

Submitted:  
08.07.2025  
Accepted:  
16.12.2025  
Published:  
31.12.2025

## Application of data science methods, including machine learning, in the classification of focal lesions in the thyroid gland

Paweł Mariusz Gadzicki<sup>1</sup>, Małgorzata Krzywicka<sup>2</sup>, Katarzyna Dobruch-Sobczak<sup>3</sup>, Bartosz Migda<sup>4</sup>, Ewelina Szczepanek-Parulska<sup>5</sup>, Agnieszka Wosiak<sup>2</sup>, Zbigniew Adamczewski<sup>1</sup>

<sup>1</sup> Department of Nuclear Medicine, Medical University of Lodz, Poland

<sup>2</sup> Institute of Information Technology, Lodz University of Technology, Poland

<sup>3</sup> Radiology Department II, The Maria Skłodowska-Curie National Research Institute of Oncology, Poland

<sup>4</sup> Diagnostic Ultrasound Lab, Department of Pediatric Radiology, Medical University of Warsaw, Poland

<sup>5</sup> Department of Endocrinology, Metabolism and Internal Medicine, Poznan University of Medical Sciences, Poland

Corresponding author: Paweł Gadzicki; e-mail: pawel.gadzicki@umed.lodz.pl

DOI: 10.15557/JoU.2025.0036

### Keywords

thyroid cancer;  
machine learning;  
AI;  
classification  
of focal lesions

### Abstract

**Aim:** The aim of the study was to train, evaluate, and optimize machine learning models for classifying focal lesions in the thyroid gland as benign or malignant based on their features. **Material and methods:** A dataset of 841 focal thyroid lesions described by 17 features (ultrasonographic and patient characteristics) was considered. Using the Python programming language, statistical and then exploratory data analyses were conducted using the libraries, including the generation of graphs and heat maps of correlations between the considered features. Binary classification models were selected to categorize the focal lesion on the basis of their characteristics into one of two classes (benign lesion, malignant lesion). The following models were used: logistic regression-based, support vector machine-based, k-nearest neighbor model, Random Forest model, and decision tree classifier. We applied formulas to select those focal lesion features that most contributed to the models' classification decisions. The final dataset consisted of 841 focal thyroid lesions described by seven ultrasonographic features and histopathological assessment of malignancy (benign or malignant). Classifiers were validated using 10-fold cross-validation. Model performance was evaluated using sensitivity, accuracy, measure-F<sub>1</sub>, precision, area under the ROC curve, PPV, NPV, specificity. **Results:** The best-performing model (in term of sensitivity) was the classifier based on a support vector machine: sensitivity = 71.17%, accuracy = 83.24%, area under the ROC curve = 84.86%, measure f1 = 69.13%, precision = 68.85%, PPV = 68.49%, NPV = 89.06%. **Conclusions:** The study demonstrates the usefulness of data science methods in predicting the malignant nature of focal lesions in the thyroid gland. It proves that classification decisions made by the studied models are based on specific ultrasonographic features associated with increased or reduced risk of malignancy.

## Introduction

The incidence of thyroid cancer is steadily increasing. Nodular disease is five times more common in women, and its risk rises with age<sup>(1)</sup>. Although thyroid nodular disease is one of the most common endocrine diseases encountered in clinical practice, the risk of malignancy of a lesion diagnosed in the thyroid gland is relatively low (about 3–10%). Ultrasound in combination with biopsy plays an essential role in the initial differential diagnosis of thyroid focal lesions. Qualification for biopsy, on the other hand, is based on the assessment of ultrasound features<sup>(2–5)</sup>.

In the context of growing morbidity, it is worth noting the increasing number of tools developed to support the diagnostic process<sup>(6,7)</sup>.

Some of these tools are based on data science, which involves extracting knowledge from data, using statistics, data analysis, machine learning, and other related methods<sup>(8)</sup> (Fig. 1).

We employed data science methods to select and optimize the best machine learning model for classifying focal lesions in the thyroid gland as benign or malignant based on their characteristics.

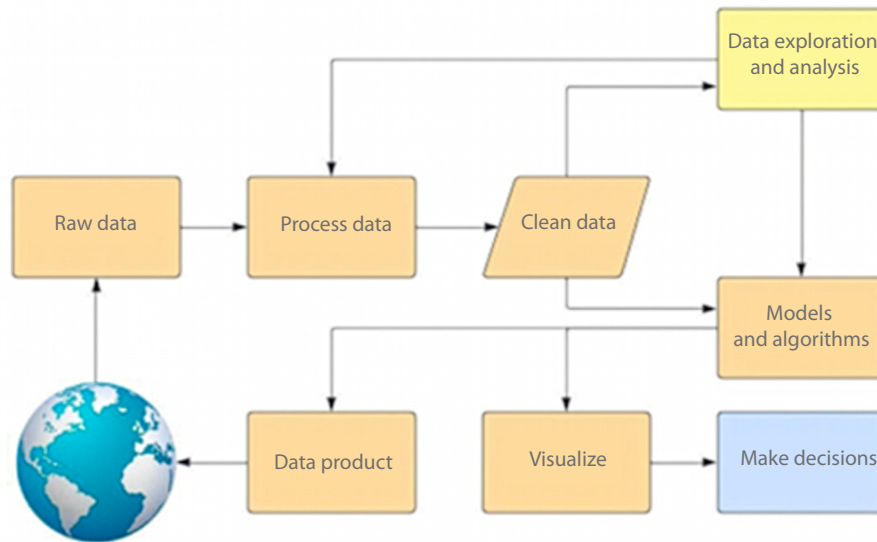


Fig. 1. Data science process

Machine learning is a branch of artificial intelligence (AI). AI refers to the ability to program computers (or more broadly, machines) to solve complicated and usually very time-consuming tasks. An example of a time-consuming and complicated task is the extraction of useful information (data mining) from large or complex clinical datasets.

Two major classes of machine learning algorithms exist: supervised and unsupervised learning. The first class is mainly used to predict outcomes using input features, whereas the second class is used to cluster unlabeled data.

Supervised algorithms are applied when learning to predict future outcomes, such as the presence of malignancy in a focal lesion based on labeled data (e.g., focal lesions with known ultrasonographic features that appear to be either malignant or benign)<sup>(9)</sup>.

## Materials and methods

The study used a dataset of the authors of previously published research<sup>(1)</sup>. The dataset consisted of 841 thyroid nodules evaluated by experienced sonographers and subsequently by histopathologists. The dataset consisted of related histopathological features, ultrasound features of nodules, and patient characteristics. The authors' database was collected from January 2009 to July 2018. All thyroid nodules diagnosed as benign or malignant on the basis of the final histopathological examination of the resected specimens were included in this study. Ultrasound examinations were performed by one of five sonographers with 9 to 22 years of experience in thyroid imaging. Final histopathological diagnoses were obtained after thyroidectomy for all 841 analyzed nodules. Among the 229 malignant neoplasms, papillary thyroid carcinoma (PTC) was the most common (184), while hyperplastic lesions were the most common among the benign lesions. The pathologists were blinded to the results of the ultrasound examination. The goal of the study was to validate the EU-TIRADS system in a multi-institutional database of thyroid nodules by analyzing the relationship between EU-TIRADS scores and histopathology results. This analysis constitutes a unique large multicenter cohort analysis of the diagnostic performance of

the EU-TIRADS scale in a previously iodine-deficient region. The authors concluded that the application of the EU-TIRADS guidelines resulted in moderate specificity. The vast majority of malignancies classified as EU-TIRADS 3, 4, and 5 would not have been recommended for biopsy because they are smaller in size than those proposed in the classification.

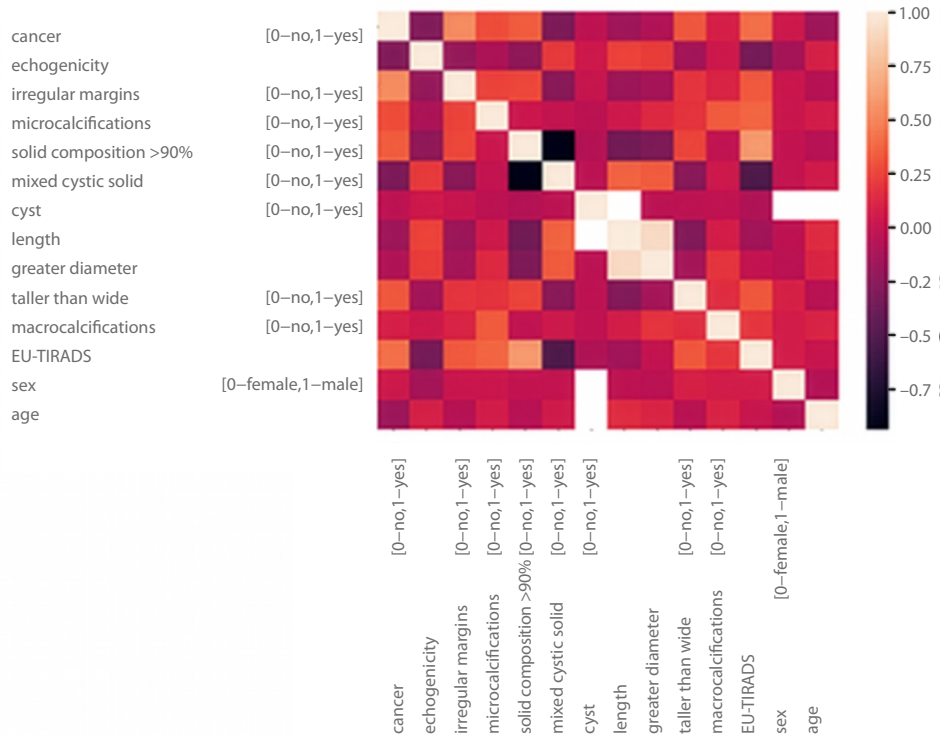
Specific methods and materials from the study whose database was used are summarized in the supplementary material (Suppl. 1).

Initially, a set of 841 nodules was considered (one lesion instance was missing from the database). The lesions were described by 17 features (ultrasound, patient-related, histopathological) and in order to use all instances of the set (all lesions examined by ultrasound specialists), features that were not reported for all surgically removed focal lesions were excluded. In the next step, using the Python programming language in Jupyter Notebook environment (Google Colab), with the NumPy and Pandas libraries, the completeness of the data and its types were checked. What is more, in the following steps, methods from the sklearn Python library were applied; they are listed and explained in greater detail in the next paragraphs and in the technical supplement (Suppl. 2).

Statistical and then exploratory analysis of the data was performed. Using the Seaborn library, graphs illustrating relationships between features and heat maps were generated (visualizing the degree of correlation between the considered features – higher correlation, higher color saturation) (Fig. 2). Binary classifiers were selected, classifying lesions into one of two categories: malignant change or benign change.

A supervised learning process was applied to train models for future prediction, such as the presence of malignancy in focal lesions. Labeled data were used (like focal lesions that appear to be either malignant or benign, with known ultrasonographic features)<sup>(9)</sup>.

The input data consisted of ultrasound features of focal lesions in the thyroid gland, and the output was the predicted presence of malignancy.



**Fig. 2.** Feature correlation heatmap for one of the datasets (14 nodule features). The lighter the color, the stronger the positive correlation; the darker the color, the stronger the negative correlation. The exception is the white color, which indicates that it is impossible to determine the correlation due to insufficient data or a large number of gaps in the set for a given feature

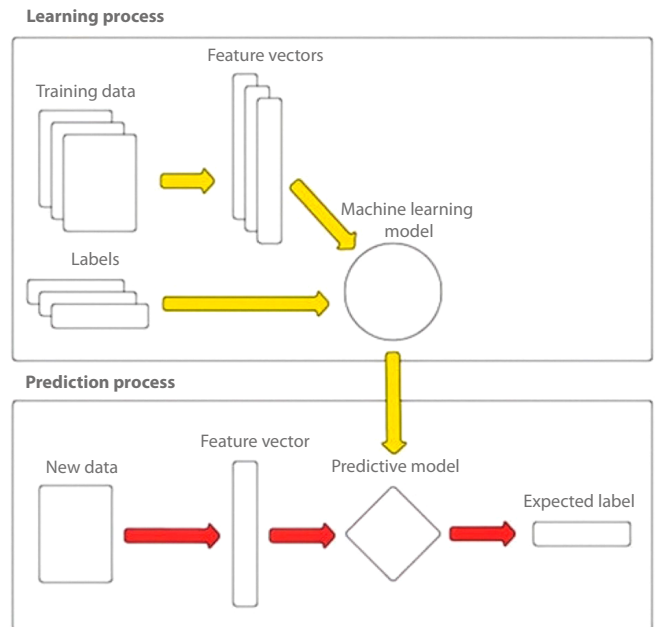
In supervised learning, analysis starts from a known training dataset, and an algorithm is used to infer predictions. The algorithm compares its output to the labels in order to modify it accordingly to match the expected values (Fig. 3).

In the field of data science, it is very important to validate the obtained models. In our work, we used 10-fold cross-validation with the proportion of predicted classes appropriate for the entire dataset (StratifiedKfold)<sup>(10)</sup>.

Another way to validate the model involved dividing the dataset into training, test, and validation sets, e.g., in the proportions of 80:10:10. This way of testing data requires more data than k-fold cross-validation to achieve good model results. The more features are considered in classification models, the larger the dataset needed to achieve good results. For this reason, validation was performed 10 times using cross-validation while maintaining the proportions of the predicted classes (benign change, malignant change).

In the context of algorithms belonging to the field of data science, it is important to select an appropriate classifier for the discussed issues and data type. Binary classifiers (assigning each lesion to one of two classes: benign or malignant) are a good choice for classifying focal lesions (Fig. 4).

In the next step, the following binary classifiers were applied: logistic regression-based (LogReg), support vector machine SVC-based classifier (SVC), k-nearest neighbor classifier (KNN), Random Forest (RF), and decision tree (DT). All classifiers were applied using dedicated methods from the sklearn Python library. Specific meth-



**Fig. 3.** Diagram of supervised machine learning

ods are described in the attached technical supplement (Suppl. 2). In addition to methods implementing classifiers, several other methods were used, which are described in the preceding and following paragraphs. Those are also, more technically, described in the attached supplement.

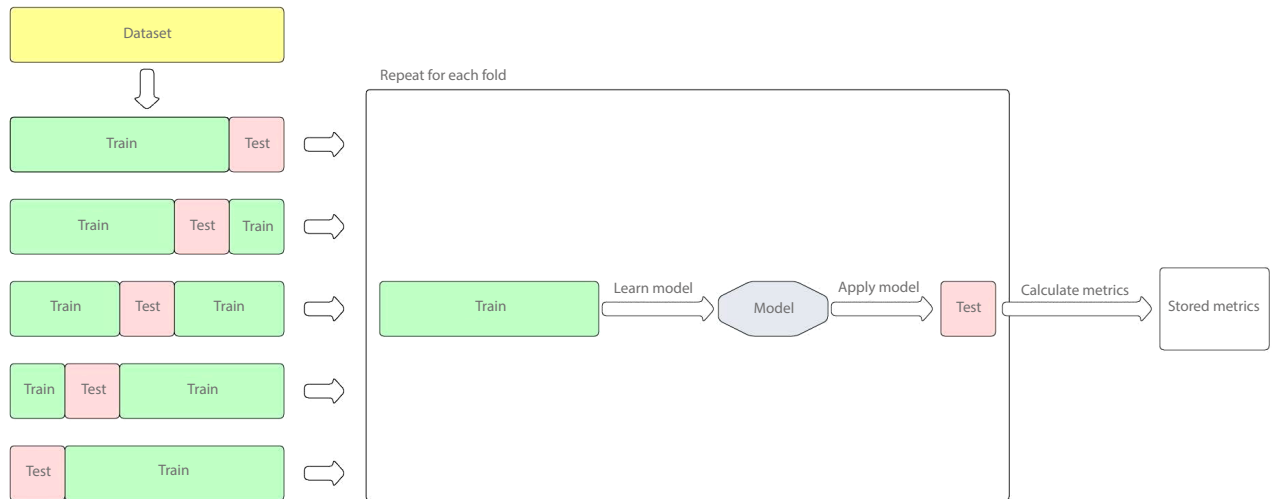


Fig. 4. K-fold cross-validation

The RFEC (recursive feature elimination with cross-validation) formula was used to identify the nodule features most important for the correct classification decision of the models (benign lesion/malignant lesion). The most significant features for the studied model were irregular margins, and a solid or almost solid composition of the lesion (i.e., >90%).

The following features describing the lesions were identified:

- irregular margins (yes/no),
- solid composition >90% (yes/no),
- mixed cystic-solid composition (yes/no),
- echogenicity – expressed on a four-point scale (I-very hypoechoic, II-hypoechoic, III-isoechoic, IV-hyperechoic),
- lesion taller than wide (yes/no),
- macrocalcifications (yes/no),
- spongiform nodule (yes/no),
- cyst (yes/no).

To ensure the reliability of classifier assessment, 10-fold cross-validation was applied while maintaining the proportion of predicted classes, appropriate for the entire dataset (StratifiedKFold). This means that the dataset was divided into 10 parts, each containing the same proportion of malignant and benign lesions. Nine of these parts were used to train the model and one for testing. This process

was repeated 10 times; each time, a different subset was used for testing. Classifier performance was calculated as the average of the results obtained.

The models were evaluated on the basis of mean values of such metrics as sensitivity, accuracy, F-measure, precision, and area under the ROC curve (AUC). Given the clinical consequences of the missing malignant lesion scenario, sensitivity was chosen as a major criterion. In the next steps, in the best models, SVC, LogReg, RF, and DT hyperparameter optimization (model “tuning”) was carried out with the help of the GridSearchCV library, thanks to which further improvement of model performance was obtained.

### Results

The results obtained using each binary classifier are presented in Table 1 and Figure 5.

### Discussion

In this study, the best results were achieved by the SVC classifier based on the support vector machine; the selection criterion was the highest sensitivity, for SVC sensitivity = 71.17% (Tab. 1).

Tab. 1. Performance metrics of the studied binary classifiers: Support vector machine-based classifier (SVC), Logistic regression-based (LogReg), Random Forest classifier (RF) and Decision Tree classifier (DT) , k-nearest neighbors (KNN)

	SVC	LogReg	RF	DT	KNN
<b>Sensitivity</b>	71.17%	68.99%	69.86%	69.86%	23.54%
<b>AUC</b>	84.86%	87.11%	84.57%	84.18%	64.91%
<b>Accuracy</b>	83.24%	83.96%	84.08%	84.19%	74.80%
<b>F-measure</b>	69.13%	69.34%	70.17%	69.93%	29.11%
<b>Precision</b>	68.85%	71.49%	70.9%	71.34%	40.40%
<b>PPV</b>	68.49%	71.17%	70.18%	69.7%	44.36%
<b>NPV</b>	89.05%	88.53%	88.74%	88.85%	80.7%
<b>Specificity</b>	87.74%	89.55%	88.89%	89.55%	75.82%

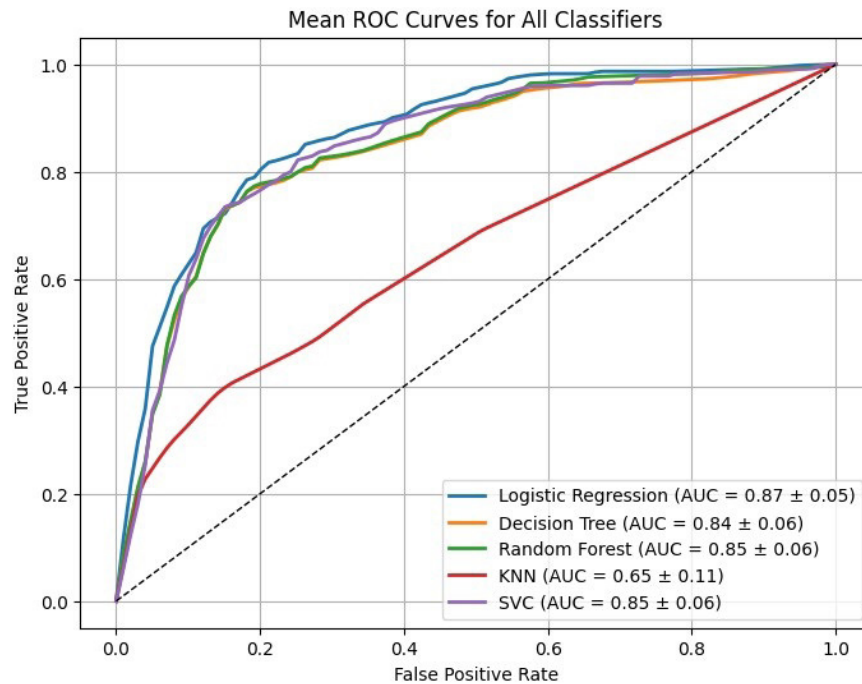


Fig. 5. Mean ROC curves for the studied binary classifiers: Support vector machine-based classifier (SVC), Logistic Regression-based (LogReg), Random Forest classifier (RF), and Decision Tree classifier (DT), k-nearest neighbor classifier (KNN)

Satisfactory performance was achieved in distinguishing between malignant and benign focal lesions (AUC ROC >80%, Accuracy >80%). We can compare our results with metrics reported by Osman Melih Topcuoglu *et al.*<sup>(11)</sup>, in which the diagnostic performance of six different currently available guidelines was compared, including the American College of Radiology Thyroid Imaging and Reporting Data System (ACR-TIRADS), Kwak-TIRADS, Korean TIRADS (K-TIRADS), European TIRADS (EU-TIRADS), American Thyroid Association (ATA), and Chinese TIRADS (C-TIRADS), in differentiating malignant from benign thyroid nodules. These guidelines aim to detect more thyroid cancers while reducing unnecessary biopsies. In that single-center study, between January 2007 and September 2023, ultrasound (US) images of thyroid nodules that were pathologically confirmed either by surgery or by fine needle aspiration biopsy (FNAB) were retrospectively evaluated and categorized according to six different currently available guidelines.

In that study, C-TIRADS offered the highest AUC, which was 84.2%. In our study, three of the four binary classification models that were trained and cross-validated achieved better results (SVC = 84.86%, LogReg = 87.11%, Random FC = 84.57%), while the DT classifier achieved comparable AUC results (84.18%).

Measured in the publication by Osman Melih Topcuoglu *et al.*, the accuracy of ACR-TIRADS, K-TIRADS, EU-TIRADS, ATA, or C-TIRADS (63.1%, 62.8%, 59.1%, 64.2%, 68.8%) was lower than the accuracy of any binary model classifiers assessed: SVC, Log Reg, RF, DT (83.24%, 83.96%, 84.08%, 84.19%).

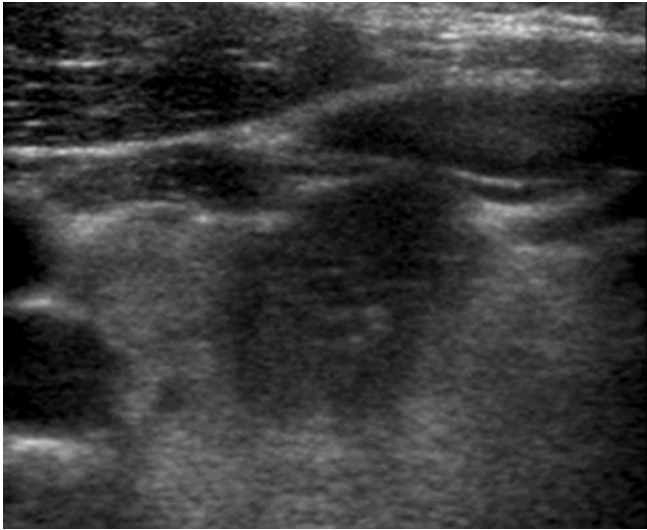
Osman Melih Topcuoglu *et al.* measured unnecessary FNAB rates (%) for ACR-TIRADS, K-TIRADS, EU-TIRADS, ATA, and C-TIRADS; they were 69.2%, 70.2%, 70.6%, 69.7%, and 71.7% respectively.

Conversely, sensitivity of any classification system (ACR-TIRADS, K-TIRADS, EU-TIRADS, ATA, or C-TIRADS, (99.8%, 97.8%, 97.6%, 97.8%, 97.5%, 92.8%)) in that study was higher than the sensitivity that our models achieved (SVC, Log.Reg, RF, DT, 71.17%, 68.99%, 69.86%, 69.86%)<sup>(11)</sup>.

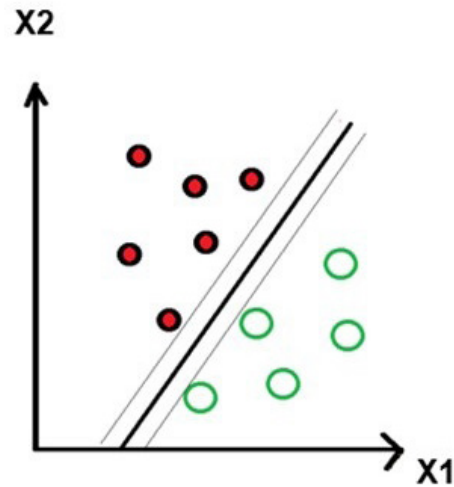
As it turned out, the models obtained the best results when classification decisions were made on the basis of features commonly associated with increased or reduced risk of malignancy. However, these features were inferred by the Recursive Feature Elimination process. The RFEC formula was used, thanks to which the nodule features most important for the correct classification decision of the models (benign lesion/malignant lesion) were identified. The most significant features for the studied model included irregular margins and solid or almost solid composition of the lesion (>90%). It corresponds to the US features most significantly associated with malignancy as identified by Dobruch-Sobczak *et al.*, irregular margins (Odds Ratio = 13.82), which was also associated with high specificity, and solid or almost solid composition (Odds Ratio = 9.82)<sup>(1,12)</sup> (Fig. 6).

The authors of the article, in which histopathological images were classified, among others, based on some of their features also expressed in numbers, referred to the algorithms we used and demonstrated their high accuracy and classification efficiency. In that study, these classic classifiers achieved higher effectiveness than the more complex deep learning models often used today, requiring greater computational power<sup>(13)</sup>. In our research, classic classifiers (SVC, LogReg, RF, and DT models) achieved consistent results, as shown in Table 1.

How the best one – in terms of sensitivity – Support Vector Machine-based classifier (SVC) works can be easily understood. This classifier uses the Support Vector Machine technique, which is



**Fig. 6.** Thyroid focal lesion: solid composition, hypoechogenicity, irregular margins, irregular shape, and extrathyroidal expansion



**Fig. 7.** Maximum-margin hyperplane and margins for an SVM trained on samples from two classes (red and green circles)

based on determining a hyperplane (represented as a line in the figure) separating (in this case, malignant from benign nodules) as far points as possible belonging to different classes (i.e., benign nodules and malignant nodules)<sup>(14–16)</sup>. Each point corresponds to an instance (focal lesions) and is described by a set of features that “locate” it in space (in Fig. 7) by two features, which makes it possible to place it in a two-dimensional space that is easy to imagine; in the studied dataset described by eight features, focal lesions are located in eight-dimensional space). The role of the hyperplane is to separate all benign from malignant nodules.

To conclude, deep learning strategies mimic the structure and function of the human brain. Instead of neurons, artificial neural networks (ANNs) are composed of linked computational units called nodes, arranged in layers. The information to be evaluated, such as a sonogram of a nodule, is presented in digital format to an input layer, and subsequently processed through a series of hidden layers comprising additional nodes. This approach can significantly improve the diagnostic accuracy of radiologists on thyroid nodule differentiation and could potentially decrease the number of unnecessary fine needle biopsies, especially when images are associated with clinical data<sup>(7,17,18)</sup>.

### Limitations of the study

Models that we train, optimize, and test have not been validated using an external dataset. Although the binary classifiers achieved satisfactory results in distinguishing malignant from benign focal lesions (AUC ROC >80%, Accuracy >80%), the best model’s sensitivity of 71.17% means that almost 30% of malignant lesions were not detected, which in a real-life scenario could have serious clinical implication. Domain knowledge (in that case, clinical insight) is cru-

cial in a data science approach; some clinical features are critical to identify certain subtypes of thyroid cancers included in the database.

### Conclusions

1. The study presents the usefulness of data science methods in predicting the malignant nature of focal lesions in the thyroid gland.
2. Classification decisions produced by the models studied are related to ultrasound features considered to reflect increased or reduced risk of malignancy. These conclusions were reached using data science reasoning.
3. Analysis of the present results and findings reported by other authors indicates that the use of the described algorithms can be an added value to the diagnostics performed by specialists.

### Conflict of interest

*The authors do not report any financial or personal connections with other persons or organizations which might negatively affect the contents of this publication and/or claim authorship rights to this publication.*

### Author contributions

*Original concept of study: PG. Writing of manuscript: PG, MK. Analysis and interpretation of data: PG, MK, AW, ZA. Final acceptance of manuscript: KDS, BM, ESP, ZA. Collection, recording and/or compilation of data: KDS, BM, ESP, ZA. Critical review of manuscript: PG, MK, KDS, BM, ESP, AW, ZA.*

## References

1. Dobruch-Sobczak K, Adamczewski Z, Szczepanek-Parulska E, Migda B, Woliński K, Krauze A *et al.*: Histopathological verification of the diagnostic performance of the EU-TIRADS classification of thyroid nodules-results of a multicenter study performed in a previously iodine-deficient region. *J Clin Med* 2019; 8: 1781. doi: 10.3390/jcm8111781.
2. 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: 225–237. doi: 10.1159/000478927.
3. Jarzab B, Dedećus M, Lewiński A, Adamczewski Z, Bakula-Zalewska E, Baldys-Waligórska A *et al.*: Diagnosis and treatment of thyroid cancer in adult patients – recommendations of polish scientific societies and the National Oncological Strategy. 2022 Update [Diagnostyka i leczenie raka tarczycy u chorych dorosłych – rekomendacje polskich towarzystw naukowych oraz Narodowej Strategii Onkologicznej. Aktualizacja na rok 2022]. *Endokrynol Pol* 2022; 73: 173–300. doi: 10.5603/EPa2022.0028.
4. Deng Y, Li H, Wang M, Li N, Tian T, Wu Y *et al.*: Global burden of thyroid cancer from 1990 to 2017. *JAMA Netw Open* 2020; 3: e208759. doi: 10.1001/jamanetworkopen.2020.8759.
5. Liu H, Ma AL, Zhou YS, Yang DH, Ruan JL, Liu XD, Luo BM: Variability in the interpretation of grey-scale ultrasound features in assessing thyroid nodules: A systematic review and meta-analysis. *Eur J Radiol* 2020; 129: 109050. doi: 10.1016/j.ejrad.2020.109050.
6. Habchi Y, Himeur Y, Kheddar H, Boukabou A, Atalla S, Chouchane A *et al.*: AI in thyroid cancer diagnosis: techniques, trends, and future directions. *Systems* 2023; 11: 519. doi: 10.3390/systems11100519.
7. Tessler FN, Thomas J: Artificial intelligence for evaluation of thyroid nodules: a primer. *Thyroid* 2023; 33: 150–158. doi: 10.1089/thy.2022.0560.
8. Hayashi C, Yajima K, Bock HH, Ohsumi N, Tanaka Y, Baba Y: Data Science, Classification, and Related Methods. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996. Springer Japan, 1998.
9. Caruana R, Niculescu-Mizil A: An empirical comparison of supervised learning algorithms. In *ACM Press*; 2006; 161–168. doi: 10.1145/1143844.1143865
10. Rodríguez JD, Pérez A, Lozano JA: Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans Pattern Anal Mach Intell* 2010; 32: 569–575. doi: 10.1109/TPAMI.2009.187.
11. Rodríguez JD, Pérez A, Lozano JA: Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE Trans Pattern Anal Mach Intell*. 2010 Mar;32(3):569-75. doi: 10.1109/TPAMI.2009.187. PMID: 20075479.
12. Topcuoglu OM, Uzunoglu B, Orhan T, Basaran EB, Gormez A, Sarica O. A real-world comparison of the diagnostic performances of six different TI-RADS guidelines, including ACR-/Kwak-/K-/EU-/ATA-/C-TIRADS. *Clin Imaging* 2025; 117: 110366. doi: 10.1016/j.clinimag.2024.110366.
13. Durante C, Hegedüs L, Na DG, Papini E, Sipos JA, Baek JH *et al.*: International Expert Consensus on US Lexicon for Thyroid Nodules. *Radiology* 2023; 309: e231481. doi: 10.1148/radiol.231481.
14. Böhlend M, Tharun L, Scherr T, Mikut R, Hagenmeyer V, Thompson LDR *et al.*: Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis. *PLoS One* 2021; 16: e0257635. doi: 10.1371/journal.pone.0257635.
15. Hornik K, Meyer D, Karatzoglou A: Support vector machines in R. *Journal of Statistical Software* 2006; 15. doi: 10.18637/jss.v015.i09.
16. Hsu CW, Chang CC, Lin CJ: A Practical Guide to Support Vector Classification. [www.csie.ntu.edu.tw/~cjlin](http://www.csie.ntu.edu.tw/~cjlin) 8.8. 2003.
17. Bennett KP, Campbell C: Support Vector Machines: Hype or Hallelujah? *ACM SIGKDD Explorations Newsletter* 2000; 2: 1–13. doi: 10.1145/380995.380999.
18. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H *et al.*: Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit Health* 2021; 3: e250–e259. doi: 10.1016/S2589-7500(21)00041-8. Erratum in: *Lancet Digit Health* 2021; 3: e413.
19. Xi NM, Wang L, Yang C: Improving the diagnosis of thyroid cancer by machine learning and clinical data. *Sci Rep* 2022; 12: 11143. doi: 10.1038/s41598-022-15342-z. Erratum in: *Sci Rep* 2022; 12: 13252. doi: 10.1038/s41598-022-17659-1.