applied sciences

MDPI

Article

Machine Learning in Differentiated Thyroid Cancer Recurrence and Risk Prediction

Matthew A. Penner [1,2,3], Derek Berger [1], Xuchen Guo [1] and Jacob Levman [1,4,*]

[1] Department of Computer Science, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada; matthew.penner@mail.utoronto.ca (M.A.P.); dberger@stfx.ca (D.B.); x2023diy@stfx.ca (X.G.)
[2] Department of Physics, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada
[3] Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada
[4] Nova Scotia Health Authority, Halifax, NS B3H 1V8, Canada
* Correspondence: jlevman@stfx.ca

Abstract

Differentiated thyroid cancer (DTC) poses significant management challenges due to the variable risk of recurrence. This study uses a dataset comprising clinical, pathological, and treatment data from 383 patients to develop and validate machine learning models, combined with feature selection algorithms, for predicting differentiated thyroid cancer recurrence. We evaluated models based on a variety of machine learning technologies (light gradient boosting machine, random forest, k-nearest neighbor, logistic regression, stochastic gradient descent, and an emerging deep learner optimized for tabular data: Gandalf) combined with several feature selection methods. Our feature selection technologies include an emerging redundancy-aware wrapper-based feature selection technique, achieving thyroid cancer recurrence prediction accuracy of 94.8 to 95.9% across two validation methods, based only on whether the patient's tumor's response was structurally incomplete, whether their tumor's stage was advanced (III, IVA, or IVB), and the patient's age. The results underline the potential for machine learning to enhance the precision of recurrence prediction in DTC while developing technologies whose predictive capacity is more easily explained. Using the same dataset, machine learning and feature selection techniques, this study also provides an analysis on predicting American Thyroid Association (ATA) risk scores. The technologies developed as part of this study have potential for improving the personalization of healthcare through the creation of models based on detailed patient-specific clinical attributes.

Keywords: machine learning; differentiated thyroid cancer; cancer recurrence; thyroid cancer; predictive modeling; risk prediction; artificial intelligence

Academic Editor: Giorgio De Nunzio

Received: 8 May 2025
Revised: 30 July 2025
Accepted: 23 August 2025
Published: 27 August 2025

Citation: Penner, M.A.; Berger, D.; Guo, X.; Levman, J. Machine Learning in Differentiated Thyroid Cancer Recurrence and Risk Prediction. *Appl. Sci.* **2025**, *15*, 9397. https://doi.org/10.3390/app15179397

1. Introduction

1.1. Overview

Thyroid cancer is the fifth most common cancer in women in 2015, and there are approximately 62,000 new cases a year for men with the most common type being differentiated thyroid cancer (DTC) [1]. The majority of thyroid cancers are well differentiated, having been derived from thyroid follicular cells [2]. Well-differentiated thyroid cancers occur in papillary, follicular, and oncocytic forms, representing 84%, 4%, and 2% of all thyroid cancers, respectively [2]. These cancers are usually curable when found early and in younger patients [2]. However, some patients may experience recurrence or metastasis

(meaning it comes back after treatment, possibly having spread to another tissue), which can negatively affect patient survival and quality of life [3]. Technologies that predict recurrence in patients with well-differentiated thyroid cancer can be useful prognostic tools, which can help identify patients who would most benefit from rigorous surveillance and monitoring, and potentially more aggressive treatment strategies.

1.2. Clinical Management

The current method for treating thyroid cancer is surgery, radioactive iodine therapy, and thyroid-stimulating hormone (TSH) suppression therapy [4]. The extent of surgery, and the need for radioactive iodine therapy, depend on the size of the tumor, its location, how aggressively it grows, etc. Surveillance includes serum thyroglobulin (Tg) measurements, TSH-stimulated Tg test, ultrasound, and more [5,6]. The amount of surveillance relied upon depends on many clinical factors [6].

The current method for prognosis uses the American Thyroid Association (ATA) Risk Stratification System, which is often referred to as the ATA risk score. The 2015 ATA guidelines help doctors decide when to perform biopsies on thyroid nodules by classifying them into risk categories. However, there is some concern that these guidelines might underestimate the risk for more serious nodules, and doctors often vary in their assessment of risk [7]. A recent study found that a microRNA-based risk score is more accurate than the traditional ATA score at predicting the progression of thyroid cancer, suggesting it could be a better tool for prognosis [8]. Research also suggests that the ATA's initial risk predictions can be significantly refined by monitoring how patients respond to their first two years of treatment, which helps tailor follow-up care more effectively [9].

1.3. Literature Review

Machine learning has been used in thyroid cancer diagnosis and management in applications such as detection, classification, prognosis, and treatment [10]. Machine learning models have also been used to predict the outcomes of thyroid cancer patients based on ultrasound-derived features [11]. AI models have been trained to predict the risk of recurrence in patients with well-differentiated thyroid cancer [3]. These models have shown promising improvements over traditional risk assessment methods [3]. AI has considerable potential to assist in thyroid cancer management, and AI technologies have helped with challenges like inaccuracy in predicting prognoses and intra-observer variability in ultrasound imaging. AI's role in improving diagnostic accuracy and moving toward personalized treatment is becoming more and more prevalent as outlined in the literature [12]. Specifically, AI's application in thyroid cancer diagnosis has improved sensitivity and specificity, providing new analytic approaches for indeterminate nodules [13]. AI has tremendous potential in improving our approach to personalized medicine, especially in differentiating follicular neoplasms and refining prognoses, where these technologies demonstrate potential to revolutionize thyroid cancer care [10]. AI can even match or outperform less experienced doctors in evaluating thyroid nodules through ultrasonographic diagnosis, making the resultant diagnoses less subjective [14]. AI not only supports radiologists in delivering more accurate diagnoses but can also improve diagnostic efficiency and reduce clinician workload [15]. Considerable research has focused on applications of AI on thyroid cancer recurrence [16–29], applying a variety of machine learning and sometimes feature selection algorithms.

The leading publication addressing the dataset relied upon in this study focused on machine learning for the prediction of thyroid cancer recurrence, "Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study" [3]. Their analysis included a comparison of many machine learning algorithms, including k-nearest neighbor

(KNN), support vector machine (SVM) tree-based models, and artificial neural networks (ANNs) with 283 samples used for training and 100 for validation. This study trained the learners three times targeting recurrence: once with all the data available, once with all the data but without the American Thyroid Association (ATA) risk score, and once with just the ATA risk score. When training on all feature measurements available in this dataset, their learning machines achieved their top predictive performance for thyroid cancer recurrence with accuracies of 96% for the SVM, 93% for the k-nearest neighbor, 96% for the decision tree, 95% for the random forest, and 96% for the artificial neural network. When the ATA measurements were excluded, models achieved improved specificity but a decrease in sensitivity, whereas when using only the ATA score, they observed an increase in sensitivity and a decrease in specificity.

Thyroid cancer recurrence prediction with machine learning, based on this dataset, has been the focus of additional research articles [18,20,22,23,27,28,30,31]. One study was focused on the development of explainable deep learning technology for thyroid cancer prediction [30]; however, they did not report predictive accuracies on held-out testing data, and so their results cannot be compared with our analysis. An additional study found that the XGBoost algorithm achieved 98% accuracy in predicting thyroid cancer recurrence [31]. Ensemble methods have also been developed for this application [27,28]. Approaches have also been developed that utilize synthetic minority oversampling (SMOTE) methods [18,20,23]; however, it should be noted that SMOTE analyses create synthetic data, and so performance metrics, such as accuracy, should not be directly compared with studies that rely purely on real-world data. Finally, an approach that relies upon few features in order to support more explainable AI was developed, which relied on distance correlation to select for response, risk, tumor grade and nodal status upon which to inform thyroid cancer recurrence prediction, achieving accuracies of 94.8 to 96.1% [22].

1.4. Problem Statement, Hypotheses and Contributions

Thyroid cancer recurrence is a major health problem, and so the creation of technologies that may assist in accurate explainable predictions of thyroid cancer recurrence and risk have the potential to improve the standard of patient care. In this analysis, we hypothesize that a thorough assessment of a wide variety of machine learning and feature selection algorithms, inclusive of a novel and emerging redundancy-aware based feature selection approach, performed with public domain software, will assist in the creation of high-performing models for the prediction of thyroid cancer recurrence. We further hypothesize that by reducing the input features relied upon to inform prediction with feature selection algorithms, we can create high-performing models whose functionality is easier to explain through the reduced feature set. By identifying a reduced feature set relied upon for high-performing AI models, we can also inform the broader research community as to combinations of clinical variables that are highly predictive of thyroid cancer recurrence. We also hypothesize that machine learning can be relied upon to predict American Thyroid Association (ATA) risk scores. The main contributions of this thyroid cancer-focused manuscript are the consideration of novel redundancy-aware step-up feature selection, the consideration of an emerging deep learning algorithm optimized for tabular data (Gandalf), and the consideration of two applications in ATA risk score prediction in addition to the standard focus on thyroid cancer recurrence prediction.

2. Materials and Methods

As an overview, this manuscript will introduce the dataset the analysis is based upon (Section 2.1); introduce the learning technologies considered (Section 2.2); outline the statistical metrics employed for technology evaluation (Section 2.3); present the results for thyroid cancer

recurrence prediction (Section 3.1) and ATA risk prediction (Sections 3.2 and 3.3); discuss the thyroid cancer recurrence prediction results (Section 4.1), the binarized ATA risk prediction results (Section 4.2), and the regression based ATA risk prediction results (Section 4.3); comparatively discuss the learning machines addressed in this analysis (Section 4.4); and conclude with a discussion of strengths, limitations, and future work (Section 4.5).

2.1. Dataset Description

The dataset relied upon in this analysis is publicly available [32] under a Creative Commons 4.0 International Attribution License. The dataset was obtained from patients with differentiated thyroid cancer (DTC) and included a minimum of 10 years of follow-up, as part of a 15-year study timeframe, with additional details available in the literature [3]. The primary target variable in this dataset is the binary variable recurrence, indicating whether thyroid cancer recurred in a given patient within the 15-year timeframe, and a minimum of 10 years follow-up after treatment. The dataset was acquired from 383 patients, and thyroid cancer recurred in 108 patients (i.e., DTC did not recur in 275 patients) over the 15-year span of the study. Table 1 provides a detailed description of the specific categorical predictor features found in this dataset along with the different possible values for each categorical feature, the number of patients/samples in this dataset with a given categorical feature value (counts), as well as the proportions (percentages) of patients with each categorical feature value. Age is the only numerical variable with mean $\pm$ standard deviation (SD): 40.87 ± 15.13.

Table 1. Categorical data.

Measurement Name	Values with Counts and Proportions (%)
Gender	Female (312, 81.5%), Male (71, 18.5%)
Smoking	No (334, 82.2%), Yes (49, 12.8%)
History (Hx) of Smoking	No (355, 92.7%), Yes (28, 7.3%)
History of Radiotherapy	No (376, 98.2%), Yes 7 (1.8%)
Thyroid Function	Euthyroid (332, 86.7%), Clinical Hyperthyroidism (20, 5.2%), Subclinical Hypothyroidism (14, 3.7%), Clinical Hypothyroidism (12, 3.1%), Subclinical Hyperthyroidism (5, 1.3%)
Physical Examination	Multinodal goiter (140, 36.6%), single nodular goiter—right (140, 36.6%), single nodular goiter—left (89, 23.2%), normal (7, 1.8%), diffuse goiter (7, 1.8%)
Adenopathy	No (277, 72.3%), Right (48, 12.5%), Bilateral (32, 8.7%), Left (17, 4.4%), Extensive (7, 1.8%), Posterior (2, 0.5%)
Pathology	Papillary (287, 74.9%), micropapillary (48, 12.5%), follicular (28, 7.3%), Hurthel cell (20, 5.2%)
Focality	Uni-Focal (247, 64.4%), Multi-Focal (136, 35.5%)
Risk	Low (246, 65.0%), Medium (102, 26.6%), High (32, 8.4%)
T—Tumor	T1a (49, 12.8%), T1b (43, 11.2%), T2 (151, 39.4%), T3a (96, 25.1%), T3b (16, 4.2%), T4a (20, 5.2%), T4b (8, 2.1%)
N—Node	N0 (268, 70.0%), N1b (93, 24.3%), N1a (22, 5.7%)
M—Metastasis	M0 (365, 95.3%), M1 (18, 4.7%)
Stage	I (333, 87.0%), II (32, 8.4%), III (4, 1.0%), IVA (3, 0.8%), IVB (11, 2.9%)
Response	Biochemical Incomplete (23, 6.0%), Excellent (208, 54.2%), Indeterminate (61, 15.9%), Structural Incomplete (91, 23.7%)

2.2. Machine Learning

The machine learning for this research was performed using df-analyze [33], which is a public domain software package that implements machine learning classifiers and regression learners, along with statistical validation techniques, on tabular datasets. This includes feature type inference, feature description (e.g., univariate associations and statistics), data cleaning (e.g., missing value handling and imputation), training, and test splitting, feature selection, hyperparameter tuning, and validation (e.g., K-fold and holdout validation) [33]. This benchmarking software package, which facilitates running a thorough competition between leading learning technologies for a given dataset, known as df-analyze, was first used in a neuroscience study on patients with schizophrenia from features extracted from magnetic resonance imaging examinations [34], and it has since been applied to mitigating bias in traffic stop outcomes [35], the diagnosis of chronic kidney disease [36], studying proteins potentially involved with learning in the cerebral cortex [37], and diagnosing, predicting treatment, and staging of pediatric appendicitis [38]. The software was configured to exhaustively compare all combinations of learning algorithms (light gradient boosting machine—lgbm, random forest—rf, logistic regression—lr, stochastic gradient descent—sgd, a tabular data optimized deep learner—gandalf [39], and the k-nearest neighbor algorithm—knn) and feature selection methods (no feature selection—none, filter-based association feature selection—assoc, filter-based prediction feature selection—pred, redundancy-aware step-up wrapper-based feature selection—wrap, embedded-based lgbm feature selection—embed_lgbm, and embedded-based linear feature selection—embed_linear) with full documentation available as part of the df-analyze software package, version 3.3.0 [33]. Note that the redundancy-aware feature selection algorithm is an emerging and novel technique that is a modification of forward stepwise feature selection designed to ensure that redundant features (ones that do not provide substantial additional predictive capacity relative to those features already included) are not added to the model [40]. The software is a single-phase selection technology that is publicly available with a detailed description online on github [40]. This is the first study to report on the performance of redundancy-aware feature selection [40] in thyroid cancer recurrence and risk prediction. This is also the first study to report on the performance of an emerging tabular data optimized deep learner, Gandalf [39], in thyroid cancer recurrence and risk prediction.

2.3. Statistical Analysis

Our statistical analysis was completed twice: implementing k-fold validation with 5 folds, as well as holdout validation, with a large 40% of samples reserved for the holdout set, to help ensure the reliability and reproducibility of the findings. Optuna hyperparameter tuning was completed with 50 runs. The first analysis was focused on predicting whether thyroid cancer would recur—a classification technology for which our statistics, outlined below, are defined in terms of the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). For the comparative classification analyses, we provide a thorough set of statistics, including the overall accuracy (acc = (TP + TN)/(TP + TN + FP + FN)), sensitivity (sens = TP/(TP + FN)), specificity (spec = TN/(TN + FP)), area under the receiver operating characteristic curve (auroc, the area under the profile of performance tradeoff between sensitivity and specificity), balanced accuracy (bal-acc = (sens + spec)/2), F1 Score (f1 = TP/(TP + 1/2(FP + FN))), negative predictive value (npv = TN/(TN + FN)), and positive predictive value (ppv = TP/(TP + FP)). The second and third analyses were blinded/agnostic to whether the thyroid cancer would recur (the target variable in the first analysis) and focused on predicting the ATA risk score. In the second analysis, we report on prediction of the ATA risk score as a classification learning task; however, it should be

noted that there are insufficient samples in the higher risk categories of the ATA risk score in this dataset to address this task as a multi-class classification problem with three classes. As such, for this experiment, we grouped the intermediate and high ATA risk categories into a single class and addressed the problem as a binary classification task, predicting low vs. elevated (intermediate/high) ATA risk scores with machine learning. In the third analysis, we report on prediction of the ATA risk score as a regression learning task targeting an ordinal encoding of the three ATA risk score categories. For this regression analysis, we provide a thorough set of statistics, including the mean absolute error (mae, which is the average unsigned amount of error in predictions), median absolute error (mdae, which is the sorted median of error in predictions), mean squared error (msqe, the square of the signed differences between predictions and ground truth), coefficient of determination or R^2 (r2 = 1 − Sum of Squares of Residuals/Total Sum of Squares), and the variance (var) explained (var-exp = 1 − var(target − predictions)/var(predictions)).

3. Results

3.1. Recurrence Prediction

The main findings of two types of validation, comprising a detailed comparison between all combinations of supported machine learning and feature selection technologies, applied to predicting thyroid cancer recurrence, are reported here. Table A1 provides the leading validation results on the holdout set for recurrence prediction. Table A2 provides the leading k-fold validation results on the holdout set. From the results provided in Tables A1 and A2, we observe that the wrapper-based redundancy-aware step-up feature selection algorithm (wrap) produces particularly high-quality results when paired with a variety of learning machines (lgbm, lr, rf, sgd, knn). The wrapper-based feature selection results are provided in Table A3.

Note that our highly accurate models (with about 95% accuracy) based on this reduced feature set from redundancy-aware wrapper-based feature selection rely on just three features: age, whether their tumor's stage was III, IVA, or IVB, and whether their tumor's response was "Structurally Incomplete". It is also noteworthy that in feature selection reports, such as the one provided in Table A3, the features are sorted in order of apparent predictive importance. Thus, this method implies that Response and Stage are more predictive of thyroid cancer recurrence than age, though age is assessed to still have value in thyroid cancer recurrence prediction.

3.2. Secondary Application: ATA Risk Score Prediction—Binary Classification

We performed an analysis predicting ATA risk scores as a binary classification task with the intermediate- and high-risk target classes grouped together to create an elevated risk category. This was performed as the elevated risk subcategories were too underrepresented in the dataset to run the df-analyze software employed. To summarize, our leading models achieved accuracies of 86.4 to 93.2% across our two validation methods based on the tabular data optimized deep learner, Gandalf [39], with no feature selection. Detailed results are provided in Table A4.

3.3. Tertiary Application: ATA Risk Score Prediction—Regression

Our thorough competition serves as a method for comparative validation between combinations of feature selection and machine learning algorithms for the task of ATA risk prediction. This task was addressed as a regression task with the target variable encoded as an ordinal, from a dataset that excluded knowledge of whether the patient's thyroid cancer would recur, and the results are provided here. Table A5 provides the leading validation results on the holdout set, and Table A6 provides the leading K-fold

validation results on the holdout set. The results demonstrate that it is possible to create continuous regression models based on an ordinal representation of ATA risk. The filter-based association (assoc) feature selection method, alongside no feature selection (none), produced the highest-performing models. The features selected by the algorithm (assoc) are provided in Table A7.

It is noteworthy that in feature selection reports, such as the one provided in Table A7, that the categorical features are sorted in order of predictive importance. Thus, in Table A7, Adenopathy and Response are highly predictive of ATA risk, whereas being a smoker or having a history of smoking are much less predictive. The final selected features from the filter-based association feature selection method were established from the ranking in Table A7, and they included Age, Adenopathy, Response, Thyroid_Function, Pathology, Physical Examination, Focality, Stage, History of Radiotherapy, Gender, Nodal status (N), History of Smoking, and Smoking. Figure 1 is provided to visualize the regression predictions of leading models (one LGBM, one ElasticNet) alongside the ground truth those learning machines were tasked with predicting.

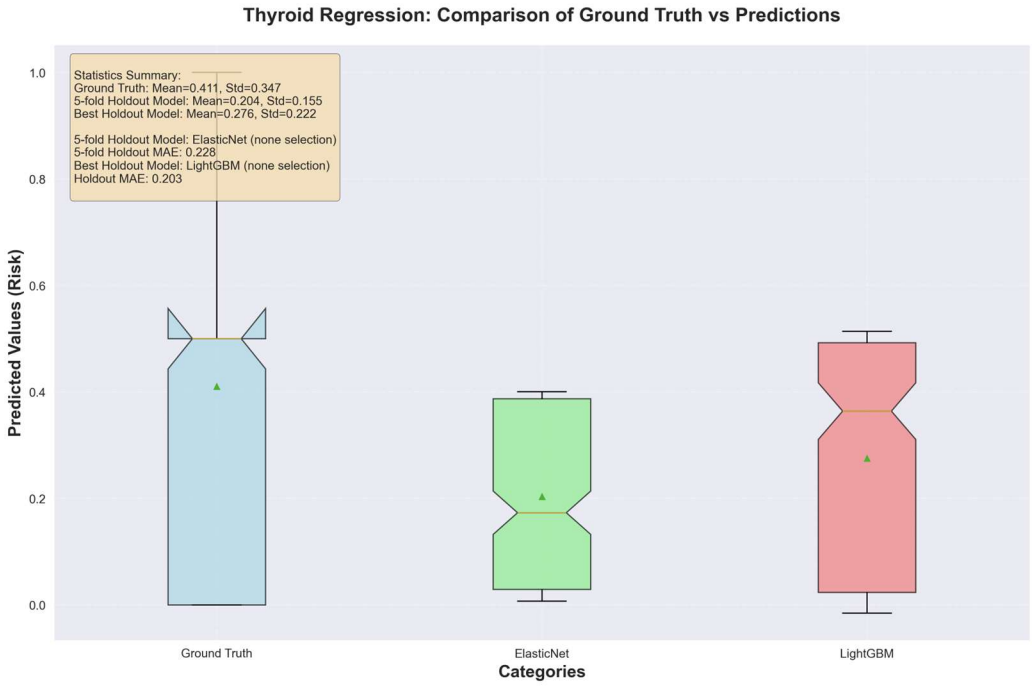

Figure 1. Boxplots of ground truth target variables and leading regression prediction models of ordinally encoded ATA risk.

4. Discussion

This study presents the results of a thorough assessment of many combinations of feature selection algorithms and learning machines for the prediction of three tasks: thyroid cancer recurrence, encoding of ATA risk score prediction as a binary classification task, and ordinal encoding of ATA risk score prediction as a regression target variable. The results demonstrate high-quality results with feature selection reports that assist in identifying leading factors that appear to contribute to thyroid cancer recurrence as well as technologies and underlying features that can potentially contribute to predictive assessments of ATA risk scores.

Once a patient has been diagnosed with thyroid cancer, our models can be relied upon clinically to predict whether or not the patient will experience tumor recurrence. This information would thus be available to the clinicians prior to commencement of the follow-up period. Patients for whom the AI predicts recurrence could theoretically, at the order

of the managing clinician, be monitored for recurrence more aggressively (more frequent examinations, additional types of examinations applied, etc.). In turn, this additional monitoring could potentially assist in identifying thyroid cancer recurrence earlier on, thus supporting earlier treatment. Earlier treatment is broadly associated with better patient outcomes; as such, AI models for thyroid cancer recurrence prediction have the potential to improve the standard of patient care.

4.1. Thyroid Cancer Recurrence

We were able to create models predicting thyroid cancer recurrence with ~95% accuracy, using just three underlying feature measurements: age, stage, and response. Our feature selection method also further informs us that knowledge of whether the tumor's stage was elevated (i.e., stages III, IVA, and IVB) or not is associated with improved accuracy of thyroid cancer recurrence prediction. Our feature selection method also further informs us that knowledge of whether the tumor response was structurally incomplete (or not) is also associated with improved accuracy of thyroid cancer recurrence prediction. These findings help with creating AI models whose functionality is far easier to explain to the clinicians and patients who might benefit from such technology. These findings also help illustrate the value of these three highly predictive features of tumor recurrence, which is potentially useful information for the broader medical community focused on thyroid cancer management. Our findings produce similar accuracies to those reported on in the literature [3,22,31]; however, our approach, based on the emerging redundancy-aware step-up feature selection algorithm, produced a small set of just three underlying features (age, stage and response) that together produce very high-quality recurrence prediction models, potentially helping to create more explainable AI technologies in this domain.

4.2. ATA Binarized Risk Prediction

Our analysis demonstrated two methods for predicting ATA risk scores, including binarized- and regression-based models. Results from binarized prediction provide early promising results for the emerging Gandalf deep learner [39], which is optimized for tabular data, and was the winning learning algorithm in this task. Historically, deep learners have been underperformers when basing predictions on tabular data; thus, these findings indicate considerable potential from future developments in deep learning applied to tabular datasets.

4.3. Regression Risk Prediction of Ordinally Encoded ATA Risk

For our regression-based models, the risk score was encoded as an ordinal variable or a sorted categorical variable. This involves assigning a numerical value to each of the risk scores (for example: low = 1, intermediate = 2, high = 3) and then training a regression-based learning algorithm to predict those values. The advantages of taking such an approach are multiple. This method allows the direct prediction of all categorical class values, even though the elevated risk categories (intermediate and high) are underrepresented in the dataset, making classical multi-class classification for three target classes impossible in our dataset. This method also allows for nuanced per-patient predictions of risk. For example, if a patient's machine learning assessed risk is in between low and intermediate, or in between intermediate and high, the regression-based learners will communicate this assessment to the clinician and patient through predictions between 1 and 2, or 2 and 3, respectively. Thus, a predicted risk score between 1 and 2 implies a patient at elevated risk relative to those assessed as low risk but at reduced risk relative to those patients assessed as intermediate risk. Similarly, predicted scores between 2 and 3 imply reduced risk relative to the high-risk patients and elevated risk relative to the intermediate-risk patients. Thus, this approach to risk prediction provides potential benefits in patient-

specific risk assessments, potentially supporting AI-based risk assessment technology to assist in improving approaches to personalized medicine. Technologies that can predict ATA risk scores have the potential to replace the sometimes subjective and clinician-specific assessment of patient risk. Predicting ATA risk scores was not the focus of previous literature studies; as such, our findings in these two applications are novel. Figure 1 was provided to visualize the distribution of predictions for leading regression models in this application. Note that although risk was encoded as 1, 2 and 3, df-analyze internally employs min–max normalization, which maps all targets and predictions to floating point values in the 0 to 1 range (hence the range of the *y* axis in Figure 1). The visualized figure implies that the models are struggling to correctly predict the high-risk category, which is severely under-sampled in this dataset. Future work will benefit from increased examples of patients with high-risk ATA categorization. Note that the leading models were created from either no feature selection or the filter-based association feature selection method. The filter-based association FS method tends to select for large numbers of features available; thus, these findings imply that the leading regression ATA risk score models are based on a large proportion of the feature measurements available, potentially implying that additional features that correlate with ATA risk will be needed to further optimize the performance of these models.

4.4. Machine Learning Algorithm Discussion

For recurrence prediction, the leading methods were tied: lr, sgd, knn, rf, and lgbm. This represents almost all machine learning algorithms considered with only the newly emerging and experimental deep learning based method, Gandalf, underperforming. In this situation, there are no obvious major advantages nor disadvantages in choosing any of the leading techniques in terms of model predictive performance/accuracy. However, the KNN classifier offers advantages in terms of example-based explainability, as the underlying technology can provide the user with the feature measurements paired with the target class labels of the k-nearest neighbor training samples to the test sample, providing a level of illustrative explainability not possible with the other techniques. Although the decision tree is a highly explainable learning machine, it consistently underperforms on its own, and so it is not included in the main trial. Conversely, the random forest and the light gradient boosting machine, which are each based on the decision tree, are strong performers; however, they continue to be a challenge to explain on a per-sample basis, as aggregating and interpreting the reason for predictions across many decision trees is difficult. Logistic regression and stochastic gradient descent provide statistical learning baselines to compare all techniques against. When performing regression-based risk prediction, elastic regression and LGBM were the leading techniques; however, Gandalf and KNN also performed well. It was noteworthy that for binarized risk prediction, the emergent deep learner built for tabular data, Gandalf, was the winner, demonstrating considerable potential from developments in deep learning technology applied to tabular data. Deep learners particularly benefit from large sample sizes for training; thus, Gandalf may be at a disadvantage in our study, which may help explain its underperformance in the first and third applications.

4.5. Strengths, Limitations and Future Work

The main limitation of this study is the limited number of samples in this dataset. Future work will apply these methods to larger datasets that are inclusive of more samples with high ATA risk scores. An additional limitation was the set of features considered for this study. Had additional feature measurements been acquired from each patient, we may have more interesting feature selection findings to report as well as potentially

helping to produce more highly accurate predictive models. An additional limitation of this study was that we were unable to assess the generalization of the study findings due to a lack of public domain datasets derived from additional populations. Thus, future work will re-apply the methods outlined in this paper on independent datasets after they become available. An additional limitation of this study is that the technology developed is dependent on the availability of the input feature measurements relied upon for basic predictive functionality; as such, in situations where this information is not available (such as clinical assessment of tumor grade), data imputation methods would need to be relied upon, adding an uncharacterized source of error. Additionally, the hosting of AI models on the internet would involve the transmission of patient information to the website hosting the AI technology, which brings with it a risk of privacy violations. Fortunately, our models do not rely on sufficient patient identifying information to uniquely identify a patient from the predictor variables that inform model predictions; as such, it is expected that in future deployments of this technology, patient privacy should not be at risk. However, in order to be extra cautious, such a website could employ encryption technologies in order to help ensure that no third party is capable of intercepting any patient data, even if it is anonymized. An additional limitation of this study was that our regression analysis produced negative values for the coefficient of determination (R2) in multiple technologies, implying potentially poorly fitted models. This finding reinforces our primary evaluation statistic, MAE, which indicates that the average absolute error is approximately 20–30% of the range of the target variable. While this component of the study presents novel methods for a regression assessment of ATA risk, the results imply that future work will be needed. Adding new feature measurements predictive of risk in future studies could improve the quality of fit to the training data and thus support reliable risk predictions. Finally, false predictions are a major limitation of the technology created, of which there are two main types: falsely indicating to a patient that they will have recurrent thyroid cancer when they will not (a problem that might involve the patient receiving additional unnecessary surveillance as well as an unnecessary scare) as well as falsely indicating to a patient that they will not have recurrent thyroid cancer when they will (a much more dangerous problem, as potentially life-saving additional surveillance may not be applied to the patient). Future work will investigate the potential to create AI models with custom operating points that minimize the more serious second type of error at the expense of some additional cases of the earlier type of less serious error. This will involve a study thoroughly evaluating the tradeoffs of AI models with varying operating points and thus varying rates of each type of error.

Strengths of this study include that it was based on an open-source dataset that included a long follow-up on each patient to assess thyroid cancer recurrence (10 years minimum as part of an overall 15-year study). Strengths also include having devoted 40% of our dataset to the holdout testing group, which helps ensure reliability in study findings. However, as a result, fewer samples were available for training our models, potentially resulting in reduced predictive performances reported to what could have been obtained had we trained on larger training datasets and relied on fewer samples for our statistical assessment of predictive performance. This issue may help explain the modest discrepancies between our reported predictive accuracies and those in the literature [3,31]. Strengths also include the use of open-source software, which facilitates the assessment of the potential of a wide variety of feature selection and machine learning technologies, the identification of small subsets of features that together produce reliable highly predictive technologies, and the consideration of ATA risk prediction by multiple methods. Strengths also include the consideration of a novel emerging redundancy-aware step-up wrapper-based feature selection algorithm and the consideration of an emerging deep learner

optimized for tabular data. Future work can address the potential for microRNA-based risk scores to inform thyroid cancer recurrence prediction as well as to correlate microRNA-based risk scores to established ATA risk scores.

Author Contributions: Conceptualization, M.A.P., and J.L.; methodology, M.A.P., D.B., X.G.; df-analyze software, D.B., X.G., and J.L.; Validation, D.B., J.L., X.G.; Formal analysis, M.A.P., D.B., X.G., J.L.; Investigation, M.A.P.; Resources, D.B. and J.L.; Data curation, M.A.P.; Writing—original draft preparation, M.A.P.; Writing—review and editing, M.A.P., and J.L.; Supervision, J.L.; Project administration, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by a Canada Foundation for Innovation grant, a Nova Scotia Research and Innovation Trust grant, an NSERC Discovery grant, a Compute Canada Resource Allocation, and a Nova Scotia Health Authority grant to J.L.

Institutional Review Board Statement: The original study detailing this dataset indicated that the "article was submitted with the ethical identifier IR.UMSHA.REC.1402.360 at Hamadan University of Medical Sciences, Hamadan, Iran" [3]. The results presented in this manuscript involved only secondary analysis of de-identified data. The dataset used in this study is publicly available and so institutional review board approval was not required to complete this retrospective analysis.

Informed Consent Statement: The original study detailing this dataset indicated that the "article was submitted with the ethical identifier IR.UMSHA.REC.1402.360 at Hamadan University of Medical Sciences, Hamadan, Iran" [3]. The results presented in this manuscript involved only secondary analysis of de-identified data, thus informed consent was not needed for this analysis.

Data Availability Statement: The dataset used in this study is publicly available and can be accessed from Kaggle at https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence (accessed on 7 May 2025). No new data were created or collected specifically for this study. Since this was a retrospective analysis of public domain data, no institutional review board approval was necessary for conducting this study.

Conflicts of Interest: Jacob Levman is the founder of Time Will Tell Technologies, Inc. The authors declare no relevant conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANN	Artificial Neural Network
ATA	American Thyroid Association
auroc	Area under the Receiver Operating Characteristic Curve
DTC	Differentiated Thyroid Cancer
FN	False Negative
FP	False Positive
KNN/knn	K Nearest Neighbor
lgbm	Light Gradient Boosting Machine
lr	Logistic Regression
M	Metastasis
N	Node
rf	Random Forest
SD	Standard Deviation
sgd	Stochastic Gradient Descent
SMOTE	Synthetic Oversampling Method
SVM	Support Vector Machine
T	Tumor

Tg	Thyroglobulin
TN	True Negative
TP	True Positive
TSH	Thyroid Stimulating Hormone

Appendix A

Table A1. Recurrence prediction competition on holdout set.

Model	Selection	Embed_Selector	acc	auroc	bal-acc	f1	npv	ppv	sens	spec
lr	wrap	none	0.948	0.945	0.919	0.933	0.958	0.944	0.919	0.852
sgd	wrap	none	0.948	0.945	0.919	0.933	0.958	0.944	0.919	0.852
knn	wrap	none	0.948	0.937	0.919	0.933	0.958	0.944	0.919	0.852
rf	wrap	none	0.948	0.961	0.919	0.933	0.958	0.944	0.919	0.852
lgbm	wrap	none	0.948	0.943	0.919	0.933	0.958	0.944	0.919	0.852
lgbm	pred	none	0.885	0.980	0.914	0.871	0.716	0.991	0.914	0.981
lgbm	embed_lgbm	lgbm	0.885	0.977	0.914	0.871	0.716	0.991	0.914	0.981
rf	pred	none	0.885	0.975	0.897	0.868	0.735	0.967	0.897	0.926
lr	pred	none	0.880	0.978	0.910	0.865	0.707	0.991	0.910	0.981
lr	embed_lgbm	lgbm	0.880	0.977	0.910	0.865	0.707	0.991	0.910	0.981
lr	none	none	0.874	0.975	0.907	0.860	0.697	0.991	0.907	0.981
rf	none	none	0.874	0.980	0.896	0.858	0.708	0.975	0.896	0.944
lr	assoc	none	0.869	0.975	0.903	0.855	0.688	0.991	0.903	0.981
rf	embed_lgbm	lgbm	0.869	0.977	0.898	0.854	0.693	0.983	0.898	0.963
lgbm	none	none	0.864	0.975	0.899	0.850	0.679	0.991	0.899	0.981
lgbm	assoc	none	0.864	0.975	0.894	0.848	0.684	0.983	0.894	0.963
gandalf	embed_lgbm	lgbm	0.843	0.940	0.846	0.819	0.676	0.935	0.846	0.852
knn	assoc	none	0.832	0.924	0.866	0.816	0.637	0.973	0.866	0.944
knn	none	none	0.832	0.924	0.866	0.816	0.637	0.973	0.866	0.944
rf	assoc	none	0.812	0.970	0.852	0.796	0.607	0.972	0.852	0.944
knn	embed_lgbm	lgbm	0.806	0.914	0.848	0.791	0.600	0.972	0.848	0.944
knn	pred	none	0.806	0.909	0.854	0.792	0.598	0.981	0.854	0.963
sgd	assoc	none	0.775	0.843	0.843	0.765	0.557	1.000	0.843	1.000
sgd	none	none	0.775	0.843	0.843	0.765	0.557	1.000	0.843	1.000
sgd	pred	none	0.775	0.843	0.843	0.765	0.557	1.000	0.843	1.000
sgd	embed_lgbm	lgbm	0.775	0.843	0.843	0.765	0.557	1.000	0.843	1.000
dummy	assoc	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	none	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	embed_lgbm	lgbm	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	wrap	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	pred	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
gandalf	assoc	none	0.681	0.670	0.542	0.538	0.387	0.738	0.542	0.222
gandalf	wrap	none	0.665	0.788	0.744	0.658	0.455	0.951	0.744	0.926
gandalf	none	none	0.293	0.925	0.507	0.237	0.286	1.000	0.507	1.000
gandalf	pred	none	0.283	0.810	0.500	0.220	0.283	nan	0.500	1.000

Table A2. Recurrence prediction competition 5-fold performance on holdout set.

Model	Selection	Embed_Selector	acc	auroc	bal-acc	f1	npv	ppv	sens	spec
knn	wrap	none	0.959	0.940	0.937	0.943	0.967	0.963	0.937	0.889
rf	wrap	none	0.948	0.950	0.919	0.929	0.963	0.949	0.919	0.853
lr	wrap	none	0.948	0.960	0.919	0.929	0.963	0.949	0.919	0.853
rf	none	none	0.948	0.997	0.924	0.930	0.950	0.955	0.924	0.871
lgbm	wrap	none	0.948	0.923	0.919	0.929	0.963	0.949	0.919	0.853
sgd	wrap	none	0.948	0.911	0.919	0.929	0.963	0.949	0.919	0.853
rf	pred	none	0.943	0.994	0.915	0.923	0.947	0.949	0.915	0.853
lr	pred	none	0.922	0.995	0.895	0.894	0.921	0.946	0.895	0.835
lgbm	none	none	0.917	0.997	0.897	0.888	0.890	0.953	0.897	0.853
lr	none	none	0.911	0.996	0.882	0.881	0.913	0.939	0.882	0.816
lr	assoc	none	0.911	0.996	0.882	0.881	0.913	0.939	0.882	0.816
rf	embed_lgbm	lgbm	0.911	0.997	0.898	0.889	0.882	0.958	0.898	0.871
lr	embed_lgbm	lgbm	0.911	0.997	0.898	0.891	0.893	0.958	0.898	0.871
gandalf	assoc	none	0.906	0.993	0.861	0.860	0.938	0.924	0.861	0.760
sgd	embed_lgbm	lgbm	0.901	0.886	0.886	0.876	0.859	0.950	0.886	0.853
knn	none	none	0.896	0.918	0.852	0.857	0.911	0.920	0.852	0.756
knn	assoc	none	0.896	0.918	0.852	0.857	0.911	0.920	0.852	0.756
gandalf	embed_lgbm	lgbm	0.896	0.905	0.848	0.861	0.895	0.908	0.848	0.740
lgbm	pred	none	0.890	0.992	0.873	0.860	0.848	0.944	0.873	0.835
lgbm	assoc	none	0.890	0.993	0.878	0.865	0.879	0.953	0.878	0.853
knn	pred	none	0.890	0.948	0.873	0.861	0.856	0.944	0.873	0.835
sgd	assoc	none	0.885	0.863	0.864	0.848	0.848	0.940	0.864	0.816
sgd	pred	none	0.885	0.869	0.869	0.855	0.838	0.944	0.869	0.835
sgd	none	none	0.880	0.848	0.848	0.833	0.866	0.930	0.848	0.778
knn	embed_lgbm	lgbm	0.880	0.953	0.876	0.857	0.830	0.958	0.876	0.871
rf	assoc	none	0.848	0.989	0.859	0.831	0.838	0.963	0.859	0.889
lgbm	embed_lgbm	lgbm	0.843	0.997	0.850	0.821	0.835	0.958	0.850	0.871
dummy	none	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	embed_lgbm	lgbm	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	assoc	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	wrap	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
dummy	pred	none	0.717	0.500	0.500	0.418	nan	0.717	0.500	0.000
gandalf	pred	none	0.643	0.780	0.507	0.409	0.643	0.712	0.507	0.200
gandalf	wrap	none	0.596	0.816	0.605	0.465	0.617	0.816	0.605	0.618

Table A3. Features selected by wrapper based step-up feature selection. Higher numbers imply more predictive value from that measurement.

Feature	Score
Response_Structural Incomplete	9.425×10^{-1}
Stage_nan	9.425×10^{-1}
Age	9.325×10^{-1}

Table A4. ATA risk binary classification prediction results on the holdout set.

Model	Selection	Embed_Selector	acc	auroc	bal-acc	f1	npv	ppv	sens	spec
gandalf	assoc	none	0.890	0.898	0.915	0.886	0.761	1.000	0.915	1.000
gandalf	none	none	0.864	0.933	0.881	0.858	0.741	0.962	0.881	0.940
knn	assoc	none	0.853	0.944	0.873	0.848	0.724	0.962	0.873	0.940
knn	none	none	0.853	0.944	0.873	0.848	0.724	0.962	0.873	0.940
sgd	assoc	none	0.838	0.974	0.865	0.833	0.696	0.970	0.865	0.955
sgd	none	none	0.838	0.974	0.865	0.833	0.696	0.970	0.865	0.955
gandalf	embed_lgbm	lgbm	0.817	0.932	0.828	0.808	0.690	0.916	0.828	0.866
sgd	pred	none	0.801	0.972	0.843	0.799	0.641	0.989	0.843	0.985
knn	pred	none	0.791	0.916	0.835	0.788	0.629	0.988	0.835	0.985
gandalf	wrap	none	0.791	0.903	0.804	0.782	0.655	0.904	0.804	0.851
sgd	embed_lgbm	lgbm	0.749	0.836	0.803	0.748	0.584	0.987	0.803	0.985
knn	wrap	none	0.712	0.805	0.713	0.699	0.571	0.822	0.713	0.716
knn	embed_lgbm	lgbm	0.712	0.833	0.771	0.712	0.551	0.973	0.771	0.970
lr	none	none	0.712	0.965	0.771	0.712	0.551	0.973	0.771	0.970
lr	assoc	none	0.712	0.967	0.771	0.712	0.551	0.973	0.771	0.970
lr	pred	none	0.707	0.966	0.767	0.707	0.546	0.972	0.767	0.970
lgbm	pred	none	0.702	0.933	0.763	0.701	0.542	0.972	0.763	0.970
lgbm	embed_lgbm	lgbm	0.696	0.911	0.756	0.696	0.538	0.958	0.756	0.955
lgbm	none	none	0.696	0.925	0.756	0.696	0.538	0.958	0.756	0.955
lr	embed_lgbm	lgbm	0.696	0.956	0.759	0.696	0.537	0.971	0.759	0.970
rf	wrap	none	0.691	0.901	0.731	0.689	0.537	0.892	0.731	0.866
rf	none	none	0.691	0.907	0.731	0.689	0.537	0.892	0.731	0.866
rf	pred	none	0.691	0.897	0.731	0.689	0.537	0.892	0.731	0.866
rf	embed_lgbm	lgbm	0.691	0.931	0.731	0.689	0.537	0.892	0.731	0.866
rf	assoc	none	0.691	0.928	0.731	0.689	0.537	0.892	0.731	0.866
lr	wrap	none	0.691	0.906	0.731	0.689	0.537	0.892	0.731	0.866
lgbm	assoc	none	0.691	0.931	0.731	0.689	0.537	0.892	0.731	0.866
sgd	wrap	none	0.691	0.848	0.731	0.689	0.537	0.892	0.731	0.866
lgbm	wrap	none	0.691	0.927	0.731	0.689	0.537	0.892	0.731	0.866
dummy	assoc	none	0.649	0.500	0.500	0.394	nan	0.649	0.500	0.000
dummy	embed_lgbm	lgbm	0.649	0.500	0.500	0.394	nan	0.649	0.500	0.000
dummy	none	none	0.649	0.500	0.500	0.394	nan	0.649	0.500	0.000
dummy	wrap	none	0.649	0.500	0.500	0.394	nan	0.649	0.500	0.000
dummy	pred	none	0.649	0.500	0.500	0.394	nan	0.649	0.500	0.000
gandalf	pred	none	0.424	0.818	0.529	0.402	0.366	0.733	0.529	0.881

Table A5. ATA risk prediction results on the holdout set. regression targeting of ordinal encoding of ATA risk.

Model	Selection	Embed_Selector	mae	mdae	msqe	r2	var-exp
lgbm	none	none	0.203	0.029	0.105	0.125	0.277
lgbm	embed_linear	linear	0.207	0.075	0.099	0.178	0.298
lgbm	pred	none	0.234	0.172	0.103	0.141	0.361
elastic	none	none	0.259	0.161	0.124	−0.027	0.329
knn	embed_linear	linear	0.293	0.200	0.186	−0.544	−0.011
elastic	assoc	none	0.295	0.206	0.153	−0.269	0.270
elastic	pred	none	0.311	0.244	0.164	−0.361	0.250
knn	pred	none	0.313	0.300	0.185	−0.534	0.140
knn	none	none	0.347	0.400	0.220	−0.826	0.064
knn	assoc	none	0.347	0.400	0.220	−0.826	0.064
lgbm	assoc	none	0.372	0.397	0.220	−0.832	0.081
elastic	embed_linear	linear	0.402	0.465	0.268	−1.228	0.001
elastic	embed_lgbm	lgbm	0.404	0.477	0.271	−1.248	0.000
elastic	wrap	none	0.404	0.477	0.271	−1.248	0.000
lgbm	wrap	none	0.404	0.477	0.271	−1.248	0.000
lgbm	embed_lgbm	lgbm	0.405	0.472	0.272	−1.263	−0.009
sgd	none	none	0.406	0.489	0.282	−1.344	0.016
knn	embed_lgbm	lgbm	0.407	0.500	0.284	−1.357	0.000
sgd	assoc	none	0.409	0.495	0.286	−1.373	0.008
sgd	pred	none	0.409	0.494	0.284	−1.363	0.004
sgd	embed_linear	linear	0.410	0.499	0.288	−1.396	0.001
sgd	wrap	none	0.411	0.500	0.289	−1.403	0.000
rf	pred	none	0.411	0.500	0.289	−1.404	0.000
rf	wrap	none	0.411	0.500	0.289	−1.404	0.000
rf	none	none	0.411	0.500	0.289	−1.404	0.000
rf	embed_linear	linear	0.411	0.500	0.289	−1.404	0.000
dummy	assoc	none	0.411	0.500	0.289	−1.404	0.000
rf	assoc	none	0.411	0.500	0.289	−1.404	0.000
dummy	wrap	none	0.411	0.500	0.289	−1.404	0.000
dummy	pred	none	0.411	0.500	0.289	−1.404	0.000
dummy	none	none	0.411	0.500	0.289	−1.404	0.000
dummy	embed_linear	linear	0.411	0.500	0.289	−1.404	0.000
dummy	embed_lgbm	lgbm	0.411	0.500	0.289	−1.404	0.000
rf	embed_lgbm	lgbm	0.411	0.500	0.289	−1.404	0.000
knn	wrap	none	0.411	0.500	0.289	−1.404	0.000
sgd	embed_lgbm	lgbm	0.411	0.499	0.289	−1.400	0.001

Table A6. ATA risk prediction results: five-fold performance on holdout set and regression targeting of ordinal encoding of ATA risk.

Model	Selection	Embed_Selector	mae	mdae	msqe	r2	var-exp
elastic	none	none	0.228	0.190	0.078	−0.536	0.015
elastic	assoc	none	0.234	0.190	0.081	−0.591	0.042
lgbm	assoc	none	0.236	0.209	0.087	−0.748	−0.087
lgbm	none	none	0.243	0.222	0.089	−0.789	−0.117
dummy	assoc	none	0.256	0.300	0.128	−1.255	0.000
dummy	embed_lgbm	lgbm	0.256	0.300	0.128	−1.255	0.000
dummy	embed_linear	linear	0.256	0.300	0.128	−1.255	0.000
dummy	none	none	0.256	0.300	0.128	−1.255	0.000
dummy	pred	none	0.256	0.300	0.128	−1.255	0.000
dummy	wrap	none	0.256	0.300	0.128	−1.255	0.000
elastic	pred	none	0.257	0.244	0.096	−0.906	−0.087
knn	none	none	0.258	0.270	0.099	−1.101	−0.290
knn	assoc	none	0.258	0.270	0.099	−1.101	−0.290
lgbm	embed_linear	linear	0.264	0.245	0.101	−1.037	−0.319
elastic	embed_linear	linear	0.276	0.282	0.113	−1.192	−0.040
knn	embed_linear	linear	0.284	0.290	0.120	−1.453	−0.357
knn	embed_lgbm	lgbm	0.291	0.300	0.124	−1.458	−0.121
lgbm	pred	none	0.297	0.274	0.123	−1.556	−0.727
knn	pred	none	0.302	0.300	0.124	−1.732	−0.613
sgd	pred	none	0.306	0.295	0.132	−1.596	−0.277
sgd	assoc	none	0.312	0.297	0.142	−1.776	−0.348
elastic	wrap	none	0.319	0.365	0.158	−1.949	0.000
lgbm	wrap	none	0.319	0.365	0.158	−1.949	0.000
elastic	embed_lgbm	lgbm	0.322	0.343	0.151	−1.935	−0.182
sgd	embed_linear	linear	0.322	0.320	0.150	−1.910	−0.289
lgbm	embed_lgbm	lgbm	0.324	0.373	0.158	−2.037	−0.196
knn	wrap	none	0.355	0.400	0.175	−2.325	−0.011
sgd	none	none	0.379	0.366	0.214	−3.081	−0.426
rf	pred	none	0.398	0.383	0.266	−4.905	0.000
rf	wrap	none	0.412	0.400	0.290	−5.568	0.000
sgd	wrap	none	0.412	0.400	0.290	−5.572	−0.000

Table A7. Continuous and categorical features selected for by filter-based association feature selection. Higher numbers imply more predictive value from that measurement.

Continuous Features (Mutual Information: Higher = More important)

	mut_info
Age	1.952×10^{-2}

Categorical Features (Kruskal-Wallace H: Higher = More important)

	H
Adenopathy	3.651×10^{2}
Response	3.331×10^{2}
Thyroid_Function	3.325×10^{2}
Pathology	3.146×10^{2}
Physical_Examination	2.994×10^{2}
Focality	2.759×10^{2}
Stage	9.192×10^{0}
Hx_Radiothreapy	6.477×10^{0}
Gender	1.474×10^{0}
N	5.113×10^{-1}
Hx_Smoking	3.327×10^{-1}
Smoking	1.034×10^{-2}

References

1. Cabanillas, M.E.; McFadden, D.G.; Durante, C. Thyroid Cancer. *Lancet* **2016**, *388*, 2783–2795. [CrossRef] [PubMed]
2. Boucai, L.; Zafereo, M.; Cabanillas, M.E. Thyroid Cancer: A Review. *JAMA* **2024**, *331*, 425–435. [CrossRef] [PubMed]
3. Borzooei, S.; Briganti, G.; Golparian, M.; Lechien, J.R.; Tarokhian, A. Machine Learning for Risk Stratification of Thyroid Cancer Patients: A 15-Year Cohort Study. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 2095–2104. [CrossRef] [PubMed]
4. Giuffrida, D.; Prestifilippo, A.; Scarfia, A.; Martino, D.; Marchisotta, S. New Treatment in Advanced Thyroid Cancer. *J. Oncol.* **2012**, *2012*, 391629. [CrossRef]
5. Mazzaferri, E.; Robbins, R.; Spencer, C.; Braverman, L.; Pacini, F.; Wartofsky, L.; Haugen, B.; Sherman, S.; Cooper, D.; Braunstein, G.; et al. A Consensus Report of the Role of Serum Thyroglobulin as a Monitoring Method for Low-Risk Patients with Papillary Thyroid Carcinoma. *J. Clin. Endocrinol. Metab.* **2003**, *88*, 1433–1441. [CrossRef]
6. Santhanam, P.; Ladenson, P. Surveillance for Differentiated Thyroid Cancer Recurrence. *Endocrinol. Metab. Clin. N. Am.* **2019**, *48*, 239–252. [CrossRef]
7. Pandya, A.; Caoili, E.; Jawad-Makki, F.; Wasnik, A.; Shankar, P.R.; Bude, R.; Haymart, M.; Davenport, M. *Limitations of the* 2015 ATA Guidelines for Prediction of Thyroid Cancer: A Review of 1947 Consecutive Aspirations. *J. Clin. Endocrinol. Metab.* **2018**, *103*, 3496–3502. [CrossRef]
8. Toraih, E.; Fawzy, M.; Hussein, M.; EL-Labban, M.; Ruiz, E.M.L.; Attia, A.-E.-A.; Halat, S.; Moroz, K.; Errami, Y.; Zerfaoui, M.; et al. MicroRNA-Based Risk Score for Predicting Tumor Progression Following Radioactive Iodine Ablation in Well-Differentiated Thyroid Cancer Patients: A Propensity-Score Matched Analysis. *Cancers* **2021**, *13*, 4649. [CrossRef]
9. Tuttle, R.; Tala, H.; Shah, J.; Leboeuf, R.; Ghossein, R.; Gonen, M.; Brokhin, M.; Omry, G.; Fagin, J.; Shaha, A. 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. Off. J. Am. Thyroid. Assoc.* **2010**, *20*, 1341–1349. [CrossRef]
10. Li, L.-R.; Du, B.; Liu, H.-Q.; Chen, C. Artificial Intelligence for Personalized Medicine in Thyroid Cancer: Current Status and Future Perspectives. *Front. Oncol.* **2021**, *10*, 604051. [CrossRef]
11. Paul, R.; Juliano, A.; Faquin, W.; Chan, A.W. An Artificial Intelligence Ultrasound Platform for Screening and Staging of Thyroid Cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **2022**, *112*, e8. [CrossRef]
12. Nagendra, L.; Pappachan, J.M.; Fernandez, C.J. Artificial Intelligence in the Diagnosis of Thyroid Cancer: Recent Advances and Future Directions. *Artif. Intell. Cancer* **2023**, *4*, 1–10. [CrossRef]

13. Habchi, Y.; Himeur, Y.; Kheddar, H.; Boukabou, A.; Atalla, S.; Chouchane, A.; Ouamane, A.; Mansoor, W. AI in Thyroid Cancer Diagnosis: Techniques, Trends, and Future Directions. *Systems* **2023**, *11*, 519. [CrossRef]
14. Ahn, J.; Lee, M.-C. Application of Artificial Intelligence to Evaluate Thyroid Nodules. *J. Clin. Otolaryngol. Head Neck Surg.* **2023**, *34*, 17–22. [CrossRef]
15. Cao, C.-L.; Li, Q.; Tong, J.; Shi, L.; Li, W.-X.; Xu, Y.; Cheng, J.; Du, T.-T.; Li, J.; Cui, X. Artificial Intelligence in Thyroid Ultrasound. *Front. Oncol.* **2023**, *13*, 1060702. [CrossRef]
16. Kim, S.Y.; Kim, Y.I.; Kim, H.J.; Chang, H.; Kim, S.M.; Lee, Y.S.; Kwon, S.S.; Shin, H.; Chang, H.S.; Park, C.S. New approach of prediction of recurrence of thyroid cancer patients using machine learning. *Medicine* **2021**, *100*, e27493. [CrossRef]
17. Grani, G.; Gentili, M.; Siciliano, F.; Albano, D.; Zilioli, V.; Morelli, S.; Puxeddu, E.; Zatelli, M.C.; Gagliardi, I.; Piovesan, A.; et al. A data-driven approach to refine predictions of differentiated thyroid cancer outcomes: A prospective multicenter study. *J. Clin. Endocrinol. Metab.* **2023**, *108*, 1921–1928. [CrossRef] [PubMed]
18. Clark, E.; Price, S.; Lucena, T.; Haberlein, B.; Wahbeh, A.; Seetan, R. Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach. *Knowledge* **2024**, *4*, 557–570. [CrossRef]
19. Park, Y.M.; Lee, B.-J. Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence. *Sci. Rep.* **2021**, *11*, 4948. [CrossRef]
20. Firat Atay, F.; Yagin, F.H.; Colak, C.; Elkiran, E.T.; Mansuri, N.; Ahmad, F.; Ardigò, L.P. A hybrid machine learning model combining association rule mining and classification algorithmsto predict differentiated thyroid cancer recurrence. *Front. Med.* **2024**, *11*, 1461372. [CrossRef] [PubMed]
21. Kim, G.H.; Lee, D.H.; Choi, J.W.; Jeon, H.J.; Park, S. Mulitmodal Neural Network for Recurrence Prediction of Papillary Thyroid Carcinoma. *Adv. Intell. Syst.* **2023**, *5*, 2200365. [CrossRef]
22. Arslan, A.K.; Colak, C. Explainable Machine Learning Models for Prediting Recurrence in Differentiated Thyroid Cancer. *Med. Rec.* **2024**, *6*, 468–473. [CrossRef]
23. Gurcan, F.; Soylu, A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers* **2024**, *16*, 3417. [CrossRef]
24. Thomas, J. Predicting Differentiated Thyroid Cancer Outcomes Using Machine Learning: A Move toward Precision Medicine. *Clin. Thyroidol.* **2024**, *36*, 64–66. [CrossRef]
25. Gu, J.; Xie, R.; Zhao, Y.; Zhao, Z.; Xu, D.; Ding, M.; Lin, T.; Xu, W.; Nie, Z.; Miao, E.; et al. A machine learning-based approach to predicting the malignant and metastasis of thyroid cancer. *Front. Oncol.* **2022**, *12*, 938292. [CrossRef]
26. Mourad, M.; Moubayed, S.; Dezube, A.; Mourad, Y.; Park, K.; Torreblanca-Zanca, A.; Torrecilla, J.S.; Cancilla, J.C.; Wang, J. Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. *Sci. Rep.* **2020**, *10*, 5176. [CrossRef] [PubMed]
27. Setiawan, K.E. Predicting recurrence in differentiated thyroid cancer: A comparative analysis of various machine learning models including ensemble methods with chi-squared feature selection, Commun. *Math. Biol. Neurosci.* **2024**, *2024*, 55.
28. Sibarani, I.J.B.; Loy, K.M.; Surharjito, S. Enhancing Predictive Accuracy for Differentiated Thyroid Cancer (DTC) Recurrence Through Advanced Data Mining Techniques. *TIN Terap. Inform. Nusant.* **2024**, *5*, 11–22. [CrossRef]
29. Xi, N.M.; Wang, L.; Yang, C. Improving the diagnosis of thyroid cancer by machine learning and clinical data. *Sci. Reports.* **2022**, *12*, 11143. [CrossRef]
30. Ahmad, M.A.; Haddad, J. An Explainable AI Model for Predicting the Recurrence of Differentiated Thyroid Cancer. In Proceedings of the 2024 Second Jordanian International Biomedical Engineering Conference (JIBEC), Amman, Jordan, 27–28 November 2024.
31. Bharath, K.; Sai Sabatha, A. Predicting Recurrence in Differentiated Thyroid Cancer: A Machine Learning Approach. In Proceedings of the International Conference on Advances in Data Engineering and Intelligent Computing Systems, Chennai, India, 18–19 April 2024.
32. Arvidsson, J. Differentiated Thyroid Cancer Recurrence. 2024. Available online: https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence (accessed on 7 May 2025).
33. Stfxecutables Df-Analyze: AutoML Command-Line Tool 2024. Available online: https://github.com/stfxecutables/df-analyze (accessed on 7 May 2025).
34. Levman, J.; Jennings, M.; Rouse, E.; Berger, D.; Kabaria, P.; Nangaku, M.; Gondra, I.; Takahashi, E. A Morphological Study of Schizophrenia with Magnetic Resonance Imaging, Advanced Analytics, and Machine Learning. *Front. Neurosci.* **2022**, *16*, 926426. [CrossRef] [PubMed]
35. Saville, K.; Berger, D.; Levman, J. Mitigating Bias Due to Race and Gender in Machine Learning Predictions of Traffic Stop Outcomes. *Information* **2024**, *15*, 687. [CrossRef]
36. Figueroa, J.; Etim, P.; Shibu, A.; Berger, D.; Levman, J. Diagnosing and Characterizing Chronic Kidney Disease with Machine Learning: The Value of Clinical Patient Characteristics as Evidenced from an Open Dataset. *Electronics* **2024**, *13*, 4326. [CrossRef]
37. Huang, X.; Gauthier, C.; Berger, D.; Cai, H.; Levman, J. Identifying Cortical Molecular Biomarkers Potentially Associated with Learning in Mice Using Artificial Intelligence. *Int. J. Mol. Sci.* **2025**, *26*, 6878. [CrossRef] [PubMed]

38. Kendall, J.; Gaspar, G.; Berger, D.; Levman, J. Machine Learning and Feature Selection in Pediatric Appendicitis. *Tomography* **2025**, *11*, 90. [CrossRef]
39. Joseph, M.; Raj, H. GANDALF: Gated adaptive network for deep automated learning of features. *arXiv* **2022**, arXiv:2207.08548.
40. Berger, D. Redundancy-Aware Feature Selection. Available online: https://github.com/stfxecutables/df-analyze/tree/experimental?tab=readme-ov-file#redundancy-aware-feature-selection-new (accessed on 23 December 2024).