Predicting Thyroid Cancer Recurrence Using Machine Learning

Jayden H. Chan, David X. Kang, Ernest K. Yip,

Hong Kong Adventist Academy

Biology B, Grade 10

Ms. Sashi

March 4th, 2025

# Abstract

This research project involves researching the factors that influence thyroid cancer recurrence and developing a simple random forest machine learning algorithm using Python and other imported modules to predict whether a patient is likely to experience the recurrence of thyroid cancer. The algorithm is trained from data, in the form of a .csv file, containing clinicopathologic features about individual patients such as their demographic and diagnosis-specific information (ex. age, adenopathy, focality, et cetera), as well as whether or not they have experienced a recurrence of thyroid cancer. After preprocessing the data using label and one-hot encoding, The algorithm uses 80% of the data to train by analyzing patterns and determining importance through Gini coefficients, then uses the other 20% of the data to make decisions using decision trees, where its accuracy is measured.

Recently, machine learning has become widely used and prevalent in the detection and recurrence predictions of thyroid cancer. In 2021, researchers at Poyang University, South Korea made significant breakthroughs, producing the first study involving the use of photoacoustic images of thyroid nodules to classify them as malignant. This project and algorithm are also inspired by research at Slippery Rock University, Pennsylvania, where researchers used similar datasets and a wide range of more advanced machine-learning algorithms to predict recurrence at near-perfect accuracy (0.99).

*Keywords:* program, pattern, recurrence, features

# Predicting Thyroid Cancer Recurrence Using Machine Learning

# Introduction

Thyroid cancer is the most common type of endocrine cancer (*Endocrine and Thyroid Cancer Care*, n.d.), affecting around 1.1% of people during their lifetime (*Cancer Stat Facts: Thyroid Cancer*, n.d.). It is a type of cancer that occurs in the thyroid gland, a member of the endocrine system, which is responsible for subconscious body regulations such as the secretion of hormones and the regulation of metabolism and heart rate (*Thyroid Cancer,* n.d.). Thyroid cancer comes in many pathologies, with papillary being the most common (accounting for 84%), follicular (around 4%), and hurthle cell (around 2%). These collectively occur from thyroid follicular cells and are labeled as well-differentiated thyroid cancer (Boucai, 2024). The tumor, known as differentiated thyroid carcinoma or abbreviated as DTC, is usually treated using surgical excision (Coca-Pelaz et al., 2023). Symptoms of thyroid cancer include a painless, expanding lump in the lower area of the neck, a sore throat, and difficulty swallowing (Demarco, 2022).

Thyroid cancer has a high survival rate (98.5% for the next 5 years [Boucai et al., 2024]), but “recurrence remains a major concern with up to 20% of patients developing recurrent disease at some point during their lifetime”(Bates et al., 2024). Recurrence can occur from anywhere between a few months to decades after recovery. This program attempts to use the data to visualize which factors or combinations of factors are more likely to correlate to the recurrence of well-differentiated thyroid cancer and simulate decision trees to predict whether or not recurrence is expected.

**使用机器学习预测甲状腺癌复发**

**简介**

甲状腺癌是最常见的内分泌癌类型（*Endocrine and Thyroid Cancer Care*，无日期），约有 1.1% 的人在一生中会受到影响（*Cancer Stat Facts: Thyroid Cancer*，无日期）。它是一种发生在甲状腺中的癌症，甲状腺是内分泌系统的成员，负责潜意识的身体调节，例如激素的分泌和新陈代谢和心率的调节（*Thyroid Cancer*，无日期）。甲状腺癌有多种病理类型，其中乳头状癌最常见（占 84%），滤泡状癌（约 4%）和赫尔细胞状癌（约 2%）。这些癌统称为甲状腺滤泡细胞癌，被称为分化良好的甲状腺癌（Boucai，2024 年）。这种肿瘤被称为分化型甲状腺癌，简称 DTC，通常通过手术切除进行治疗（Coca-Pelaz 等人，2023 年）。甲状腺癌的症状包括颈部下部无痛、不断扩大的肿块、喉咙痛和吞咽困难（Demarco，2022 年）。

甲状腺癌的存活率很高（未来 5 年存活率为 98.5% [Boucai 等人，2024 年]），但“复发仍然是一个主要问题，多达 20% 的患者在其一生中的某个时刻会出现复发性疾病”（Bates 等人，2024 年）。复发可能发生在康复后数月至数十年的任何时间。该程序尝试使用数据来可视化哪些因素或因素组合更有可能与分化良好类型的甲状腺癌复发相关，并模拟决策树以预测是否预计会复发。

# 

# Summary of Research

Depending on the type of thyroid cancer, the average probability of recurrence can be between 6.6 - 28%, although this can vary based on different factors (Ywata, 2021). Demographically, older people naturally face higher recurrence risks due to poorer outcomes from treatment and more aggressive behavior of tumor cells. Because of their hormonal differences (e.g. like lack of estrogen protection), males have higher recurrence rates, Additionally, other lifestyle & environmental factors can also influence these probabilities. Notably, smoking has a paradoxical effect on recurrence risks; though it promotes DNA damage and chronic inflammation in existing tumors, it may reduce thyroid risk through anti-estrogen effects.

However, the most influential factors result from tumor characteristics and pathological features. Besides the patient’s response to treatment, the extent or severity of the thyroid cancer is measured by stage, which in turn is measured by T (tumor size), N (how it affects surrounding lymph nodes), and M (presence in distant sites). Another critical characteristic, type of tumor growth, can usually be measured using physical examinations. Pathological characteristics such as severe types of adenopathy (lymph node enlargement) and pathology (type of tumor) can often directly correlate to recurrence, making it likely a crucial factor in selection algorithms.

# Hypothesis

According to our research, the most important factors in determining thyroid cancer include stage, patient response, adenopathy, and pathology. These factors will likely play the largest role in determining whether or not a patient has a significant risk for recurrence. With sufficient data, the program should be able to effectively predict recurrence with an accuracy >90%.

# Experimental Procedure

First, machine-learning and data manipulation libraries are imported from Python into DeepNote, the collaborative notebook environment used by this project. Then, the .csv data file is imported into the program and saved into a pandas dataframe. In order to begin the machine learning, all the fields within the dataframe must be converted to numeric values. Within the dataframe, two-value categorical features are cleaned and assigned to a numeric value, 0 to 1, while others are replaced using a dictionary, called label encoding. Finally, to prevent the algorithm from accidentally interpreting the arithmetic relationship between the numbers, one-hot encoding is specially used for the “Thyroid Function” column.

Once all the data is cleaned, a random forest classifier is imported. The program defines the “Recurrence” table as the one the machine must categorize. 80% of the patient’s data are used to train the machine, which calculates how often a factor correlates to recurrence and uses gini coefficients to determine the relevance of the factor. After that, the algorithm is tested on the other 20% of patient data and predicts a value, where it is measured based on its accuracy. Using data analysis libraries, we are able to see which factors have a higher gini coefficient (and thus are viewed as more important) as well as the accuracy of the algorithm.

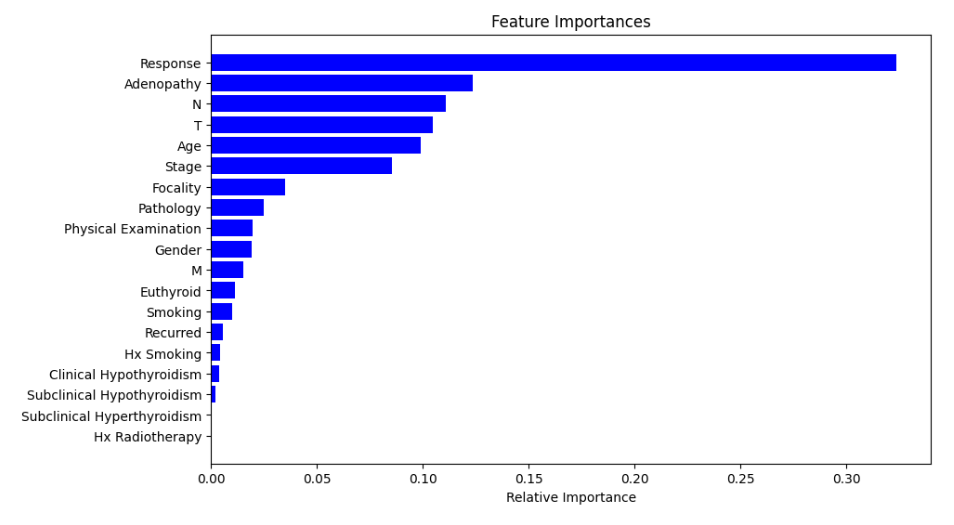
After this, a smaller data set is generated based on the top three factors ranked by the algorithm. Using only these three factors, a second simpler algorithm is produced, where its results are also measured. As a control group or for comparison, another data set is generated to train a third algorithm based on random columns.

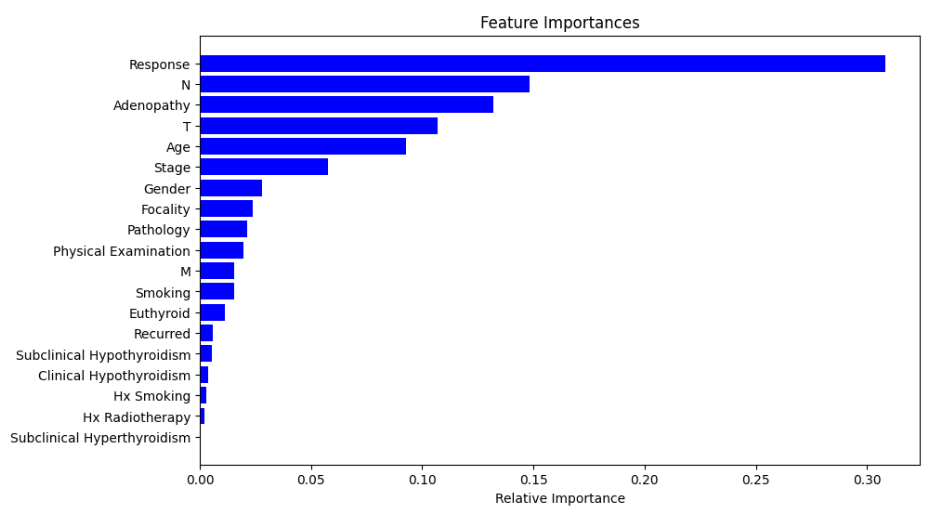
# Results

Here are the following accuracy scores of the algorithms developed:

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 (All) | 96.10% | 96.10% | 94.81% | 96.10% | 96.10% | 97.40% | 96.10% | 96.10% | 96.10% | 94.81% | 95.972% |
| 2 (Top) | 97.40% | 97.40% | 97.40% | 97.40% | 97.40% | 94.81% | 94.81% | 94.81% | 97.40% | 94.81% | 96.364% |
| 3  (Ran) | 87.01% | 87.01% | 87.01% | 87.01% | 84.42% | 87.01% | 84.42% | 83.12% | 84.42% | 85.71% | 85.714% |

Sample Graphs of Factor Importance:

****



**Further Research**

It is possible to consider using other popular machine learning algorithms and compare them against the current random forest classifier. With more time, another possible addition would be the use of SMOTE to increase the accuracy for underrepresented values, such as hurthle cell thyroid cancer. Furthermore, it is possible could create even more models with varying datasets to verify which individual factors are more important.

**Conclusion**

Our hypothesis was mostly correct; we predicted many top factors for a program (Adenopathy, stage characteristics), but there were many of our predictions that were also less important (Physical examination, pathology). The program was successfully able to predict recurrence at a reliable accuracy; proving how invaluable machine learning can be in helping detect not just thyroid cancer recurrence, but expanding to the entire medical field.

# References

Bates, M. F., Lamas, M. R., Randle, R. W., Long, K. L., Pitt, S. C., Schneider, D. F., & Sippel, R. S. (2018). Back so soon? Is early recurrence of papillary thyroid cancer really just persistent disease?. Surgery, 163(1), 118–123. <https://doi.org/10.1016/j.surg.2017.05.028>.

Borzooei, S., Briganti, G., Golparian, M. et al. (2024). Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. <https://doi.org/10.1007/s00405-023-08299-w>.

Boucai, L., Zafereo, M., & Cabanillas, M. E. (2024). Thyroid Cancer: A Review. JAMA, 331(5), 425–435. <https://doi.org/10.1001/jama.2023.26348>.

Cancer Research UK. (n.d.). Number stages for papillary and follicular thyroid cancer. Retrieved August 19, 2024.

Cancer Research UK. (n.d.). TNM staging for thyroid cancer. Retrieved August 19, 2024.

*Cancer Stat Facts: Thyroid Cancer*. (n.d.) SEER National Cancer Institute.

<https://seer.cancer.gov/statfacts/html/thyro.html>.

Clark, E., Price, S., Lucena, T., Haberlein, B., Wahbeh, A., & Seetan, R. (2024). Predictive

Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach. Knowledge, 4(4), 557-570. <https://doi.org/10.3390/knowledge4040029>.

Coca-Pelaz, A., Rodrigo, J. P., Shah, J. P., Nixon, I. J., Hartl, D. M., Robbins, K. T., Kowalski, L. P., Mäkitie, A. A., Hamoir, M., López, F., Saba, N. F., Nuyts, S., Rinaldo, A., & Ferlito, A. (2023). Recurrent Differentiated Thyroid Cancer: The Current Treatment Options. Cancers, 15(10), 2692. <https://doi.org/10.3390/cancers15102692>.

Demarco, Cynthia. (2022). Papillary thyroid cancer: 8 questions, answered. *The University of*

*Texas MD Anderson Cancer Center.*

<https://www.mdanderson.org/cancerwise/what-is-papillary-thyroid-cancer--8-questions--answered.h00-159543690.html#:~:text=Papillary%20thyroid%20cancer%20is%20the,of%20all%20thyroid%20cancer%20cases>.

*Endocrine and Thyroid Cancer Care.* (n.d.). UPMC Hillman Cancer Center.

<https://hillman.upmc.com/cancer-care/endocrine-thyroid#:~:text=Thyroid%20cancer%20is%20the%20most,very%20treatable%20form%20of%20cancer>.

N'ara S., Amit M., Fridman E., Gil Z. (2016). Contemporary Management of Recurrent Nodal Disease in Differentiated Thyroid Carcinoma. PubMed Central. Retrieved August 19, 2024.

*Thyroid Cancer.* (n.d.). Cleveland Clinic.

<https://my.clevelandclinic.org/health/diseases/12210-thyroid-cancer>.

Ywata de Carvalho, A., Kohler, H. F., Gomes, C. C., Vartanian, J. G., & Kowalski, L. P. (2021).

Predictive factors for recurrence of papillary thyroid carcinoma: analysis of 4,085 patients. Acta otorhinolaryngologica Italica : organo ufficiale della Societa italiana di otorinolaringologia e chirurgia cervico-facciale, 41(3), 236–242. <https://doi.org/10.14639/0392-100X-N1412>.

Data for this project is derived from:

<https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence>

By running through complex datasets with multiple factors, our program will sort out (or seek) patterns by pointing out the core determining factors of thyroid cancer recurrence. Using those factors, the program can make a prediction on whether thyroid cancer would reoccur.

The program successfully predicted reoccurrence with an average accuracy of around 96%

fun fact: Smoking actually has a paradoxical effect on the body. It may reduce thyroid risk through anti-estrogen effects but promotes DNA damage and chronic inflammation in existing tumors

科學：小寒期間，地球是離太陽最近的不是最遠，這是一個常見的誤解。這是因為地球的北面（中國在的地方），是向上傾斜的，所以陽光不能直接指向北半球，但在南半球是相反的。北半球夏天的時候是南半球的冬天。