Machine Learning for Risk Stratification in Thyroid Cancer Patients

Cole Khamnei (cck2139) and Shreekrishna Rout (sr4196)

Outline:

1. Introduction

Thyroid cancer, while often associated with lower mortality rates than other solid organ cancers, carries a significant risk of recurrence. Specifically, well-differentiated subtypes of thyroid cancer such as papillary and follicular carcinomas have a high risk of recurrence [1]. Accurately estimating the risk of recurrence and mortality is a common goal for most malignancies, as accurate risk prediction can better guide treatment regimens, follow-up care, and earlier and more specialized palliative care involvement. These benefits of accurate risk assessment tools

can both extend the duration of life and improve the quality of life, which are fundamental goals of medicine.

Many modern risk assessment tools have utilized machine learning-based approaches to improve prediction accuracy over traditional statistical methods. As we know from lectures, different machine learning models and methods can better model complex nonlinear relationships between features. For thyroid cancer patients, machine learning can provide a quantitative approach that complements or improves upon pre-existing guidelines such as American Thyroid Association (ATA) risk stratification systems. Medical association guidelines are often based on simple statistical models such as "point systems" where different features are assigned point values and the sum of point values for a patient's given feature set is associated with a given risk (similar to logistic regression). These point models are often highly interpretable and readily calculable which makes them easy for physicians and patients, however, machine learning models can potentially provide accuracy improvements. Additionally, machine learning approaches can discover previously unknown risk factors and feature structures, such as subgroups, which can be iteratively explored through further research.

The chosen paper explores the application of various machine learning models for predicting the risk of thyroid cancer recurrence. The models used in the paper include SVM, K-nearest neighbors classifier, decision trees, random forest, and ANN. We reimplemented the referenced paper, aiming to reproduce their results, and then build on their work. To extend the authors' work, we trained logistic regression and XGBoost models to attempt to improve predicted risk accuracy.

2. Dataset and Related Work

2.1 Dataset Description

The dataset from the initial paper consists of 383 thyroid cancer patients from a 15-year cohort study to better understand thyroid recurrence risk factors. The enrolled subjects were followed for a minimum of 10 years, which enabled an assessment of recurrence. Errors in assessing recurrence were minimized by choosing a large 10-year window, during which the majority of recurrence cases will happen. Information collected in this includes demographics, clinical risk features, pathology features, and treatment-related factors. Demographic variables included age at diagnosis (continuous feature) and sex (binary feature). Clinical history features included smoking status (binary feature) and a previous history of therapeutic or environmental head or neck radiation exposure history (binary feature).

Pathology-related features included thyroid cancer subtypes (categorical: papillary, micropapillary, follicular, Hurthle cell carcinoma), focality (binary: unifocal or multifocal), and TNM staging. TNM staging is a common cancer staging system that varies for each cancer type and subtype and is an assessment of the Tumor size (T), lymph node involvement (N), and metastasis (M). Treatment-specific features in the study included the treatment response (categorical: excellent, indeterminate, structurally incomplete, and biochemically incomplete responses). The ATA risk level (low, intermediate, or high risk) of different subjects was included as a feature, and the recurrence status of thyroid cancer for the subjects was the binary outcome to be studied.

The comprehensive nature of this study and resulting data is a significant advantage over previous studies as it contains both clinical, pathological, and demographic features. Cancer recurrence is multifactorial and can be challenging to predict, so including both clinical and pathological features was shown to improve risk prediction over previous methods. A key limitation of this study is the small sample size, especially considering the heterogeneity of the subtypes studied. The small sample size also restricts the use of more complex models, such as neural networks, due to problems with overfitting.

2.2 Paper Results and Tools

The referenced machine learning paper includes classic models such as K-nearest neighbors (KNN), support vector machines (SVM), decision trees, random forests, and artificial neural networks (ANNs).

The SVM had the highest ROC area under the curve (AUC) of 0.9971 of all models with a sensitivity of 93.33% and specificity of 97.14%. The random forest and ANN models also performed very well with comparable performances, RF AUC was 0.9938 and ANN had an AUC of 0.9964. The KNN and decision tree models were comparative underperformers, but still effective. The high AUCs in the paper suggest the models may utilize all or most of the possible decisive information contained in the features for recurrence prediction. Additionally, the referenced paper demonstrated that the features collected in the study can outperform the ATA risk factor by itself.

For specific preprocessing tools, we followed the data preprocessing techniques outlined in the paper for continuous and categorical variables. Continuous variables were standard scaled and categorical variables were one-hot encoded.

3. Reproducing Results

3.1 Methodology

We took many steps in our approach in order to reproduce the results from the paper. We initially needed to scale the continuous features using the StandardScaler object while also performing one-hot encoding on the categorical variables in order to create binary values. We also split the dataset into a training dataset and a validation dataset where the training dataset contained 283 samples while the validation dataset contained 100 samples. We also followed the paper in terms of training the same models of K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, and Artificial Neural Networks by implementing the Scikit-learn library. Moreover, we optimized the hyperparameters for each model using the GridSearchCV method that consisted of five-fold cross-validation. In particular, we made evaluations on the trained model using the metrics of AUC (area under the curve), accuracy, sensitivity, and specificity. Based on this approach, we were able to follow the same steps that closely aligned with the methodology of the original study. This allowed us to easily compare our results with the original study results. It was important to use cross-validation since we wanted to minimize overfitting by having more groups associated with training the model. We also opted to use the Scikit-learn library since this library provides thorough documentation, implementation consistency, and commonly accepted machine learning techniques.

3.2 Results

After running the models from the paper, we obtained results that were fairly consistent with the results achieved in the paper. We also ran the models with the feature sets of ATA risk only, ATA Risk excluded, and the full feature set.

The Artificial Neural Network model with ATA risk only had an accuracy of 88.0%, an AUC score of 89.21%, a sensitivity level of 86.20%, and a specificity level of 88.73%. The Artificial Neural Network model without ATA risk had an accuracy of 96.0%, an AUC score of 97.96%, a sensitivity level of 93.10%, and a specificity level of 97.18%. The Artificial Neural Network model with a full feature set had an accuracy of 95.0%, an AUC score of 98.49%, a sensitivity level of 89.65%, and a specificity level of 97.18%.

The Decision Tree model with ATA risk only had an accuracy of 88.0%, an AUC score of 89.21%, a sensitivity level of 86.20%, and a specificity level of 88.73%. The Decision Tree model without ATA risk had an accuracy of 94.0%, an AUC score of 93.54%, a sensitivity level of 89.65%, and a specificity level of 95.77%. The Decision Tree model with a full feature set had an accuracy of 93.0%, an AUC score of 91.64%, a sensitivity level of 82.75%, and a specificity level of 97.18%.

The K-Nearest Neighbors model with ATA risk only had an accuracy of 88.0%, an AUC score of 89.21%, a sensitivity level of 86.20%, and a specificity level of 88.73%. The K-Nearest Neighbors model without ATA risk had an accuracy of 94.0%, an AUC score of 95.04%, a sensitivity level of 82.75%, and a specificity level of 98.59%. The K-Nearest Neighbors model

with a full feature set had an accuracy of 94.0%, an AUC score of 95.07%, a sensitivity level of 86.20%, and a specificity level of 97.18%.

The Random Forest model with ATA risk only had an accuracy of 88.0%, an AUC score of 89.21%, a sensitivity level of 86.20%, and a specificity level of 88.73%. The Random Forest model without ATA risk had an accuracy of 97.0%, an AUC score of 99.17%, a sensitivity level of 93.10%, and a specificity level of 98.59%. The Random Forest model with a full feature set had an accuracy of 95.0%, an AUC score of 98.73%, a sensitivity level of 89.65%, and a specificity level of 97.18%.

The Support Vector Machine model with ATA risk only had an accuracy of 88.0%, an AUC score of 89.21%, a sensitivity level of 86.20%, and a specificity level of 88.73%. The Support Vector Machine model without ATA risk had an accuracy of 96.0%, an AUC score of 98.73%, a sensitivity level of 89.65%, and a specificity level of 98.59%. The Support Vector Machine model with a full feature set had an accuracy of 93.0%, an AUC score of 98.00%, a sensitivity level of 86.20%, and a specificity level of 95.77%.

We can see that the above results indicate that these models were effective in predicting the recurrence of thyroid cancer in patients. We also found that models using multidimensional data including features like treatment response and TNM staging result in the models making more accurate predictions.

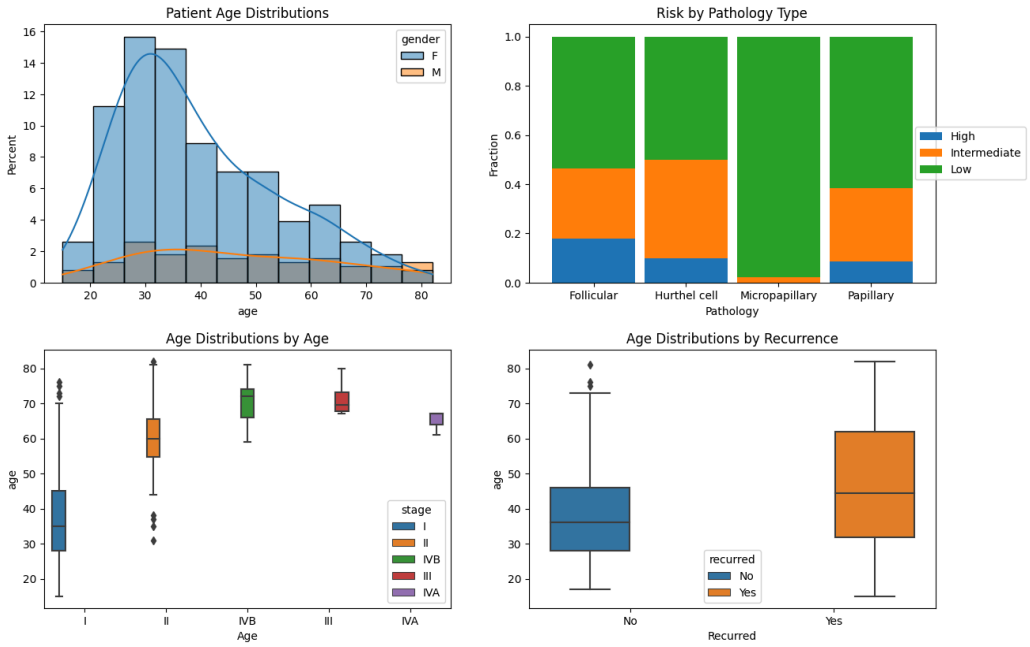

Figure 1. Patient Demographic Plots

Table 1: Training Split Demographics:

Features		Train	Validation	Total
Gender	F	224 (79%)	88 (88%)	312 (81%)
	M	59 (21%)	12 (12%)	71 (19%)
Smoking	No	245 (87%)	89 (89%)	334 (87%)
	Yes	38 (13%)	11 (11%)	49 (13%)
History Of Smoking	No	259 (92%)	96 (96%)	355 (93%)
	Yes	24 (8%)	4 (4%)	28 (7%)
History Of Radiotherapy	No	277 (98%)	99 (99%)	376 (98%)
	Yes	6 (2%)	1 (1%)	7 (2%)
Thyroid Function	Euthyroid	245 (87%)	87 (87%)	332 (87%)
	Clinical Hyperthyroidism	17 (6%)	3 (3%)	20 (5%)
	Clinical Hypothyroidism	6 (2%)	6 (6%)	12 (3%)
	Subclinical Hyperthyroidism	4 (1%)	1 (1%)	5 (1%)
	Subclinical Hypothyroidism	11 (4%)	3 (3%)	14 (4%)
Physical Examination	Single nodular goiter-left	64 (23%)	25 (25%)	89 (23%)
	Multinodular goiter	105 (37%)	35 (35%)	140 (37%)
	Single nodular goiter-right	106 (37%)	34 (34%)	140 (37%)
	Normal	4 (1%)	3 (3%)	7 (2%)
	Diffuse goiter	4 (1%)	3 (3%)	7 (2%)
Adenopathy	No	206 (73%)	71 (71%)	277 (72%)
	Right	36 (13%)	12 (12%)	48 (13%)
	Extensive	6 (2%)	1 (1%)	7 (2%)
	Left	10 (4%)	7 (7%)	17 (4%)
	Bilateral	23 (8%)	9 (9%)	32 (8%)
	Posterior	2 (1%)	0 (0%)	2 (1%)
Pathology	Micropapillary	30 (11%)	18 (18%)	48 (13%)
	Papillary	217 (77%)	70 (70%)	287 (75%)
	Follicular	23 (8%)	5 (5%)	28 (7%)
	Hurthel cell	13 (5%)	7 (7%)	20 (5%)

Features		Train	Validation	Total
Focality	Uni-Focal	178 (63%)	69 (69%)	247 (64%)
	Multi-Focal	105 (37%)	31 (31%)	136 (36%)
Risk	Low	182 (64%)	67 (67%)	249 (65%)
	Intermediate	78 (28%)	24 (24%)	102 (27%)
	High	23 (8%)	9 (9%)	32 (8%)
T	T1a	31 (11%)	18 (18%)	49 (13%)
	T1b	29 (10%)	14 (14%)	43 (11%)
	T2	115 (41%)	36 (36%)	151 (39%)
	T3a	75 (27%)	21 (21%)	96 (25%)
	T3b	12 (4%)	4 (4%)	16 (4%)
	T4a	14 (5%)	6 (6%)	20 (5%)
	T4b	7 (2%)	1 (1%)	8 (2%)
N	N0	197 (70%)	71 (71%)	268 (70%)
	N1b	70 (25%)	23 (23%)	93 (24%)
	N1a	16 (6%)	6 (6%)	22 (6%)
M	M0	272 (96%)	93 (93%)	365 (95%)
	M1	11 (4%)	7 (7%)	18 (5%)
Stage	I	244 (86%)	89 (89%)	333 (87%)
	II	26 (9%)	6 (6%)	32 (8%)
	IVB	7 (2%)	4 (4%)	11 (3%)
	III	3 (1%)	1 (1%)	4 (1%)
	IVA	3 (1%)	0 (0%)	3 (1%)
Response	Indeterminate	46 (16%)	15 (15%)	61 (16%)
	Excellent	152 (54%)	56 (56%)	208 (54%)
	Structural Incomplete	65 (23%)	26 (26%)	91 (24%)
	Biochemical Incomplete	20 (7%)	3 (3%)	23 (6%)
Recurred	No	204 (72%)	71 (71%)	275 (72%)
	Yes	79 (28%)	29 (29%)	108 (28%)

4. Proposed Techniques

We also decided to test additional machine learning algorithms in order to add more results to the original research paper. We opted to use the models of logistic regression and AdaBoost as additional models. The logistic regression model calculates the weighted score for each of the values before applying the sigmoid function in order to sort the predictions into various class predictions. The AdaBoost model is an algorithm that forms an ensemble of many weak classifiers and can be compared to the Random Forest model and Decision Tree model. We wanted to use the AdaBoost model because it can be important for feature importance selection.

4.1 Evaluation Plan

We will evaluate each model using the same metrics included in the original study such as AUC, accuracy, sensitivity, and specificity. We will generally focus on looking at the strengths of each model by handling categorical and numerical features and accounting for the efficiency of computations. We will also focus on the time spent related to computational efficiency since we want to determine the reasonable possibility of running these algorithms in clinical settings so that doctors and healthcare professionals can make decisions in real-time settings.

Table 2: Implemented Models from Paper:

	Feature Set	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy
ANN	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	93.10%	97.18%	93.10%	97.18%	97.96%	96.0%
	Full	89.65%	97.18%	92.85%	95.83%	98.49%	95.0%
DecisionTree	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	89.65%	95.77%	89.65%	95.77%	93.54%	94.0%
	Full	82.75%	97.18%	92.30%	93.24%	91.64%	93.0%
KNN	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	82.75%	98.59%	96.0%	93.33%	95.04%	94.0%
	Full	86.20%	97.18%	92.59%	94.52%	95.07%	94.0%
RandomForest	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	93.10%	98.59%	96.42%	97.22%	99.17%	97.0%
	Full	89.65%	97.18%	92.85%	95.83%	98.73%	95.0%
SVM	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	89.65%	98.59%	96.29%	95.89%	98.73%	96.0%
	Full	86.20%	95.77%	89.28%	94.44%	98.00%	93.0%

Table 3: New Models Evaluation Results:

	Feature Set	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy
AdaBoost	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	93.10%	94.36%	87.09%	97.10%	97.52%	94.0%
	Full	93.10%	94.36%	87.09%	97.10%	97.28%	94.0%
LR	ATA Risk	86.20%	88.73%	75.75%	94.02%	89.21%	88.0%
	ATA Risk Excluded	93.10%	97.18%	93.10%	97.18%	98.30%	96.0%
	Full	93.10%	95.77%	90.0%	97.14%	98.34%	95.0%

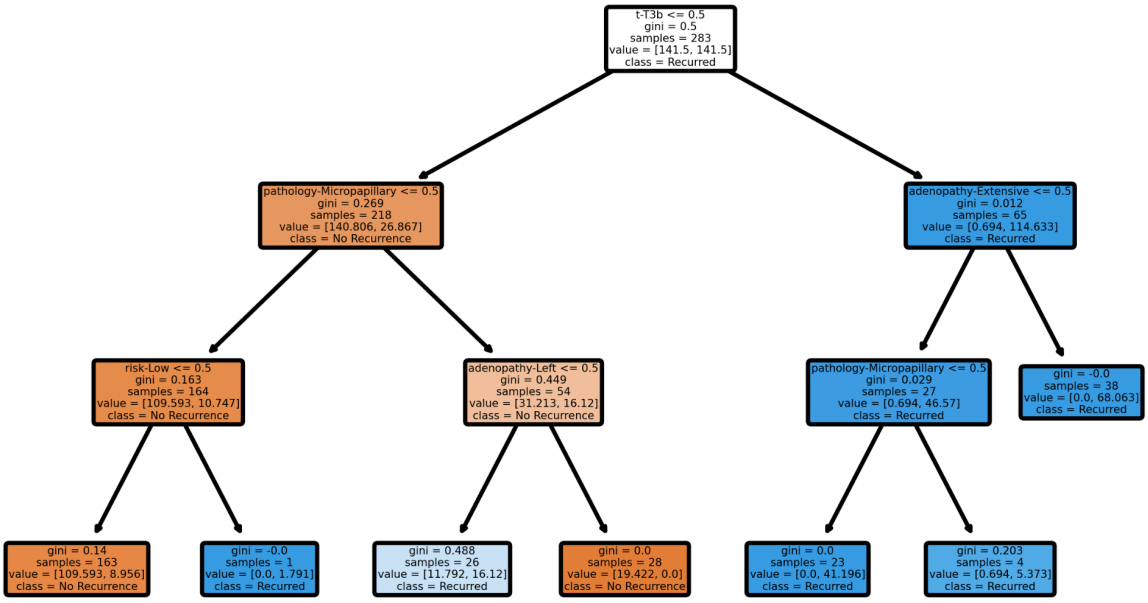

Figure 3: Example Decision Tree Feature Splits

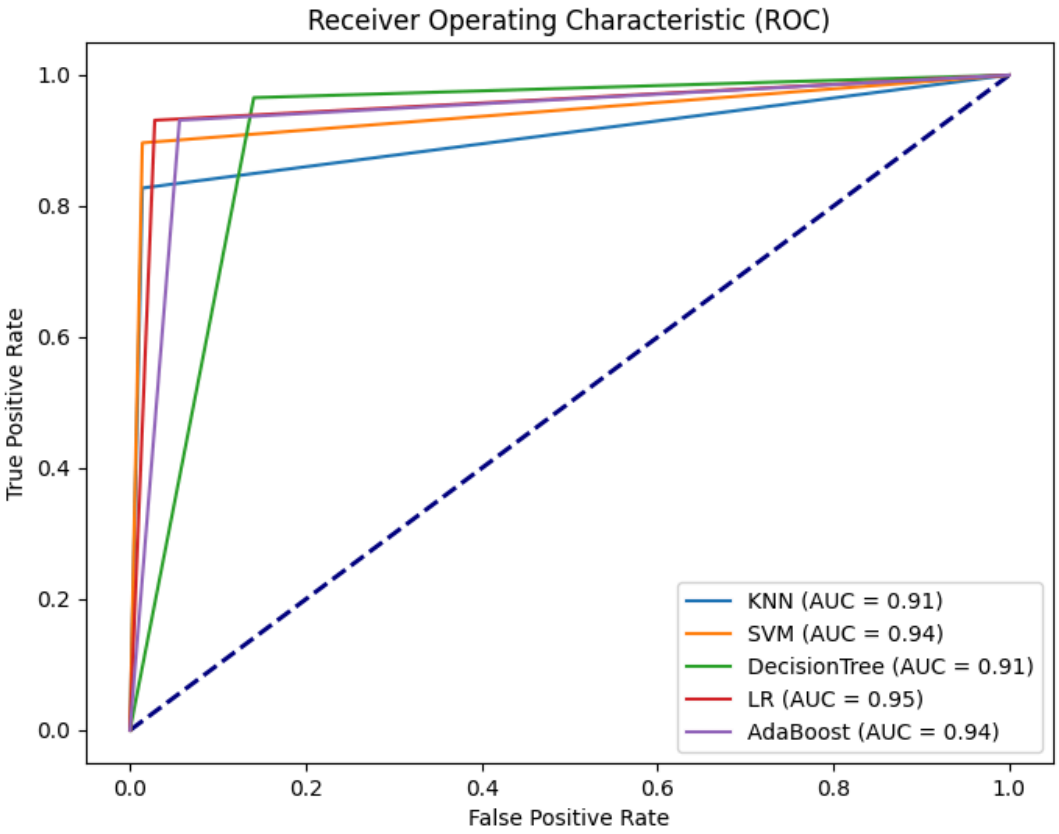

Figure 4: Model ROC Curves

5. Discussion and Conclusion

5.1 Comparative Analysis

We saw that the data preprocessing steps dramatically improved the effectiveness of the models. In particular, we saw that standard scaling and one-hot encoding helped significantly with the data preprocessing procedure. We also saw that the models generally had high accuracy and AUC levels. The original study models of K-nearest neighbors, Support Vector Machine, and Decision Tree had AUC levels of 0.91, 0.94, and 0.91 respectively. In contrast, the new models that we reintroduced had slightly better performance levels where the logistic regression model had an AUC level of 0.95 while the AdaBoost model had an AUC level of 0.94. Overall, many

of the models were able to provide fairly accurate results with respect to the prediction classes of thyroid cancer recurrence in the dataset of patients. This was very important because we wanted to be able to form certain conclusions based on the results obtained from these models so that we could evaluate the effect of certain categorical or continuous variables on the prediction class. We also wanted to be able to highlight the strongest trends by ranking the importance of features according to the different models since their different approaches might prioritize certain features over other features. In general, it was important for us to analyze not only the final results obtained by the wide variety of these models but also properly delve into the specific steps of the methodology that each model took to get to their final results. We wanted to be able to understand whether the steps that these models took were easily replicable and reasonable, especially with respect to other datasets. Finally, we saw that these models proved to be fairly robust in terms of providing calculations and using different groups as the training dataset before undergoing validation and testing.

5.2 Advantages and Limitations

The models used to obtain the results above had many advantages throughout the process. Models like Support Vector Machine, Random Forest, k-Nearest Neighbors, and AdaBoost were very effective in detecting and explaining nonlinear relationships between different features and the prediction classes. These models were also very computationally powerful in integrating diverse features in order to extend components of risk stratification systems. On the other hand, the limitations include a small dataset that results in the models being less general. There are also no external validation datasets, which means that the confidence level of the model is also not that high.

5.3 Future Work

Some possible future work includes validating the models that we developed on outside datasets to make the models more generalizable and robust across different datasets. We would want to make our models more general because we don't want the bias of our current dataset to dramatically skew the prediction results of our models, especially when we want to apply these models to other patient datasets. We can also look into adding genetic markers, imaging data, and other features since that could potentially improve the prediction accuracy of our models. We will also look to explore additional models like XGBoost and Naive Bayes classifier models more in-depth using the same procedural steps as what we took for the other models above. This is especially important because we want to pursue different approaches that will help us to analyze potential performance improvements. Additionally, we can look into combining machine learning models with each other in order to develop efficient decision support systems for clinical practice. In particular, our study can look at using machine learning to analyze risk stratification for thyroid cancer patients, which would result in more personalized medical approaches. By implementing this approach, medical professionals would be able to utilize these studies to look at a more comprehensive method of predicting thyroid cancer rates across a large pool of afflicted patients. This is also especially important in tracking signs of thyroid cancer in patients in order to have a treatment plan in place for these patients.

Works Cited:

1. Haugen, B. R., et al. "2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer." *Thyroid*, vol. 26, no. 1, 2016, pp. 1–133.
2. Powers, A. E., et al. "Changes in Trends in Thyroid Cancer Incidence in the United States, 1992 to 2016." *JAMA*, vol. 322, no. 24, 2019, pp. 2440–2441.
3. Schlumberger, M., & Leboulleux, S. "Current Practice in Patients with Differentiated Thyroid Cancer." *Nature Reviews Endocrinology*, vol. 17, no. 3, 2021, pp. 176–188.
4. Ouyang, F. S., et al. "Comparison Between Linear and Nonlinear Machine-Learning Algorithms for the Classification of Thyroid Nodules." *European Journal of Radiology*, vol. 113, 2019, pp. 251–257.
5. Tuttle, R. M., et al. "Estimating Risk of Recurrence in Differentiated Thyroid Cancer After Total Thyroidectomy and Radioactive Iodine Remnant Ablation." *Thyroid*, vol. 20, no. 12, 2010, pp. 1341–1349.