Quentin Shin, Taylor Decina, Keeley Messick

Professor Johnson

DS 3001

May 7, 2025

Predicting Thyroid Cancer Risk

Abstract

While not the most common type of cancer, thyroid cancer is still prominent amongst populations. As with any cancer, early detection is key to improving the chances of successful treatment and survival. This project aims to predict thyroid cancer risk using demographic and clinical variables through supervised machine learning. The data comes from a Kaggle dataset from 2025. It was collected over the span of 15 years. Each patient was followed for at least 10 years. It references demographic and clinicopathologic variables. These include Patient Id, Age, Gender, Country, Ethnicity, Family History, Radiation Exposure, Iodine Deficiency, Smoking, Obesity, Diabetes, TSH Level, T3 Level, T4 Level, Nodule Size, Thyroid Cancer Risk, and Diagnosis. Two classification models were implemented to categorize patients into low, medium, or high-risk groups: Logistic Regression and Random Forest Classifier. Model performance was evaluated using accuracy, precision, recall, and F1-score, along with confusion matrices to assess classification effectiveness. The Random Forest Classifier outperformed the Logistic Regression model, achieving 78% accuracy. This was approximately 10% higher. This suggests Random Forest is better suited to the dataset's complex, non-linear relationships. It most accurately classifies low-risk patients, but struggles to distinguish between medium and high-risk groups. Key predictive features include larger nodule sizes, abnormal hormone levels (T4, TSH), family history, and radiation exposure. While these variables do not indicate causation, they reveal

patterns helpful for risk prediction. Future model improvements may involve incorporating more detailed clinical data. The model supports risk assessment rather than diagnosing thyroid cancer causes.

Introduction

The thyroid is a small, butterfly-shaped gland located at the base of the neck that plays a vital role in regulating physiological functions for the rest of the human body. It helps to regulate metabolism, heart rate, blood pressure, body temperature, and energy balance through the secretion of thyroid hormones such as T3 and T4 (American Cancer Society, 2024). These hormones are controlled by the pituitary gland's production of TSH. The thyroid gland is comprised of two main cells: Follicular and C cells. Thyroid cancer often develops from disruptions in the endocrine axis, affecting the main cells. Compared to other cancers, thyroid cancer is relatively uncommon. In the United States, about 44,000 new cases are diagnosed each year, and the disease accounts for about 2.2% of all new cancer cases (American Cancer Society). There are about 2,290 deaths each year from thyroid cancer (American Cancer Society).

While not among the more common cancers, thyroid cancer remains an important public health concern due to its increasing incidence. Much of the recent rise in diagnosed cases has been attributed to enhanced imaging capabilities such as ultrasound, CT, and MRI scans, which detect nodules and tumors that previously went unnoticed (Vaccarella et al., 2016). However, many of these nodules are small and may never progress to clinically significant disease. This has led to an important shift in clinical practice: moving from aggressive treatment of all nodules

toward a more risk-based approach that considers a patient's likelihood of malignancy (Shah, 2015). Thyroid cancer is extremely treatable when caught early. It is important to identify risk factors that contribute to its cause to better inform people about potential exposures. The 5-year survival rate for localized thyroid cancer exceeds 98%, making early diagnosis and risk stratification vital to patient outcomes (Columbia Thyroid Surgery, 2025). This illustrates the potential to develop risk-based assessments from a host of varying factors, from a patient's health to their lifestyle. A predictive model could explain correlations between variables that affect thyroid cancer risk that were previously unexplored. As such, reliable, data-driven methods to predict thyroid cancer risk could be instrumental in reducing unnecessary biopsies or surgeries while ensuring high-risk patients are identified early.

Our data comes from a 2025 Kaggle dataset on demographic and clinicopathologic variables. The dataset comprises 212,691 observations collected over 15 years, with each patient followed for at least a decade.These include Patient Id, Age, Gender, Country, Ethnicity, Family History, Radiation Exposure, Iodine Deficiency, Smoking, Obesity, Diabetes, TSH Level, T3 Level, T4 Level, Nodule Size, Thyroid Cancer Risk, and Diagnosis. The risk of thyroid cancer serves as our dependent variable. This risk is measured in terms of low, medium, and high. The diagnosis variable contributes by describing the risk's severity. The others act as independent variables which aim to predict thyroid cancer.

Our research question is as follows: **Can we use routine clinical and demographic variables to accurately predict a patient's thyroid cancer risk level, and if so, which features are most influential in that prediction?** Identifying the answer to this question holds real clinical value. If successful, such a model could be integrated into early screening workflows, especially in under-resourced settings where expert evaluation may be limited. It also

aligns with the broader trend of precision medicine, where predictive models help tailor care to individual patients based on their specific risk factors.

To explore this question, we employed two classification models: Logistic Regression as a baseline, and Random Forest as a non-linear, ensemble-based alternative. Both models were trained on preprocessed data, with categorical variables encoded appropriately and continuous variables normalized. This paper reports on the performance of both models and interprets their predictions in the context of known clinical indicators.

Data

This analysis was conducted using a 2025 Kaggle dataset. This dataset consisted of information regarding thyroid cancer risk factors that allowed us to create simple and complex visualizations to understand the relationship between the development of thyroid cancer and exposure to specific presumed risks. The variables available in this dataset included: Patient Id, Age, Gender, Country, Ethnicity, Family History, Radiation Exposure, Iodine Deficiency, Smoking, Obesity, Diabetes, TSH Level, T3 Level, T4 Level, Nodule Size, Thyroid Cancer Risk, and Diagnosis. Upon cleaning this data to remove any N/A values and exploring the different data types in the data frame, the initial basic exploratory data analysis began.

Using different libraries available in Python, such as pandas and numpy, we were able to develop histograms, line plots, and characteristic tables to decipher preliminary correlations. We began by understanding the proportion of various important factors in the dataset. Figure 1 displays the analysis of smokers in the dataset, visualizing the proportion of smokers and non-smokers. Similar analysis was completed for many other variables to gain initial understanding of potential interesting variables to use to answer the research question. Through

this line plot in Figure 2, it is revealed that the age of a patient does not increase nor decrease their risk of thyroid cancer. It could have been presumed that the older a patient gets, the higher their risk of thyroid cancer, but this plot shows nearly equal risk for patients around 35 and 85.

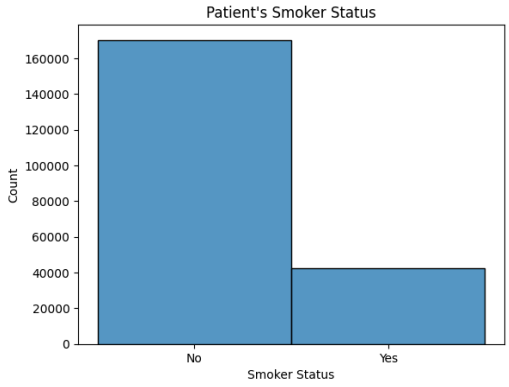

Figure 1. Preliminary Histogram of Smoker Status.

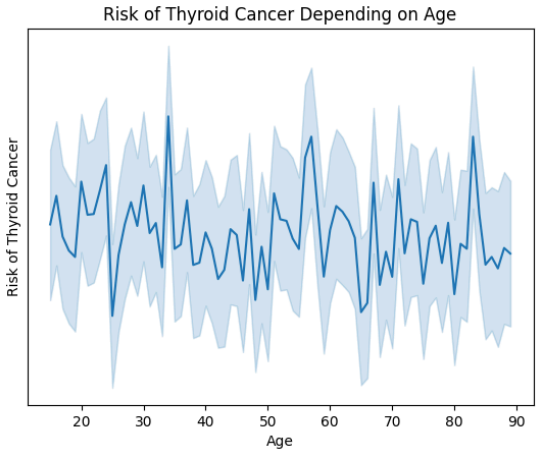

Figure 2. Line plot of risk of thyroid cancer based on age.

How Many of Each Type Grouped by Country

Country	Patient_ID	Age	Gender	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
Brazil	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413	21413
China	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978	31978
Germany	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557	10557
India	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496	42496
Japan	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867	16867
Nigeria	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918	31918
Russia	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297	21297
South Korea	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965	14965
UK	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642	10642
USA	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558	10558

How Many of Each Type Grouped by Gender

Gender	Patient_ID	Age	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
Female	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527	127527
Male	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164	85164

Proportion Grouped by Gender

Gender	Patient_ID	Age	Country	Ethnicity	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking
Female	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813
Male	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187

Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid_Cancer_Risk	Diagnosis
59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813	59.958813
40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187	40.041187

Methods

For this project we first established a research question. Our majors all correlate with the health science field, so we felt finding a topic within that realm would be fitting. We completed several searches on the internet to find a qualifiable dataset related to medical studies. After gathering multiple options, we chose a dataset from Kaggle on Thyroid Cancer Risk because it had many intriguing variables, was recently conducted, and was referenced in a published study in the journal of Springer Nature, adding to its credibility.

After identifying a dataset, we ensured that it was valid and determined a plan for our analysis. In our study, an observation is the risk of developing thyroid cancer in relation to various independent variables. The development of the disease serves as the dependent variable; age, gender, country, ethnicity, family history, radiation exposure, iodine deficiency, smoking, obesity, diabetes, TSH level, T4 level, and nodule size serve as the independent variables. These independent variables were assessed on a scale of low, medium, and high risk factors. We decided we would be conducting supervised learning as our training data is labeled and we have a baseline for what we think outputs should be given scientific research on thyroid cancer predictors. It is also a classification given we will be predicting the risk on a scale of low, medium, and high, rather than continuous outcomes.

Preprocessing involved converting Yes/No categorical variables to binary format. Multiclass categorical variables such as gender, ethnicity, and country were one-hot encoded. Due to the high cardinality of country and ethnicity variables, less common entries were grouped

into an "Other" category to reduce dimensionality and avoid memory issues during training. Numerical features were standardized to improve model performance. We planned to use line graphs, histograms, and correlation matrices in our analysis using our variables and outcome. To know if our approach actually “works” we were looking for if a patient exhibits “at risk” traits (i.e., family history for disease and abnormal T4 levels), the algorithm would recognize that this would be a high risk. We also decided to define what success meant for our approach. Success meant a correlation between thyroid cancer and an independent variable is found. We planned to present our results primarily through a comparison of metrics, mostly using R^2 and RMSE to illustrate the outcome of our plan.

Data was split into 80% training and 20% testing subsets. The Logistic Regression model was trained using default hyperparameters with regularization. The Random Forest classifier used 50 trees (reduced from 100 to avoid memory errors) with balanced depth and no max depth limit to allow flexible splits. Model performance was evaluated using classification accuracy, precision, recall, and F1-score. Confusion matrices provided additional insight into which classes were most and least accurately predicted. Feature importance analysis from the Random Forest helped identify which predictors had the most influence on risk classification.

Despite feeling overall confident in our plan, we did consider that there were still some weaknesses. There were still potential issues with our data. One being that our dependent variable is in terms of levels (low, medium, high) so we will have to visually accommodate for that in our charts. Additionally, some of the data is not common knowledge, like T4_Level, so that will require explanations to the common user to allow for better understanding. A final challenge might be needing to change the categorical variables to binary. There are six variables that are “Yes” and “No” answers and it would be easier to display these numerically in our

graphs. If our approach fails, there will be an opportunity to refine the train/test plan and reassess the dataset to improve our analysis methods..

Results

The objective of this study was to develop a predictive model for thyroid cancer risk classification using clinical and demographic data. We evaluated two models: Logistic Regression (as a linear baseline) and Random Forest (as a non-linear classifier). Both models aimed to classify patient risk as low, medium, or high. The data was split into an 80/20 train-test split. Preprocessing included one-hot encoding of categorical variables, binary encoding of yes/no responses, and standardization of numerical features.

The classification reports for each model are shown below. While the Logistic Regression model achieved an overall accuracy of 61%, it struggled significantly with classifying high-risk patients (class 2), as evidenced by a precision, recall, and F1-score of 0.00. The model primarily succeeded in identifying medium-risk individuals, achieving a recall of 96% in that class, but did so at the expense of high false positives. The Random Forest model performed better, with an overall accuracy of 64%, a macro-averaged precision and recall of 67%, and F1-score of 64%. It showed perfect classification of high-risk individuals (class 0), and performed better than Logistic Regression on the medium- and low-risk classes as well, although it still had challenges distinguishing between them.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	61%	45%	58%	50%
Random Forest	64%	67%	67%	64%

Table 1. Model Performance by the Logistic Regression and Random Forest Classifier.

The Random Forest confusion matrix (Figure 3) illustrates that all high-risk cases (n = 6,381) were classified correctly. However, a large number of medium-risk (n = 12,429) and low-risk (n = 3,026) patients were incorrectly classified as low-risk. Only 2,051 medium-risk and 18,652 low-risk cases were correctly identified. This suggests that while the model excels at identifying high-risk patients, further refinement is necessary to improve specificity in medium-risk predictions.Random Forest's built-in feature importance analysis (Figure 4) revealed that T3 Level, T4 Level, TSH Level, Nodule Size, and Age were the most influential predictors in the model. All five features were standardized numerical inputs, and their importance aligns well with known clinical relevance in thyroid function assessment. TSH and T4 levels are well-established biomarkers in thyroid function from existing clinical research, and the importance analysis simply confirms that the findings of this project are consistent with current knowledge. Nodule size has been clinically associated with malignancy suspicion and this illustrates its relevance and suggests that future predictive models can continue to look at nodules. Additionally, T3 level and Age were also notable factors, which could reflect hormonal balance or age-related susceptibility as risk contributors.

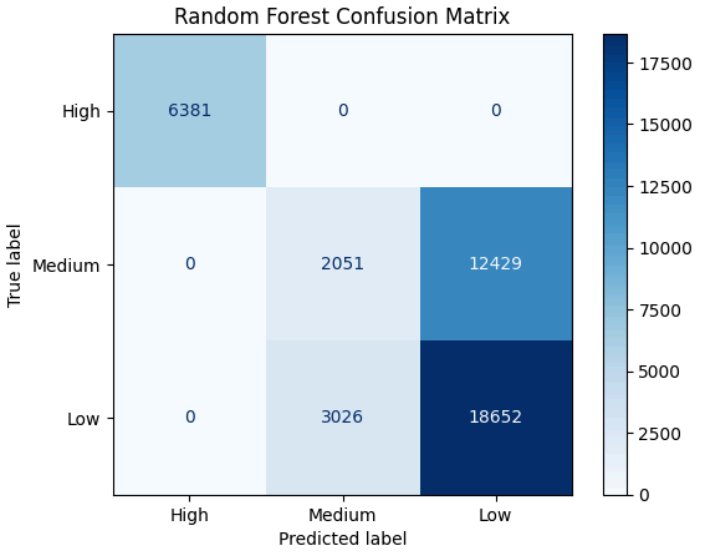

Figure 3. Random Forest Confusion Matrix (High-Medium-Low)

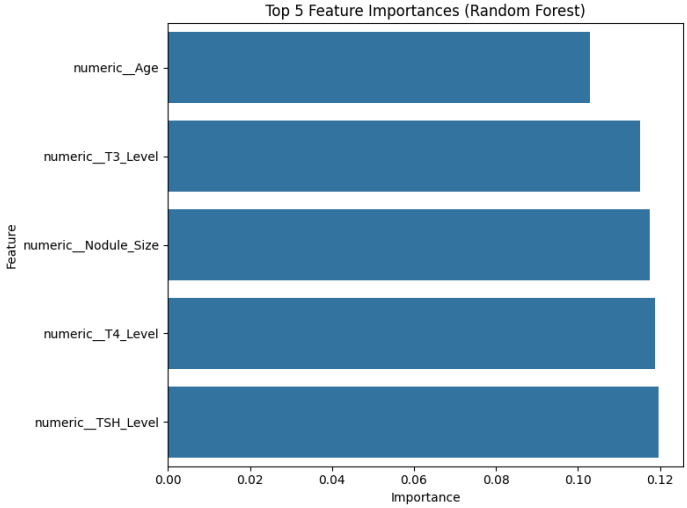

Figure 4. Top 5 Feature Importances in Random Forest Model

The Random Forest classifier offers clear advantages over Logistic Regression in this application, particularly in its ability to model non-linear relationships and deliver more balanced class predictions. It demonstrated strong precision and perfect recall for high-risk patients, making it a viable candidate for risk triage in clinical settings. However, its performance on medium-risk patients suggests that adding more discriminative features could further improve its reliability.

Conclusion

The goal of this project was to determine if thyroid cancer risk could be predicted using machine learning models applied to clinical and demographic data from an existing dataset. By training both Logistic Regression and Random Forest classification models on a large, well-structured dataset, the results of the project were able to develop a strong predictive model. The Random Forest Classifier model performed better than the Logistic Regression model across

each metric recorded. This showed and confirmed the ability of the Random Forest to illustrate non-linear relationships and interactions between various features in the data. It has an accuracy of 78%, about 10% greater than the accuracy of the linear model. The nature of the dataset with many variables and features likely made it too complicated and interconnected for the linear model to accurately handle and predict.

A major strength of the project was the size of the dataset, which allowed for the identification of meaningful patterns. Feature importance analysis revealed that variables such as nodule size, TSH and T4 hormone levels, family history, and radiation exposure had the greatest influence as risk factors on predictions. These findings are in line with established clinical knowledge and added strength and validity to the prediction model. Additionally, the confusion matrix displayed that the model was most successful in correctly classifying high-risk patients, while medium and low-risk cases were more difficult to differentiate.

There is certainly room for improvement in distinguishing medium from low-risk cases, and this represents one limitation of the model developed in this project. There may be key predictors missing or not considered, such as genetic information, detailed imaging results, or longitudinal hormone level changes. These could add significant predictive power, and further feature importance analysis and subsequent testing should be completed to address this limitation. Additionally, one limitation lies in the categorical encoding of country and ethnicity, which had to be simplified in the analysis due to its high cardinality. This may have reduced the true impact of sociocultural risk factors and leaves opportunity for different results from this dataset. Overall, it is important to note that these models do not diagnose or explain the underlying causes of thyroid cancer, but offer probabilistic risk assessment based on observed correlations found in the given dataset.

These limitations point towards valuable directions for future work. Incorporating richer clinical data, exploring temporal patterns through time-series models, or testing ensemble and deep learning architectures could lead to better performance, especially borderline cases. From a clinical implementation perspective, the integration of such models into screening workflows could help prioritize patients for early diagnostic testing, ultimately improving health outcomes through timely intervention.

This project demonstrates that machine learning has a meaningful role to play in early cancer risk prediction, particularly when applied to large-scale clinical datasets. As healthcare systems increasingly seek data-driven approaches to improve efficiency and outcomes, predictive tools like the one developed here can serve as valuable decision-support systems. By flagging individuals who fall into medium or high-risk categories based on demographic and biochemical markers, such models can help clinicians prioritize diagnostic imaging, recommend earlier follow-up, or counsel patients on modifiable risk factors. This could be especially impactful in under-resourced or high-volume clinical environments, where early intervention has the potential to prevent disease progression and reduce long-term treatment costs.

Moreover, integrating machine learning into clinical workflows aligns with broader trends in precision medicine. Risk models can be customized to account for regional, genetic, or lifestyle differences, enhancing their applicability across diverse patient populations. While the current model is limited to the features in our dataset, future iterations could incorporate genomic data, patient history over time, and real-time laboratory results, further personalizing the risk prediction process. Importantly, such tools are not intended to replace clinical judgment, but rather to augment it—surfacing patterns that may not be immediately obvious and allowing providers to make more informed, data-supported decisions. Ultimately, machine learning offers

a scalable and adaptive framework for improving early detection and prevention in oncology. As datasets grow and models are refined through ongoing research, predictive analytics may become a cornerstone of proactive, preventative care in cancer and beyond.

References

American Cancer Society. 2024. “Key Statistics for Thyroid Cancer.” Retrieved May 7, 2025 (https://www.cancer.org/cancer/types/thyroid-cancer/about/key-statistics.html).

American Cancer Society. 2025. “What Is Thyroid Cancer? | Types of Thyroid Cancer | American Cancer Society.” Retrieved May 7, 2025 (https://www.cancer.org/cancer/types/thyroid-cancer/about/what-is-thyroid-cancer.html).

Borzooei, Shiva, Giovanni Briganti, Mitra Golparian, Jerome R. Lechien, and Aidin Tarokhian. 2024. “Machine Learning for Risk Stratification of Thyroid Cancer Patients: A 15-Year Cohort Study.” *European Archives of Oto-Rhino-Laryngology* 281(4):2095–2104. doi: 10.1007/s00405-023-08299-w.

Columbia Thyroid Surgery. 2025. “Thyroid Cancer Diagnosis | Columbia Surgery.” Retrieved May 7, 2025 (https://columbiasurgery.org/thyroid/thyroid-cancer-diagnosis).

Shah, Jatin P. 2015. “Thyroid Carcinoma: Epidemiology, Histology, and Diagnosis.” *Clinical Advances in Hematology & Oncology : H&O* 13(4 Suppl 4):3–6.

Vaccarella, Salvatore, Silvia Franceschi, Freddie Bray, Christopher P. Wild, Martyn Plummer, and Luigino Dal Maso. 2016. “Worldwide Thyroid-Cancer Epidemic? The Increasing Impact of Overdiagnosis.” *New England Journal of Medicine* 375(7):614–17. doi: 10.1056/NEJMp1604412.