European Archives of Oto-Rhino-Laryngology (2024) 281:2095–2104
https://doi.org/10.1007/s00405-023-08299-w

HEAD AND NECK

Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study

Shiva Borzooei[1] · Giovanni Briganti[2,3] · Mitra Golparian[4] · Jerome R. Lechien[5] · Aidin Tarokhian[4]

Received: 21 August 2023 / Accepted: 17 October 2023 / Published online: 30 October 2023

Abstract

Purpose The objective of this study was to train machine learning models for predicting the likelihood of recurrence in patients diagnosed with well-differentiated thyroid cancer. While thyroid cancer mortality remains low, the risk of recurrence is a significant concern. Identifying individual patient recurrence risk is crucial for guiding subsequent management and follow-ups.

Methods In this prospective study, a cohort of 383 patients was observed for a minimum duration of 10 years within a 15-year timeframe. Thirteen clinicopathologic features were assessed to predict recurrence potential. Classic (K-nearest neighbors, support vector machines (SVM), tree-based models) and artificial neural networks (ANN) were trained on three distinct combinations of features: a data set with all features excluding American Thyroid Association (ATA) risk score (12 features), another with ATA risk alone, and a third with all features combined (13 features). 283 patients were allocated for the training process, and 100 patients were reserved for the validation of stage.

Results The patients' mean age was 40.87 ± 15.13 years, with a majority being female (81%). When using the full data set for training, the models showed the following sensitivity, specificity and AUC, respectively: SVM (99.33%, 97.14%, 99.71), K-nearest neighbors (83%, 97.14%, 98.44), Decision Tree (87%, 100%, 99.35), Random Forest (99.66%, 94.28%, 99.38), ANN (96.6%, 95.71%, 99.64). Eliminating ATA risk data increased models specificity but decreased sensitivity. Conversely, training exclusively on ATA risk data had the opposite effect.

Conclusions Machine learning models, including classical and neural networks, efficiently stratify the risk of recurrence in patients with well-differentiated thyroid cancer. This can aid in tailoring treatment intensity and determining appropriate follow-up intervals.

Keywords Machine learning · Artificial intelligence · Thyroid cancer · Recurrence

Introduction

The incidence of thyroid cancers has steadily risen in recent years, primarily due to advancements in diagnostic methods [1, 2]. Although the mortality rate of differentiated thyroid cancers remains low, recurrence may not be uncommon [3]. Therefore, developing a model that enables healthcare practitioners to accurately predict an individual patient's risk of recurrence is crucial [4]. This knowledge provides a valuable opportunity to implement personalized treatment approaches and modify follow-up regimens [5]. Risk stratification plays a particularly significant role in determining the initiation of radioiodine therapy [6].

The prognostic guideline established by the American Thyroid Association (ATA) is currently the most extensively accepted system for estimating the risk of recurrence in

✉ Aidin Tarokhian
tarokhianaidin@gmail.com

1 Department of Endocrinology, Faculty of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran

2 Chair of AI and Digital Medicine, Faculty of Medicine, University of Mons, Mons, France

3 Department of Clinical Sciences, Faculty of Medicine, Université de Liège, Liège, Belgium

4 Hamadan University of Medical Sciences, Pajoohesh Blvd., Hamadan, Iran

5 Department of Otolaryngology-Head Neck Surgery, Elsan Hospital, Paris, France

differentiated thyroid cancers. These guidelines integrate a comprehensive array of clinicopathologic features to stratify patients into low, intermediate, and high-risk categories, providing a spectrum of qualitative risk assessment [7].

Applying machine learning models, particularly deep learning models, to medical data has garnered significant interest in recent years. The ability of non-parametric models to effectively capture and interpret non-linear data has contributed to improved predictive capabilities. The technology can be applied across a wide range of clinical applications, such as classifying patients into distinct risk groups for an event, such as thyroid cancer recurrence [8].

Machine learning models have already demonstrated significant potential in the field of thyroid disease. The classification of thyroid nodules into benign and malignant categories, achieved through convolutional neural networks trained on thousands of ultrasound images, has yielded state-of-the-art accuracy [9, 10]. A study demonstrated that by applying a multivariate model, researchers successfully predicted the presence of BRAF V600E-positive thyroid carcinoma, thus offering valuable prognostic insights [11]. However, as of now, no study has been undertaken to ascertain the capabilities of artificial intelligence in predicting the recurrence of thyroid cancers.

This study aimed to develop machine learning models that can accurately estimate the likelihood of recurrence in well-differentiated thyroid carcinomas. Furthermore, we assessed patients' risk of recurrence using models exclusively trained on the American Thyroid Association (ATA) prognostic system. Finally, we integrated the ATA risk assessment with other clinicopathologic factors into a unified model.

Materials and methods

Ethical considerations

According to the institutional review board, data utilized in this study underwent a rigorous deidentification process before their usage. The study itself was conducted in strict adherence to the local ethical guidelines and the principles outlined in the Declaration of Helsinki. Furthermore, a dedicated board of experts at Hamedan University of Medical Sciences reviewed and approved the research protocol. No approvals were necessary from the International Review Board (IRB).

Setting and subjects

This study involved a retrospective cohort of 383 patients who received histopathological diagnoses of thyroid cancers from a single medical center. Among these patients, 283 individuals were randomly chosen for the training process, while the remaining 100 patients were allocated to validate the trained models. Only patients with differentiated thyroid cancers, including papillary, micropapillary, follicular, and Hürthle cell carcinoma, were included in the study [12]. A minimum follow-up period of 10 years was maintained for all patients, starting from the time of surgery and initial diagnosis. The study spanned a period of 15 years.

Clinicopathological features

The collected patient data included the following variables: age at diagnosis in years, biological sex (Female and Male), current smoking status, past smoking history, history of radiation therapy to the head and neck region, thyroid function (classified as euthyroid, clinical or subclinical hypo/hyperthyroidism), presence of goiter (diffuse, single nodular goiter on the left or right lobe, multinodular, or normal), presence of adenopathy on physical examination (no adenopathy, anterior right, anterior left, bilateral, posterior, or extensive involving all the aforementioned locations), pathological subtype of cancer (papillary, micropapillary, follicular, Hürthle cell), focality (unifocal, multifocal), risk assessment according to ATA guidelines (low, intermediate, high), TNM staging (individual T, N, and M scores, and final stage), initial treatment response (excellent, biochemical incomplete, structurally incomplete, indeterminate), and recurrence status (including both locoregional and distant metastasis).

Statistics

- *Software and Tools* The data analysis in this study was conducted utilizing the scikit-learn library (version 1.3.0) for developing classical machine learning models and data preprocessing [13]. Deep learning models were constructed using PyTorch (version 2.0) [14]. The visualization of graphs and figures was accomplished using Matplotlib (version 3.7.2) and Seaborn (version 0.12.2) [15].
- *Training* In the study, a total of 283 patients were sampled randomly for the training phase. One hot encoding was employed for handling categorical data [16]. All data were standardized to optimize the performance of models that rely on spatial distances using a standard scaler. The classic models employed in this study encompassed k-nearest neighbors, tree-based models, and support vector machines (SVM).

These three models were selected due to their representation of distinct methodological approaches. Support Vector Machine (SVM) is a mathematical model that identifies the hyperplane that optimizes the margin between distinct

classes within the feature space. Mathematically, SVM is rooted in convex optimization, enabling it to discern the most discriminative features within the data set. Notably, SVM exhibits proficiency, particularly when the imperative is identifying pivotal features. Furthermore, it effectively manages modest-to-moderate-sized data sets [17].

In contrast, decision trees and Random Forests are robust tools renowned in classification. Their methodology involves the recursive partitioning of the data set into subsets based on the most salient features. From a mathematical standpoint, these models employ criteria such as Gini impurity or information gain to ascertain the ideal feature for each division. This approach proves advantageous in revealing feature importance and efficiently delineating intricate decision boundaries [18].

K-Nearest Neighbors (KNN), a non-parametric algorithm, is founded on distance metrics for data point classification. Its flexibility allows for adaptability to diverse data distributions, adhering to the mathematical principle that data points sharing similar feature vectors are likely to share the same class. This inherent adaptability positions KNN as a fitting choice for predicting thyroid carcinoma recurrence, as it effectively captures the intricate non-linear relationships within the data [19].

All models underwent cross-validation and grid search techniques to identify the optimal hyperparameters. Specifically, a cross-validation approach with a fold size of 2 was utilized. The best-performing model, along with its corresponding hyperparameters, was reported. Deep neural networks were developed, incorporating multiple layers and various adjustments. The architecture that yielded the best results was also documented.

This study explored a range of neural network architectures, encompassing hidden layers ranging from 1 to 5, with the number of neurons per layer varying between 2 and 128. The training process employed static learning rates from 0.0001 to 0.1, and the training duration spanned epochs from 10 to 1000. Activation functions such as ReLU, LeakyReLU, and SeLU were incorporated into the network designs. Furthermore, stochastic gradient descent and the Adam optimizer were assessed for suitability in training the models.

- *Data sets* Three sub-data sets were used to train each model. The first data set comprised 12 clinical and demographic patient attributes, except the ATA risk score. The second data set, on the other hand, exclusively contained a single feature related to the ATA risk group. The third data set combined the first and second data sets (13 features). Classical models were individually trained on each data set. However, the Artificial Neural Network (ANN) was exclusively trained on the third data set, encompassing the complete set of attributes for learning.
- *Internal validation* The models underwent internal validation using 100 additional patients from the same cohort. For each model, various performance metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and area under the curve (AUC) were calculated and reported.
- *External Validation* It is important to note that external validation of the models could not be conducted due to the unavailability of comparable data from other centers. Nevertheless, requests by third parties for the model weights and data for external validation of the study findings are encouraged and welcomed.

Results

Descriptive analysis

The data of 283 patients were considered. The mean age of patients in the training set was 40.26 ± 15.17 years, with 82% being female. Most individuals had no history of smoking or radiation exposure to the head and neck region. Euthyroidism was observed in 86% of the patients. Among the thyroid dysfunctions detected, clinical hyperthyroidism accounted for the highest prevalence at 6%, followed by subclinical hypothyroidism at 4%. The most common pathologic subtype was papillary thyroid cancer, accounting for 75% of cases, followed by micropapillary, follicular, and Hurthel cell carcinoma, each comprising 12%, 7%, and 6%, respectively. Sixty-five percent of patients had uni-focal involvement (see Fig. 1).

According to the ATA classification, patients were categorized as follows: low risk (66%), intermediate risk (27%), and high risk (7%). Notably, the majority of cases (88%) were confined to Stage I of the disease, with 8% at Stage II, 1% at Stage III, and 1% at Stage IVA, while 2% were at Stage IVB.

Over half of the patients (54%) exhibited an excellent treatment response, with the second most common response being a structurally incomplete response at 27%, followed by indeterminate and biochemical incomplete responses at 17% and 6%, respectively. In terms of recurrence, 78 patients (28%) experienced thyroid cancer recurrence, which included both local and distal recurrences.

The validation set patients exhibited characteristics closely similar to those of the training set (Table 1). This was not preplanned or stratified in any manner and was purely random. Eventually, all patients' characteristics from both the validation and training sets were combined and summarized in Table 1 for detailed reference.

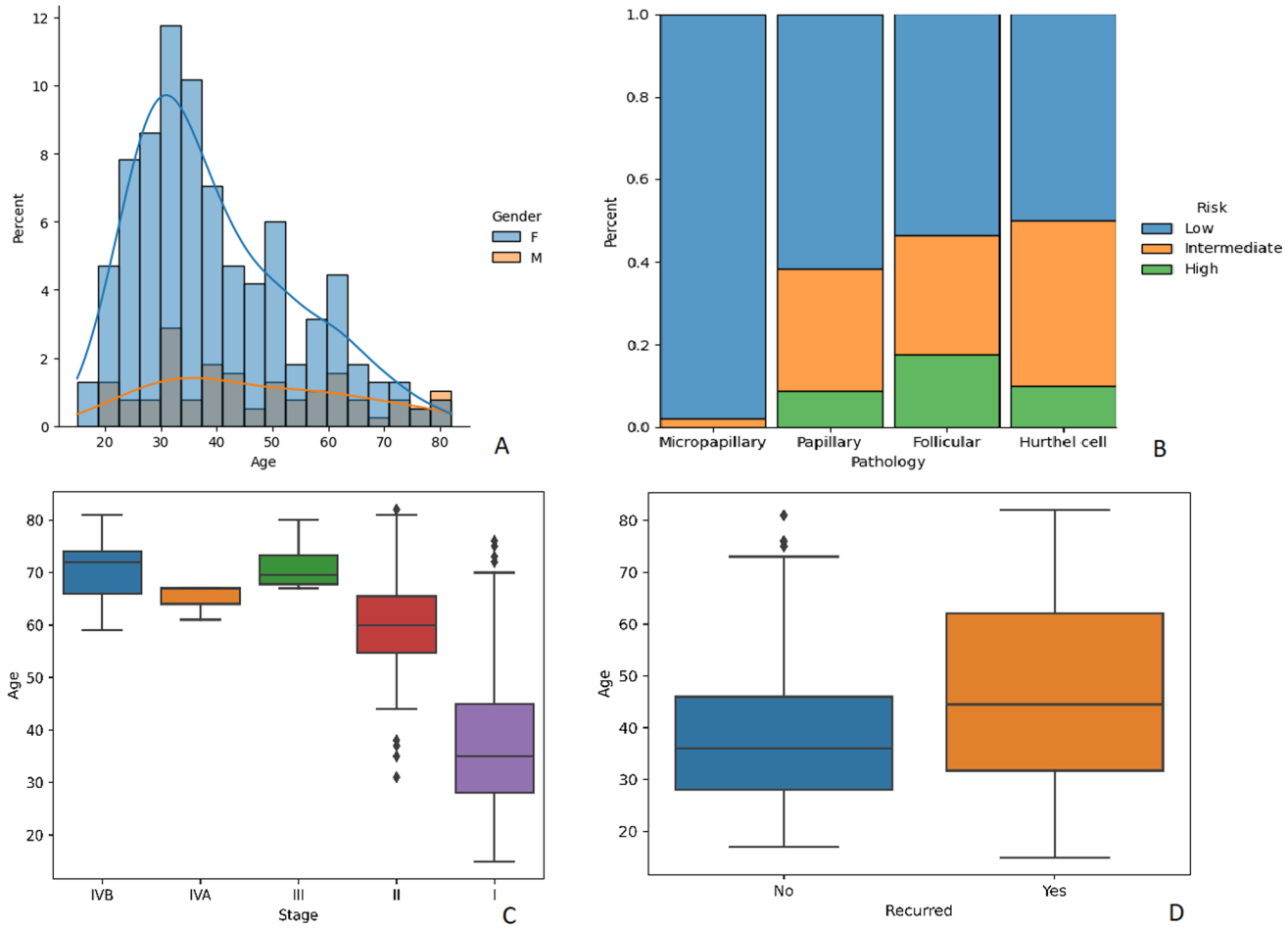

Fig. 1 **A** Patient age and sex distribution showed right skewness. **B** Comparison of pathological subtypes and ATA risk score revealed that patients with Follicular cancer were more likely to be categorized as high risk. No high-risk patients were found in the micropapillary group. **C**, **D** Patients with higher median age tended to have higher stages at diagnosis and recurrence events

Models' metrics and performance

After training the models, their performance (Table 2) was assessed using a validation data set comprising 100 patients. The individual models' results did not differ significantly from each other. Models trained on the ATA risk-only data set exhibited identical outcomes, with an AUC of 92.59%, accuracy of 89%, sensitivity of 93.33%, specificity of 87.14%, PPV of 75.67%, and NPV of 96.82%.

Training the models on the first data set, which excluded ATA risk, improved performance for most algorithms. The models showed notably better specificity and slightly lower sensitivity. In addition, accuracy and AUC also improved. Among these models, the SVM algorithm demonstrated the best performance, achieving 96% accuracy (with only four misclassified patients), an AUC of 99.23%, sensitivity of 90%, specificity of 98.57%, PPV of 96.42%, and NPV of 95.83%.

Furthermore, the models trained on all features, including ATA risk, outperformed the first two models. Again, SVM algorithms demonstrated the best performance with an AUC of 99.71%, sensitivity of 93.33%, specificity of 97.14%, PPV of 93.33%, and NPV of 97.14%. Accuracy was also high at 96%.

The artificial neural network (ANN) model was trained on the full-featured data set (40 input neurons), utilizing one hidden layer comprising 37 neurons with the ReLU activation function. In addition, a dropout rate with a probability of 0.5 was incorporated to reduce overfitting. The architecture was determined through cross-validation of various optimizers, learning rates, epochs, numbers of hidden layers and their neurons, dropout ratios, and activation functions. In addition, due to the limited size of the training data set for the ANN model, simplicity was prioritized as a vital feature to mitigate overfitting. The ANN results were better or comparable to all classic models except the SVM model.

Table 1 Patients demographic and clinicopathological characteristics

Item	Training patients (283)	Internal validation patients (100)	Total (383)
Age	40.26 ± 15.17 years	42.6 ± 14.97 years	40.87 ± 15.13 years
Gender			
Male	51 (18%)	20 (20%)	71 (19%)
Female	232 (82%)	80 (80%)	312 (81%)
Currently smoking			
Yes	35 (12%)	14 (14%)	49 (13%)
No	248 (88%)	86 (86%)	334 (87%)
History of smoking			
Yes	20 (7%)	8 (8%)	28 (7%)
No	263 (93%)	92 (92%)	355 (93%)
History of radiation exposure			
Yes	5 (2%)	2 (2%)	7 (2%)
No	278 (98%)	98 (98%)	376 (98%)
Thyroid function			
Euthyroid	244 (86%)	88 (88%)	332 (87%)
Subclinical hypothyroidism	12 (4%)	2 (2%)	14 (4%)
Clinical hypothyroidism	8 (3%)	4 (4%)	12 (3%)
Subclinical hyperthyroidism	3 (1%)	2 (2%)	5 (1%)
Clinical hyperthyroidism	16 (6%)	4 (4%)	20 (5%)
Pathology			
Papillary	213 (75%)	74 (74%)	287 (75%)
Micropapillary	34 (12%)	14 (14%)	48 (13%)
Follicular	20 (7%)	8 (8%)	28 (7%)
Hurthel cell	16 (6%)	4 (4%)	20 (5%)
Focality			
Uni-focal	184 (65%)	63 (63%)	247 (64%)
Multi-focal	99 (35%)	37 (37%)	136 (36%)
ATA Risk			
Low	186 (66%)	63 (63%)	249 (65%)
Intermediate	76 (27%)	26 (26%)	102 (27%)
High	21 (7%)	11 (11%)	32 (8%)
Adenopathy			
None	208 (73%)	69 (69%)	277 (72%)
Right-sided	32 (11%)	16 (16%)	48 (13%)
Left-sided	11 (4%)	6 (6%)	17 (4%)
Bilateral	28 (10%)	4 (4%)	32 (8%)
Posterior	2 (1%)	0 (0%)	2 (1%)
Extensive	2 (1%)	5 (5%)	7 (2%)
Goiter			
None	6 (2.5%)	1 (1%)	7 (1.5%)
Single nodular right	97 (34%)	43 (43%)	140 (37%)
Single nodular left	71 (25%)	18 (18%)	89 (23%)
Multi nodular	103 (36%)	37 (37%)	140 (37%)
Diffuse	6 (2.5%)	1 (1%)	7 (1.5%)
Tumor			
T1a	34 (12%)	15 (15%)	49 (13%)
T1b	33 (12%)	10 (10%)	43 (11%)
T2	112 (40%)	39 (39%)	152 (40%)
T3a	74 (25%)	22 (22%)	96 (25%)
T3b	11 (4%)	5 (5%)	16 (4%)

Table 1 (continued)

Item	Training patients (283)	Internal validation patients (100)	Total (383)
T4a	13 (5%)	7 (7%)	20 (5%)
T4b	6 (2%)	2 (2%)	8 (2%)
Node			
N0	200 (70%)	68 (68%)	268 (70%)
N1a	17 (6%)	5 (5%)	22 (6%)
N1b	66 (24%)	27 (27%)	93 (24%)
Metastasis			
M0	272 (96%)	93 (93%)	365 (95%)
M1	11 (4%)	7 (7%)	18 (5%)
Stage			
I	249 (88%)	84 (84%)	333 (87%)
II	23 (8%)	9 (9%)	32 (8%)
III	3 (1%)	1 (1%)	4 (1%)
IVa	2 (1%)	1 (1%)	3 (1%)
IVb	6 (2%)	5 (5%)	11 (3%)
Treatment response			
Excellent	153 (54%)	55 (55%)	208 (54%)
Indeterminate	49 (17%)	12 (12%)	61 (16%)
Biochemical incomplete	16 (6%)	7 (7%)	23 (6%)
Structural incomplete	65 (23%)	26 (26%)	91 (24%)
Recurrence			
Yes	78 (28%)	30 (30%)	108 (28%)
No	205 (72%)	70 (70%)	275 (72%)

T (Tumor), N (Node), M (Metastasis)

For a comprehensive overview of the different model metrics and their respective training hyperparameters, please refer to Table 2.

Feature selection

Decision tree models were used to identify and visualize the most important differentiating factors in thyroid cancer recurrence. The assessment of feature importance was conducted using the Gini impurity criteria, facilitating the calculation of feature significance. Importance scores were assigned to the most influential features. The sum of all scores is equal to one (see Fig. 2).

Upon examining the entire data set, it was discerned that several predictors held notable significance in forecasting thyroid cancer recurrence. These included structurally incomplete treatment response (score = 0.843), gender (score = 0.014), low-risk categorization (score = 0.054), age (0.072), Hurthel cell pathology (score = 0.013), and excellent treatment response (score = 0.004), as illustrated in Fig. 2. Notably, no constraints were imposed on the number of nodes or the maximum depth of partitioning, allowing the decision tree models to adapt and capture the patterns.

Discussion

This study demonstrates the promising potential of machine learning algorithms in predicting recurrence risk and stratifying patients with differentiated thyroid cancers. Compared to classical risk stratification using ATA risk categories, machine learning models can not only predict recurrence qualitatively but also quantitatively. They offer the advantage of expressing the chance of patient relapse in percentages, unlike the categorical nature of ATA risk categories [20].

Furthermore, machine learning can be applied to the existing rule-based criteria, such as ATA risk categories, to improve their interpretability. By doing so, qualitative and semi-quantitative systems can be transformed into more accurate and objective models, ultimately enhancing risk assessment and patient outcomes.

This study reaffirms the accuracy of the ATA risk model as a reliable tool for detecting thyroid cancer recurrence [21]. Its sensitivity, utilizing only a single categorical feature categorized as high, intermediate, and low, is surprisingly high (93.33%). However, it can result in overdiagnosis and a high false positive rate when used in isolation.

Incorporating various demographic and clinicopathologic features, particularly the initial treatment response

Table 2 Detailed models' metrics and training hyperparameters

Model Name	Model Subtype	Best Hyperparameters	Metrics					
			Sensitivity	Specificity	PPV	NPV	AUC	Accuracy
K-Nearest Neighbors	ATA risk excluded	Neighbors: 4 Weights: distance	80%	98.57%	96%	92%	0.9747	93%
	ATA risk only	Neighbors: 5 Weights: distance	93.33%	87.14%	75.67%	96.82%	0.9259	89%
	Full features	Neighbors: 6 Weights: distance	83%	97.14%	92.59%	93.15%	0.9842	93%
Support Vector Machine	ATA risk excluded	C: 0.01 Class_Weights: balanced Kernel: Linear Gamma: Scale	90%	98.57%	96.42	95.83	0.9923	96%
	ATA risk only	C: 0.01 Class_Weights: balanced Kernel: Linear Gamma: Scale	93.33%	87.14%	75.67%	96.82%	0.9259	89%
	Full features	C: 1 Class_Weights: None Kernel: RFB Gamma: Scale	93.33%	97.14%	93.33%	97.14%	0.9971	96%
Decision-Tree	ATA risk excluded	Criteria: gini Class_Weights: None Max_depth: 3 Max_features: None	90%	97.14%	93%	95.77%	0.9704	95%
	ATA risk only	Criteria: gini Class_Weights: balanced Max_depth: 3 Max_features: auto	93.33%	87.14%	76%	96.82%	0.9259	89%
	Full features	Criteria: gini Class_Weights: None Max_depth: 3 Max_features: None	87%	100%	100%	94.59%	0.9935	96%
Model Name	Model Subtype	Best Hyperparameters	Metrics					
			Sensitivity	Specificity	PPV	NPV	AUC	Accuracy
Random Forest	ATA risk excluded	class_weight: balanced, criteria: entropy, max_depth: 4, max_features: log2, n_estimators: 80	93.33%	97.14%	93.33%	97.14%	0.9876	96%
	ATA risk only	class_weight: balanced, criteria: gini, max_depth: 2, max_features: auto, n_estimators: 60	93.33%	87.14%	75.67%	96.82%	0.9259	89%
	Full features	class_weight: balanced, criteria: gini, max_depth: 4, max_features: sqrt, n_estimators: 60	96.66%	94.28%	87.87%	98.50%	0.9938	95%
Artificial Neural Network	Full Features	Input layer: 40 Hidden layer: 1 layer 37 neurons, ReLu activation Drop out: 0.5 probability Output layer: 1 neuron Optimizer: Adam Learning rate: 0.001 Epoch: 50	96.6%	95.71%	90.6%	98.52%	0.9964	96%

PPV positive predictive value, *NPV* negative predictive value, *AUC* area under curve

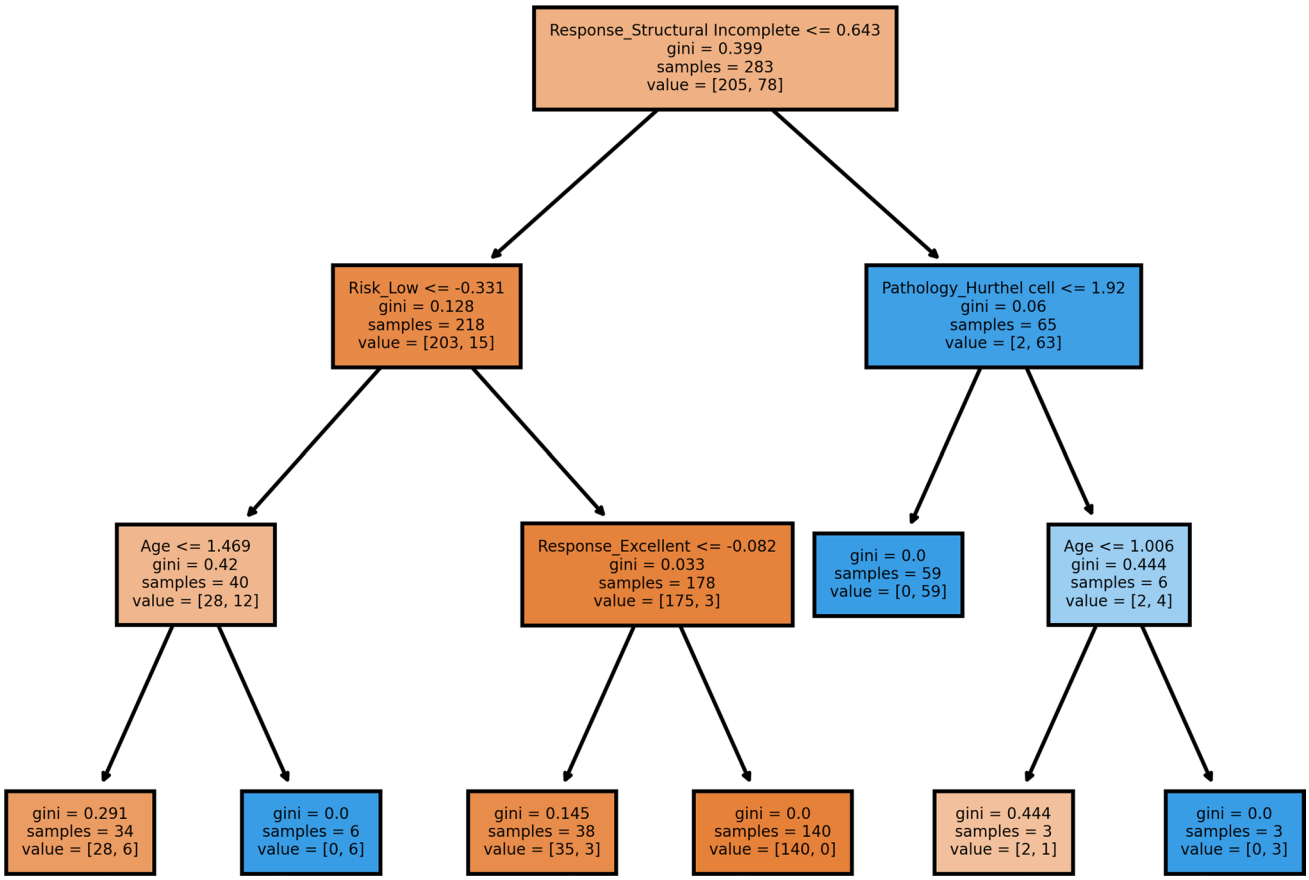

Fig. 2 Most relevant predicting features for the recurrence in the full data set

of patients, enhances prediction specificity but may also increase false negative results. However, combining these features with ATA risk balanced sensitivity and specificity. This suggests that future ATA risk system modifications could improve its accuracy.

The comparable or even superior performance of classical algorithms compared to artificial neural networks highlights the crucial role of data in predictive modeling [22]. The quality and representativeness of the data are of utmost importance. The relatively accurate models in this study can be attributed to incorporating diverse clinical, surgical, pathological, and demographic features of patients.

The fact that the SVM model outperformed k-nearest neighbors (KNN), artificial neural networks (ANN), and tree-based models in predicting thyroid cancer recurrence can be attributed to several mathematical and algorithmic factors [17, 23]:

1. Feature Selection and Dimensionality Reduction: SVMs are effective at feature selection and dimensionality reduction. KNN and tree-based models, on the other hand, do not inherently offer the same level of feature selection.
2. Robustness to Noise: SVMs are generally robust to noise and outliers in the data. In medical data sets, it is common to have noisy or incomplete data, and SVMs can handle these situations better than KNN, ANN, or decision trees, which can be sensitive to noise.
3. Optimized Margin Separation: SVMs aim to maximize the margin between different classes. This can lead to better generalization performance on unseen data. KNN might be sensitive to the specific distribution of data points, and tree-based models can overfit the training data, while SVMs focus on creating a more robust decision boundary.
4. Balanced Trade-off with Regularization Parameter (C): SVMs have a regularization parameter (C) that balances the margin width and the misclassification error. The choice of this parameter can be fine-tuned mathematically to find the right trade-off between overfitting and underfitting. This is a critical mathematical decision that influences model performance.

It is important to note that many factors, including data preprocessing, feature engineering, and the specific characteristics of the data set, can influence the performance

of machine learning models. Therefore, the superior performance of the SVM in this case may be attributed to the interplay of these factors, not just mathematical properties alone.

Currently, there has not been a comparable study on predicting thyroid cancer recurrence risk using machine learning. Nonetheless, we can gauge the model's performance in our current study by comparing it to models developed for predicting recurrence risk in more common cancers. Machine learning models have proven effective in predicting colon and breast cancer recurrence with AUC scores of 0.815 and 0.94, respectively [24, 25]. Our model demonstrates comparable, if not superior, results (AUC = 0.99).

This study has a few limitations worth noting. First, the models developed here could not be externally validated due to the lack of comparable data from other institutions. In addition, the internal validation data set closely resembles the training data set, which could lead to potential model overfitting. To address this problem, considering real-world patient characteristics from previously published studies on thyroid cancer could be beneficial. For example, a prior study reported a female-to-male ratio of 3.36 [26]. Data from SEER (The Surveillance, Epidemiology, and End Results) database showed that papillary thyroid cancer is the most common pathologic subtype (90%), and the majority of patients are diagnosed at I and II stages (94%) [27]. Another study involved 1093 patients with an average age of 49, 78% female, and 88% papillary thyroid cancer. Riskwise, 67.5% were low ATA risk, 29.8% intermediate risk, and 2.7% high risk [28]. In a Canadian study, 53% of patients fell into the low-risk category, 27% intermediate risk, and 20% high risk, with favorable treatment responses for most [29]. This suggests that the patient population in our study (both training and internal validation data sets) closely aligns with characteristics seen in studies worldwide (Table 1). This may help alleviate concerns about overfitting and external validation. Nonetheless, the applicability of these models beyond our specific data set remains uncertain.

Conducting tests on the model, which has been developed through this study, with a broader and more diverse population (e.g., African–Americans, Hispanics, non-Asians) will provide a clearer picture of its ability to perform effectively in real-world situations. This expanded testing will help assess whether the model's predictive accuracy extends beyond the specific data set it was initially trained on, making it a valuable tool for a broader range of clinical and medical applications.

It should be noted that the ATA (American Thyroid Association) model utilized in this study relied on three predefined risk categories: low, intermediate, and high risk. However, individual clinicopathologic data typically incorporated into the ATA classification system were not included in the original data set used for model development. This may have implications for the accuracy and applicability of the ATA model's predictions.

Fewer cases with specific pathological subtypes (Follicular and Hurthel cell) and mostly early stage diagnoses (Stage I and II) might affect the model's real-world utility. However, the study might have tackled these biases due to roughly equal representation in the training and validation patient sets (Table 1).

These limitations highlight the need for caution when interpreting and applying the findings of this study. Future research should aim to address these limitations by including a comparable external data set to enhance the generalizability of the models.

Conclusion

This study demonstrates the usefulness of both classic machine learning models and neural networks in predicting the recurrence risk and stratifying patients with differentiated thyroid cancers. In addition, it highlights the capacity of machine learning methods to enhance existing rule-based clinical tools, such as the ATA risk prediction model, by rendering them more accurate and interpretable.

Acknowledgements We sincerely thank Professor Peter Szolovits for his valuable comments on the presenting article. The article was submitted with the ethical identifier IR.UMSHA.REC.1402.360 at Hamadan University of Medical Sciences, Hamadan, Iran.

Author contributions All authors have made significant contributions to this work. SB, MG, and AT collected the initial and follow-up information of the patients and participated in data cleaning and preprocessing. AT and MG contributed to the development and coding of the machine learning models. GB and GRL reviewed the data set and article for potential errors and biases regarding machine learning principles. In addition, all authors have been involved in the writing of the article.

Data availability The data sets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Declarations

Conflict of interest All authors declare no financial or non-financial competing interests. No funding was received for conducting this study.

Ethics declaration The author Jerome R. Lechien is also guest editor of the special issue on 'ChatGPT and Artifcial Intelligence in Otolaryngology-Head and Neck Surgery'. He was not involved with the peer review process of this article.

References

1. Powers AE, Marcadis AR, Lee M, Morris LGT, Marti JL (2019) Changes in trends in thyroid cancer incidence in the United States, 1992 to 2016. JAMA 322(24):2440–2441. https://doi.org/10.1001/jama.2019.18528
2. Aschebrook-Kilfoy B, Kaplan EL, Chiu BC-H, Angelos P, Grogan RH (2013) The acceleration in papillary thyroid cancer incidence rates is similar among racial and ethnic groups in the United States. Ann Surg Oncol 20:2746–2753
3. Li M, Brito JP, Vaccarella S (2020) Long-term declines of thyroid cancer mortality: an international age–period–cohort analysis. Thyroid 30(6):838–846
4. Shaha AR (2012) Recurrent differentiated thyroid cancer. Endocr Pract 18(4):600–603
5. Tuttle RM, Alzahrani AS (2019) Risk stratification in differentiated thyroid cancer: from detection to final follow-up. J Clin Endocrinol Metab 104(9):4087–4100
6. Luster M, Clarke S, Dietlein M, Lassmann M, Lind P, Oyen W et al (2008) Guidelines for radioiodine therapy of differentiated thyroid cancer. Eur J Nucl Med Mol Imaging 35:1941–1959
7. Lee J, Lee SG, Kim K, Yim SH, Ryu H, Lee CR et al (2019) Clinical value of lymph node ratio integration with the 8th edition of the UICC TNM classification and 2015 ATA risk stratification systems for recurrence prediction in papillary thyroid cancer. Sci Rep 9(1):13361
8. Ouyang F-s, Guo B-l, Ouyang L-z, Liu Z-w, Lin S-j, Meng W et al (2019) Comparison between linear and nonlinear machine-learning algorithms for the classification of thyroid nodules. Eur J Radiol 113:251–257
9. Li L-R, Du B, Liu H-Q, Chen C (2021) Artificial intelligence for personalized medicine in thyroid cancer: current status and future perspectives. Front Oncol 10:604051
10. Verburg F, Reiners C (2019) Sonographic diagnosis of thyroid cancer with support of AI. Nat Rev Endocrinol 15(6):319–321
11. Yoon J, Lee E, Koo JS, Yoon JH, Nam K-H, Lee J et al (2020) Artificial intelligence to predict the BRAFV600E mutation in patients with thyroid cancer. PLoS ONE 15(11):e0242806
12. Schlumberger M, Leboulleux S (2021) Current practice in patients with differentiated thyroid cancer. Nat Rev Endocrinol 17(3):176–188
13. Bisong E, Bisong E (2019) Introduction to Scikit-learn. Building machine learning and deep learning models on Google cloud platform: a comprehensive guide for beginners, 1st edn. Apress, Ottawa, pp 215–29
14. Imambi S, Prakash KB, Kanagachidambaresan G (2021) PyTorch. Programming with TensorFlow: solution for edge computing applications, Springer, Cham, pp 87–104
15. Bisong E, Bisong E (2019) Matplotlib and Seaborn. Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners, 1st edn. Apress, Ottawa, pp 151-167
16. Yu L, Zhou R, Chen R, Lai KK (2022) Missing data preprocessing in credit classification: One-hot encoding or imputation? Emerg Mark Financ Trade 58(2):472–482
17. Yue S, Li P, Hao P (2003) SVM classification: Its contents and challenges. Applied Mathematics-A Journal of Chinese Universities 18:332–342
18. Clark LA, Pregibon D (2017) Tree-based models. In: Statistical models in S. Routledge, pp 377–419.
19. Taunk K, De S, Verma S, Swetapadma A (2019) A brief review of nearest neighbor algorithm for learning and classification. In: 2019 International Conference on intelligent computing and control systems (ICCS): IEEE; pp 1255–60
20. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. N Engl J Med 380(14):1347–1358
21. Tuttle RM, Tala H, Shah J, Leboeuf R, Ghossein R, Gonen M et al (2010) Estimating risk of recurrence in differentiated thyroid cancer after total thyroidectomy and radioactive iodine remnant ablation: using response to therapy variables to modify the initial risk estimates predicted by the new American Thyroid Association staging system. Thyroid 20(12):1341–1349
22. Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12
23. Anguita D, Ghio A, Greco N, Oneto L, Ridella S (2010) Model selection for support vector machines: advantages and disadvantages of the machine learning theory. In: The 2010 International Joint Conference on neural networks (IJCNN): IEEE, pp 1–8
24. El Haji H, Souadka A, Patel BN, Sbihi N, Ramasamy G, Patel BK et al (2023) Evolution of breast cancer recurrence risk prediction: a systematic review of statistical and machine learning-based models. JCO Clin Cancer Inform 7:e2300049
25. Mazaki J, Katsumata K, Ohno Y, Udo R, Tago T, Kasahara K et al (2021) A novel prediction model for colon cancer recurrence using auto-artificial intelligence. Anticancer Res 41(9):4629–4636
26. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM (2017) Trends in thyroid cancer incidence and mortality in the United States, 1974–2013. JAMA 317(13):1338–1348. https://doi.org/10.1001/jama.2017.2719
27. Seib CD, Sosa JA (2019) Evolving understanding of the epidemiology of thyroid cancer. Endocrinol Metab Clin N Am 48(1):23–35. https://doi.org/10.1016/j.ecl.2018.10.002
28. Kelly A, Barres B, Kwiatkowski F, Batisse-Lignier M, Aubert B, Valla C et al (2019) Age, thyroglobulin levels and ATA risk stratification predict 10-year survival rate of differentiated thyroid cancer patients. PLoS ONE 14(8):e0221298. https://doi.org/10.1371/journal.pone.0221298
29. Wu J, Hu XY, Ghaznavi S, Kinnear S, Symonds CJ, Grundy P et al (2022) The prospective implementation of the 2015 ATA guidelines and modified ATA recurrence risk stratification system for treatment of differentiated thyroid cancer in a canadian tertiary care referral setting. Thyroid 32(12):1509–1518. https://doi.org/10.1089/thy.2022.0055