancy rate reported between 7% and 15% in the general population.[28] In our dataset, which consists of 5928 patients, only 10.5% (n = 623) of cases were malignant, confirming the naturally occurring class imbalance. This skewed distribution can negatively affect the

performance of classification models, often resulting in models that are biased toward the majority class, benign TNs, and thus less reliable for identifying malignant cases.[29]

One way to address this issue is by undersampling the majority class to achieve balance, however, this approach sacrifices a large portion of the data, potentially reducing variability and limiting the model's ability to generalize.[30] In our study, we applied the SMOTE to balance the training dataset. The SMOTE addresses class imbalance by generating synthetic samples of the minority class, thereby improving predictive accuracy without discarding valuable data.[29] This method has been successfully applied in other domains facing similar imbalance challenges, such as breast cancer detection and cervical cancer classification.[31,32] This approach is particularly effective when working with low-dimensional data, where the number of features is relatively small compared to the number of samples.[29] Although research has shown that oversampling techniques like the SMOTE often outperform undersampling methods in terms of classification outcomes, in our case, the models' ability to correctly classify malignant nodules was lower than in previously mentioned studies, despite the use of the SMOTE.[26,27,32,33]

The size of the dataset is widely recognized as having a key influence on the models' performance and generalizability. Larger cohorts tend to improve classification accuracy and reduce the risk of overfitting.[28,34] In our study, we analyzed data from a cohort of 5928 patients, which provided an extensive and diverse set of features for model training. This large-scale dataset allowed the algorithm to learn from a wide variety of clinical presentations and imaging patterns, thereby enhancing its ability to correctly diagnose new, unseen cases. Therefore, our best-performing model classified thyroid nodules with 90.5% accuracy, slightly higher than the accuracy reported in similar studies based on smaller datasets; for instance, Xi et al achieved 79.3% accuracy using a cohort of 724 patients.[15] The substantial volume of data used in our work is thus a major strength, supporting the reliability of the obtained results and the clinical relevance of the proposed approach. Nevertheless, while overall accuracy was high, it was largely driven by the model's ability to correctly detect benign nodules. Improving sensitivity remains a critical challenge to ensure malignant cases are reliably identified. Future efforts should focus on increasing recall through integration of additional clinical and imaging data, such as cytology results or more detailed vascular and structural ultrasound features. Strengthening the model's sensitivity without compromising specificity would help avoid missed cancers while still preventing overtreatment of benign lesions.

Recent studies, including multicenter analyses based on raw ultrasound images, have demonstrated the potential of deep learning and convolutional neural networks for TN classification with comparable or superior performance to that of radiologists in internal and external test sets.[35,36] However, such approaches often require large, well-annotated image datasets, are less interpretable, and remain difficult to integrate into routine clinical workflows.[37,38] In contrast, the present study deliberately focuses on feature-based ML models using routinely assessed ultrasound descriptors, prioritizing interpretability and real-world applicability over maximal standalone diagnostic performance. Our study lies not in algorithmic innovation, but in the large-scale, real-world evaluation of commonly used ML models applied to a limited and clinically interpretable set of ultrasound features, highlighting both their potential utility and their current limitations in preoperative risk stratification.

In this research, the following ultrasonographic characteristics of nodules, typically associated with malignancy, were taken into account: hypoechogenicity, presence of microcalcifications, irregular shape, irregular margins, and high vascularity.[39–41] By conducting feature importance analysis, we wanted to find out which of these characteristics play a more significant role in determining the malignancy of nodules. The results of our study identified high vascularity and microcalcifications as the most critical determinants in diagnosing TC using the tested models. Our results align with Xi et al's findings, who also identified these features as highly predictive.[15] However, a study by Li et al reports that the presence of irregular margins contributed the most to the models' judgement.[42] Identifying the most significant features of TC is key to improving diagnostic accuracy and efficiency. Importantly, the moderate sensitivity observed in this study underscores the challenges of relying solely on a limited set of ultrasound features for malignancy prediction and supports the role of ML models as decision-support tools rather than stand-alone diagnostic systems.

In TC diagnostics, missed malignancies carry significant clinical risks including disease progression, potential metastasis, and delayed treatment.[43] With 48 out of 125 malignant cases undetected in our test cohort, the observed sensitivity remains insufficient for stand-alone clinical decision-making. Therefore, this approach should be viewed as complementary to, rather than replacement for, established risk stratification frameworks. Current guidelines including

EU-TIRADS and ATA recommendations provide structured approaches to nodule evaluation based on multiple sonographic features.[44,45] Our model could potentially integrate with these systems as an adjunctive decision-support tool, particularly when clinical-sonographic assessment yields indeterminate results. For instance, nodules classified as EU-TIRADS 4 (intermediate risk, 6–17% malignancy) or Bethesda III/IV cytology might benefit from ML-based probability estimates to refine surgical decision-making. However, given the current sensitivity limitations, the model's primary utility lies in confirming benign classification (high specificity: 93.9%) rather than ruling out malignancy.

Beyond conventional ultrasonography, emerging imaging modalities have been investigated for thyroid nodule diagnosis, aiming to capture tissue characteristics not accessible with standard ultrasound. Second harmonic generation (SHG) microscopy has demonstrated the ability to visualize microstructural features of thyroid nodules, including heterogeneity of collagen organization associated with malignant pathology.[46,47] Recent studies further showed that combining such advanced imaging data with supervised machine learning can improve diagnostic discrimination by explicitly leveraging high-dimensional, image-derived features.[19] While these approaches remain outside routine clinical practice, they illustrate the broader applicability of machine learning as an analytical framework for integrating diverse imaging modalities. In this context, the machine learning strategy presented in the current study can be viewed as a clinically accessible implementation of this approach, based on routinely available ultrasound features.

Limitations of the Study

This study has several limitations. First, the US characteristics of the nodules were extracted manually by clinicians, which may introduce bias and affect data reliability.[48] Recent studies have demonstrated the potential of deep learning models, such as convolutional neural networks (CNNs), to automatically identify TNs from ultrasound images and correctly classify them.[49–51] These algorithms analyze complex image features that are beyond human perception. When compared to experienced radiologists, these models have shown comparable accuracy and even superior diagnostic performance in detecting benign nodules.[52] Second, we acknowledge the potential issue of era bias due to the 15-year study period, as changes in ultrasound technology and evolving diagnostic standards (eg., EU-TIRADS) over time may have introduced variability affecting data interpretation and model performance. Moreover, we recognize that the use of vascularity as a simplified binary feature (high vs. low) may not fully capture the complexity of blood flow patterns within thyroid nodules; this approach, though consistent with clinical practice, could introduce bias and affect feature importance ranking, highlighting the need for more detailed, prospective, multi-center data in future studies. Additionally, the retrospective nature of the study and the long observation period may introduce variability related to operator experience. The analysis was also based on a limited set of routinely assessed ultrasound features, which may have constrained model sensitivity but was intentionally chosen to reflect real-world clinical practice. These limitations are unavoidable in retrospective study designs. Finally, all patient records analyzed in this study came from a single medical center. Future studies should focus on external validation by training and testing these models on datasets from multiple centers to enhance accuracy and minimize random diagnostic errors.

Recent studies highlight the importance of a multimodal diagnostic approach in TN assessment, as reliance on a single diagnostic modality is associated with relevant clinical uncertainty, particularly in indeterminate nodules.[53] US-based evaluation, although fundamental in initial risk stratification, is limited by substantial overlap between benign and malignant sonographic features and by inter-observer variability, which restricts its standalone diagnostic value. These limitations contribute to the persistent clinical dilemma of indeterminate nodules, where neither ultrasound nor cytology alone can reliably exclude malignancy. In this context, the assessment and integration of potential markers or risk factors of malignancy are considered essential for improving diagnostic accuracy.[54–56] Consequently, current evidence supports the use of ultrasound-derived features as part of an integrated diagnostic framework rather than as an isolated decision-making tool.[57]

Conclusion

Our study demonstrates that feature-based ML models can support the diagnostic assessment of TNs using non-invasive ultrasound features. The proposed approach achieved satisfactory performance in identifying benign lesions, which is

clinically relevant for reducing overtreatment and avoiding unnecessary surgical procedures. However, the current sensitivity of the model remains insufficient for stand-alone clinical decision-making, particularly with respect to reliable detection of malignant nodules. Therefore, the model should be considered a decision-support tool rather than an independent method replacing established diagnostic pathways. Further improvements will require expansion of the dataset – especially with a higher proportion of malignant cases – and integration of additional diagnostic inputs, including extended ultrasound characteristics and cytological data from FNAB. Such multimodal approaches may enhance diagnostic reliability and, in the longer term, contribute to more cost-effective and individualized management of patients with TNs.

Abbreviations

TC, thyroid cancer; TNs, thyroid nodules; FNAB, fine-needle aspiration biopsy; AI, artificial intelligence; ML, machine learning; CA, classification accuracy; Prec, precision; Rec, recall; ROC-AUC, receiver operating characteristic-area under the curve; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives; MLP, Multilayer Perceptron (Neural Network); GBMs, Gradient Boosting Machines.

Ethics Approval

This study was conducted in accordance with the Declaration of Helsinki and approved by the Bioethics Committee of the Wroclaw Medical University, Poland. Approval No. KB-241/2023; 24 February 2023.

Author Contributions

All authors made a significant contribution to the work reported, whether that was in the conception, study design, execution, acquisition of data, analysis, or interpretation; took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research was funded by subsidy of the Wroclaw Medical University, Poland (internal subsidy number: SIMPLE SUBZ.A530.26.029).

Disclosure

The author(s) report no conflicts of interest in this work.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229–263. doi:10.3322/caac.21834
2. Chen DW, Lang BHH, McLeod DSA, Newbold K, Haymart MR. Thyroid cancer. *The Lancet*. 2023;401(10387):1531–1544. doi:10.1016/S0140-6736(23)00020-X
3. Chouhan L, Manmohan M, Pasoria S, Tavri O. Sonographic, cytological and histopathological characteristics of spectrum of thyroid nodules: a comparative analysis. *IJSR*. 2024;13(6):813–819. doi:10.21275/SR24610164446
4. Ali SZ, Baloch ZW, Cochand-Priollet B, Schmitt FC, Vielh P, VanderLaan PA. The 2023 bethesda system for reporting thyroid cytopathology. *Thyroid*. 2023;33(9):1039–1044. doi:10.1089/thy.2023.0141
5. Jin X, Jing X. Cytologic assessment of thyroid nodules – updates in 2023 bethesda reporting system, diagnostic challenges and pitfalls. *Hum Pathol Rep*. 2024;36(1):300743. doi:10.1016/j.hpr.2024.300743
6. Almquist M, Muth A. Surgical management of cytologically indeterminate thyroid nodules. *Gland Surg*. 2019;8(Suppl 2):S105–S111. doi:10.21037/gs.2019.01.03
7. Mavromati M, Saiji E, Demarchi MS, et al. Unnecessary thyroid surgery rate for suspicious nodule in the absence of molecular testing. *Eur Thyroid J*. 2023;12(6):e230114. doi:10.1530/ETJ-23-0114
8. Barranco H, Fazendin J, Lindeman B, Chen H, Ramonell KM. Thyroid hormone replacement following lobectomy: long-term institutional analysis 15 years after surgery. *Surgery*. 2023;173(1):189–192. doi:10.1016/j.surg.2022.05.044
9. Corso C, Gomez X, Sanabria A, Vega V, Dominguez LC, Osorio C. Total thyroidectomy versus hemithyroidectomy for patients with follicular neoplasm. A cost-utility analysis. *Int J Surg*. 2014;12(8):837–842. doi:10.1016/j.ijsu.2014.07.005
10. Lin JS, Aiello Bowles EJ, Williams SB, Morrison CC. *Screening for Thyroid Cancer: A Systematic Evidence Review for the U.S. Preventive Services Task Force*. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017. Report No.: 15-05221-EF-1.

11. Nickel B, Tan T, Cvejic E, et al. Health-related quality of life after diagnosis and treatment of differentiated thyroid cancer and association with type of surgical treatment. *JAMA Otolaryngol Head Neck Surg*. 2019;145(3):231–238. doi:10.1001/jamaoto.2018.3870
12. Yan K, Li T, Marques JAL, Gao J, Fong SJ. A review on multimodal machine learning in medical diagnostics. *Math Biosci Eng*. 2023;20 (5):8708–8726. doi:10.3934/mbe.2023382
13. Fiorentino V, Pizzimenti C, Franchina M, et al. The minefield of indeterminate thyroid nodules: could artificial intelligence be a suitable diagnostic tool? *Diagn Histopathol*. 2023;29(8):396–401. doi:10.1016/j.mpdhp.2023.06.013
14. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 2019;20(5):e262–e273. doi:10.1016/S1470-2045(19)30149-4
15. Xi NM, Wang L, Yang C. Improving the diagnosis of thyroid cancer by machine learning and clinical data. *Sci Rep*. 2022;12(1):11143. doi:10.1038/s41598-022-15342-z
16. Guerrisi A, Seri E, Dolcetti V, et al. A machine learning model based on thyroid US radiomics to discriminate between benign and malignant nodules. *Cancers*. 2024;16(22):3775. doi:10.3390/cancers16223775
17. Guo YY, Li ZJ, Du C, et al. Machine learning for identifying benign and malignant of thyroid tumors: a retrospective study of 2,423 patients. *Front Public Health*. 2022;10:960740. doi:10.3389/fpubh.2022.960740
18. Prete A, Borges de Souza P, Censi S, Muzza M, Nucci N, Sponziello M. Update on fundamental mechanisms of thyroid cancer. *Front Endocrinol*. 2011:102.
19. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European thyroid association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *Eur Thyroid J*. 2017;6(5):225–237. doi:10.1159/000478927
20. Khayat M, Halawani R, Dhahi T, et al. Ultrasound characteristics of thyroid nodules: differentiating benign from malignant nodules using histopathology as the gold standard. *SJR*. 2024;3(RSSA):64–75.
21. Karagülle M, Arslan FZ, Şimşek S, et al. Investigation of the effectiveness of microvascular doppler ultrasound and Q-Pack in the discrimination of malign thyroid nodules from benign. *Ultrasound Q*. 2023;39:37–46.
22. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence – Volume 2; 1995: 1137–1143; San Francisco, CA. Morgan Kaufmann Publishers.
23. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. *IJDKP*. 2015;5(2):1–11.
24. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45(4):427–437. doi:10.1016/j.ipm.2009.03.002
25. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
26. Alhajlan M, Al-Masabi M, Al Mansour M, et al. The accuracy of fine-needle aspiration cytology and ultrasonography in assessing thyroid nodules in correlation with histopathology: a retrospective study. *Ann Med Surg*. 2024;86(12):7002–7009. doi:10.1097/MS9.0000000000002676
27. Mehanna H, Nankivell P, Boelaert K, et al. Diagnostic performance of ultrasound vs ultrasound-guided FNAC in thyroid nodules: data from the ElaTION trial. *J Clin Endocrinol Metab*. 2025;110(7):1997–2006. doi:10.1210/clinem/dgae682
28. Prusa J, Khoshgoftaar TM, Seliya N. The effect of dataset size on training tweet sentiment classifiers. IEEE 14th International Conference on Machine Learning and Applications (ICMLA); Miami, FL. Institute of Electrical and Electronics Engineers; 2015.
29. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform*. 2013;14:1–16.
30. Araf I, Idri A, Chairi I. Cost-sensitive learning for imbalanced medical data: a review. *Artif Intell Rev*. 2024;57:1–72. doi:10.1007/s10462-023-10652-8
31. Fallahi A, Jafari S. An expert system for detection of breast cancer using data preprocessing and bayesian network. *IJAST*. 2011;34:65–70.
32. Karamti H, Alharthi R, Anizi AA, et al. Improving prediction of cervical cancer using KNN imputed SMOTE features and multi-model ensemble learning approach. *Cancers*. 2023;15(17):4412. doi:10.3390/cancers15174412
33. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J Biomed Inform*. 2019;90:103089. doi:10.1016/j.jbi.2018.12.003
34. Althnian A, AlSaeed D, Al-Baity H, et al. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Appl Sci*. 2021;11(2):796. doi:10.3390/app11020796
35. Kim YJ, Choi Y, Hur SJ, et al. Deep convolutional neural network for classification of thyroid nodules on ultrasound: comparison of the diagnostic performance with that of radiologists. *Eur J Radiol*. 2022;152:110335. doi:10.1016/j.ejrad.2022.110335
36. Ko SY, Lee JH, Yoon JH, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head Neck*. 2019;41 (4):885–891. doi:10.1002/hed.25415
37. Wei X, Gao M, Yu R, et al. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. *Med Sci Monit*. 2020;26(e926096). doi:10.12659/MSM.926096.
38. Tao Y, Yu Y, Wu T, et al. Deep learning for the diagnosis of suspicious thyroid nodules based on multimodal ultrasound images. *Front Oncol*. 2022;12:1012724. doi:10.3389/fonc.2022.1012724
39. Hong YJ, Son EJ, Kim EK, Kwak JY, Hong SW, Chang HS. Positive predictive values of sonographic features of solid thyroid nodule. *Clin Imaging*. 2010;34(2):127–133. doi:10.1016/j.clinimag.2008.10.034
40. Frates MC, Benson CB, Doubilet PM, Cibas ES, Marqusee E. Can color Doppler sonography aid in the prediction of malignancy of thyroid nodules? *J Ultrasound Med*. 2003;22(2):127–131. doi:10.7863/jum.2003.22.2.127
41. Papini E, Guglielmi R, Bianchini A, et al. Risk of malignancy in nonpalpable thyroid nodules: predictive value of ultrasound and color-Doppler features. *J Clin Endocrinol Metab*. 2002;87(5):1941–1946. doi:10.1210/jcem.87.5.8504
42. Li W, Hong T, Fang J, et al. Incorporation of a machine learning pathological diagnosis algorithm into the thyroid ultrasound imaging data improves the diagnosis risk of malignant thyroid nodules. *Front Oncol*. 2022;12:968784. doi:10.3389/fonc.2022.968784
43. Schlumberger M, Leboulleux S. Current practice in patients with differentiated thyroid cancer. *Nat Rev Endocrinol*. 2021;17(3):176–188. doi:10.1038/s41574-020-00448-z
44. Ringel MD, Sosa JA, Baloch Z, et al. 2025 American thyroid association management guidelines for adult patients with differentiated thyroid cancer. *Thyroid*. 2025;35(8):841–985. doi:10.1177/10507256251363120

45. Durante C, Hegedüs L, Czarniecka A, et al. 2023 European thyroid association clinical practice guidelines for thyroid nodule management. *Eur Thyroid J*. 2023;12(5):e230067. doi:10.1530/ETJ-23-0067
46. Hristu R, Eftimie LG, Stanciu SG, et al. Quantitative second harmonic generation microscopy for the structural characterization of capsular collagen in thyroid neoplasms. *Biomed Opt Express*. 2018;9(8):3923–3936. doi:10.1364/BOE.9.003923
47. Padrez Y, Hristu R, Timoshchenko I, Eftimie LG, Rutkauskas D, Golubewa L. Supervised machine learning thyroid carcinoma diagnosis using wide-field SHG microscopy. *IEEE Access*. 2025;13:112021–112038. doi:10.1109/ACCESS.2025.3583435
48. Nguyen DT, Kang JK, Pham TD, Batchuluun G, Park KR. Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. *Sensors*. 2020;20(7):1822. doi:10.3390/s20071822
49. Maarouf AA, Meriem H, Hachouf F. Deep learning and handcrafted features for thyroid nodule classification. *Int J Imaging Syst Technol*. 2024;34 (e23215). doi:10.1002/ima.23215
50. Ma J, Wu F, Jiang T, Zhao Q, Kong D. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2017;12(11):1895–1910. doi:10.1007/s11548-017-1649-7
51. Kim J, Kim MH, Lim DJ, et al. Deep learning technology for classification of thyroid nodules using multi-view ultrasound images: potential benefits and challenges in clinical application. *Endocrinol Metab*. 2025;40(2):216–224. doi:10.3803/EnM.2024.2058
52. Wang L, Yang S, Yang S, et al. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *Worl J Surg Onc*. 2019;17:1–9.
53. Lebbink CA, Links TP, Czarniecka A, et al. 2022 European thyroid association guidelines for the management of pediatric thyroid nodules and differentiated thyroid carcinoma. *Eur Thyroid J*. 2022;11(6):e220146. doi:10.1530/ETJ-22-0146
54. Cardisciani L, Policardo F, Tralongo P, Fiorentino V, Rossi ED. What psammoma bodies can represent in the thyroid. What we recently learnt from a story of lack of evidence. *Pathologica*. 2022;114(5):373–375. doi:10.32074/1591-951X-815
55. Pizzimenti C, Fiorentino V, Ieni A, et al. BRAF-AXL-PD-L1 signaling axis as a possible biological marker for RAI treatment in the thyroid cancer ATA intermediate risk category. *Int J Mol Sci*. 2023;24(12):10024. doi:10.3390/ijms241210024
56. Kitahara CM, Schneider AB. Epidemiology of thyroid cancer. *Cancer Epidemiol Biomarkers Prev*. 2022;31(7):1284–1297. doi:10.1158/1055-9965. EPI-21-1440
57. Patel J, Klopper J, Cottrill EE. Molecular diagnostics in the evaluation of thyroid nodules: current use and prospective opportunities. *Front Endocrinol*. 2023;14:1101410. doi:10.3389/fendo.2023.1101410