Analysis of Thyroid Cancer Risk with Ordinal Logistic Regression Model Using R

Ann Abigail Halim
Computer Science and Statistics
BINUS University
Jakarta, Indonesia
ann.halim@binus.ac.id

Carmenita Angelica
Computer Science and Statistics
BINUS University
Jakarta, Indonesia
carmenita.angelica@binus.ac.id

Joycelin
Computer Science and Statistics
BINUS University
Jakarta, Indonesia
joycelin@binus.ac.id

***Abstract*—Thyroid cancer incidence is increasing globally, emphasizing the importance of early risk identification and preventive strategies. This study applies an ordinal logistic regression model using R to classify thyroid cancer risk into low, medium, and high categories based on demographic, lifestyle, and physiological factors. The dataset, sourced from Kaggle, comprises over 212,000 records with variables including age, gender, ethnicity, country, family history, radiation exposure, iodine deficiency, and hormone levels (TSH, T3, T4).**

Exploratory and inferential analysis identified key categorical predictors significantly associated with higher risk, particularly family history, radiation exposure, iodine deficiency, ethnicity, and country. Three models were developed to evaluate the predictive power of all variables, numerical only, and categorical only features. The model using only categorical variables achieved a classification accuracy of 51.356%, comparable to the full model, demonstrating that categorical indicators alone are sufficient for effective risk stratification.

These findings suggest that ordinal logistic regression is a practical and interpretable tool for early thyroid cancer risk assessment. Future improvements could involve enhancing model accuracy through advanced algorithms and balanced datasets for more precise risk differentiation.

Keywords—Thyroid cancer, ordinal logistic regression, categorical data analysis, risk classification, R programming, predictive modeling

I. INTRODUCTION

A. Introduction

Thyroid cancer is a type of malignant tumour that develops in the thyroid gland, a tiny, butterfly-shaped organ at the base of the neck that is important in controlling metabolism through hormone production [1]. The American Cancer Society emphasizes that thyroid cancer grows when genetic mutations in thyroid cells trigger excessive expansion and multiplying, that eventually results in the growth of a cancer [2]. Each of those four categories of thyroid cancer, papillary, follicular, medullary, and anaplastic, each of which has a different prognostic significance [3].

Over the past few decades, the number of people diagnosed with thyroid cancer worldwide has increased significantly. GLOBOCAN 2022 states that there were approximately 821,000 new cases of thyroid cancer worldwide [4]. This is why it is one of the most common endocrine malignancies. This increase is partly due to new ways of finding the disease, such as high-resolution imaging and fine-needle aspiration cytology [5]. However, there has also been a substantial increase in cases, especially among women and younger people. Interestingly, the number of new cases continues to increase, but the global mortality rate is relatively low, with approximately 47,500 deaths reported in the same year [4]. This disparity illustrates that most thyroid tumour, especially papillary types, grow slowly most of the time. However, there is still a large disparity in how well people with cancer survive because many low and middle income countries still have problems getting a quick diagnosis and appropriate therapy [6]. This highlights the importance of healthcare resources and public health initiatives to be more equitable around the world.

Effective treatment of thyroid cancer requires a multidisciplinary approach including early identification, surgery, radioactive iodine treatment, and in certain cases focused medicines. Early identification is crucial as, particularly in cases of early stages or a limited cancer, it significantly improves outcomes. High-resolution ultrasonic waves and fine-needle aspiration biopsies are two diagnostic tools that really help early identification [5]. Furthermore, guiding treatment decisions and patient monitoring is knowledge of the risk factors: radiation exposure, aberrant thyroid hormone levels (TSH, T3, and T4), and family history of thyroid disease [7]. The comprehensive review also emphasizes environmental and genetic risk factors such as exposure to ionizing radiation, obesity, lack of physical activity, and genetic predisposition, such as family history, and how these assessments combined with imaging and cytology can aid in personalized treatment planning and long-term outcomes [8].

Researchers are utilizing statistical models, such as ordinal logistic regression to classify patients based on risk profiles, to combat the increasing load of thyroid cancer. Diverse patient data, including medical history, and lifestyle habits, are analyzed by models and pinpointed ones with higher risk levels. This data-driven method allows healthcare providers to better prioritize examination and intervention for

patients, supporting attempts of early detection and late-stage diagnoses reduction [9].

B. Related Works

The rising prevalence of thyroid cancer worldwide has prompted heightened research into the identification of risk factors and early diagnosis by statistical and machine learning models. A multitude of studies have concentrated on creating predictive models for thyroid cancer to facilitate early intervention and enhance prognosis, particularly in resource-limited environments. Wang, et al. indicate that enhancements of deep learning in ultrasonography methodologies have markedly increased diagnostic sensitivity for thyroid nodules [10].

Ordinal logistic regression (OLR) is frequently utilized for multi-class classification tasks, including cancer risk stratification. OLR has demonstrated efficacy in identifying significant predictors of the stadium of thyroid cancer [11]. A study by Girardi et al. employed logistic regression on clinical data to forecast thyroid malignancy, showcasing significant interpretability and clinical significance in their results [12].

Exposure to ionizing radiation during childhood and adolescence significantly increases the risk of thyroid cancer, as demonstrated by large-scale cohort studies of populations affected by the Chornobyl accident. These findings are supported by dose-response analyses showing a strong, linear relationship between thyroid radiation dose and cancer risk, with an excess relative risk of 5.25 per Gray among those exposed at young ages [13]. Family history constitutes a significant risk factor for hereditary non-medullary thyroid cancer (HNMTC), as evidenced by epidemiologic studies showing that when three or more first-degree relatives are affected, there is a >94% likelihood of hereditary predisposition. Although the specific susceptibility remains unidentified, the association with more aggressive disease underscores the clinical importance of identifying at-risk families [14].

Lifestyle factors, including smoking and alcohol consumption, have shown inconsistent associations with thyroid cancer. Kitahara et al. indicated that cigarette smoking is inversely correlated with the risk of thyroid cancer, potentially attributable to nicotine's influence on thyroid-stimulating hormone (TSH) levels, although the underlying mechanism is still debated [15]. Conversely, Ma, et al. underlined that obesity has been positively associated with the risk of thyroid cancer, presumably via hormonal and inflammatory mechanisms [16].

Numerous research have concentrated on biochemical markers, specifically thyroid hormones T3, T4, and TSH. Increased TSH levels, even within the normal range, correlate with a heightened risk of differentiated thyroid carcinoma, as demonstrated in a population-based cohort study by Fiore [17]. Integrating these physiological indicators with demographic and lifestyle information strengthens the reliability of predictive models.

Algorithms such as Random Forests, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM) have been widely used to improve the prediction of thyroid nodule malignancy. Comparative studies suggest that while ensemble methods like Random Forests and GBM often achieve higher accuracy by capturing complex patterns, simpler models such as logistic regression remain valuable for interpretability and clinical utility, justifying their role as baseline references in performance evaluations [18].

Data-driven approaches have become increasingly vital in supporting clinicians with personalized diagnostic decisions. For instance, machine learning models, such as Support Vector Classification (SVC) and logistic regression, have been employed to enhance preoperative thyroid nodule screening, as demonstrated in Khodabandelu et al. study using ultrasonography features from 431 nodules. While all models showed strong accuracy (91-92.1%) and AUC (92.6-93.2%), the balanced dataset improved geometric mean scores by 7.4 percentage points for Model 2 versus Model 1. Key malignant predictors included microcalcifications and non-isoechoic patterns, with no gender-based prevalence differences observed. These findings highlight that addressing class imbalance through techniques like SMOTE enhances model reliability for thyroid cancer detection without compromising overall accuracy [19].

Publicly accessible datasets on platforms such as Kaggle have markedly expedited thyroid cancer research by offering uniform data for model evaluation. Zeeshan's "Thyroid Cancer Risk Dataset" has been utilized in educational and research settings to investigate risk factor modeling and to create practical tools [20].

The necessity for accessible and interpretable risk classification models is particularly vital in low- and middle-income countries, where access to nuclear imaging and sophisticated biopsies may be constrained. P. Li, et al. pointed out that basic statistical models utilizing accessible clinical factors might significantly aid in early diagnosis in resource-limited areas [21].

Finally, research demonstrates how machine learning innovations can bridge critical gaps in preoperative PTC management, exemplifying the multidisciplinary nature of modern oncology research. By integrating clinical biomarkers, imaging characteristics, and algorithmic modeling into an accessible web platform, we provide a comprehensive solution for LLNM risk stratification that addresses both diagnostic challenges and therapeutic decision-making. The development of such predictive tools reflects the growing imperative of combining technological advancements with clinical expertise in tackling thyroid cancer's global rise. As evidenced by our robust performance metrics and clinically interpretable risk factors, this approach offers a template for transforming complex data into actionable insights at the point of care [22].

C. Objectives

1) To identify and evaluate significant categorical variables that are statistically associated with thyroid cancer risk levels.

2) To compare the influence of lifestyle, medical history, and physiological indicators on thyroid cancer risk using statistical testing and modeling.

3) To develop a risk classification model using ordinal logistic regression that stratifies patients into low, medium, and high risk categories based on demographic and clinical features.

4) To assess and compare the predictive performance of models built using different types of features (all, only numerical, and only categorical), and determine the optimal trade-off between model complexity and interpretability.

D. Methodology

1) Data Collection

The dataset used in this project is secondary data obtained from Kaggle, specifically from the "Thyroid Cancer Risk Dataset" available at kaggle.com/datasets/mzohaibzeeshan/thyroid-cancer-risk-dataset. The data was last updated in February 2025. It consists of 212,691 rows and 17 columns. The target variable used for analysis is *Thyroid_Cancer_Risk*.

2) Data Preprocessing

a) Data Loading

- The Patient_ID variable was removed prior to analysis as it served solely as a unique identifier and held no analytical value.
- Converted the categorical variables that will not be used as dependent variables, such as *Family_History*, *Gender*, *Smoking*, *etc*., into factor type.
- Checked for missing values, null data, or data inconsistencies and handling them appropriately using either imputation or deletion.

b) Exploratory Data Analysis (EDA)

- Performed frequency distribution for categorical variables.
- Generated summary statistics for continuous variables.
- Create visualizations (bar charts, box plots, etc.) to explore patterns.
- Find out the relationship between each of the categorical indicator variables and the target variable *Thyroid_Cancer_Risk* using Chi-Square Test.

3) Model Building

To evaluate the impact of different types of predictors on model performance, we constructed three separate ordinal logistic regression models using the polr() function in R. The first model included all available predictors, the second used only numerical variables (e.g., age, hormone levels), and the third relied solely on categorical variables (e.g., gender, ethnicity, family history). This comparison enabled us to assess the relative predictive power and interpretability of each feature group.

4) Model Evaluation

- Confusion Matrix: It is employed to evaluate the model's efficacy in categorizing the samples into three classifications: Low, Medium, and High.
- Accuracy: It is determined as the proportion of correct predictions to the total number of forecasts.

II. Result and Discussion

A. Dataset Description

1) Dataset Structure

- Age : int, numerical variable in integer
- Gender: Factor, nominal binary categorical variable (Male and Female)
- Country: Factor, nominal multi-categorical variable (Brazil, China, Germany, India, Japan Nigeria, Russia, South Korea, UK, and USA)
- Ethnicity: Factor, nominal multi-categorical variable (African, Asian, Caucasian, Hispanic, and Middle Eastern)
- Family_History: Factor, nominal binary categorical variable (Yes and No)
- Radiation_Exposure: Factor, nominal binary categorical variable (Yes and No)
- Iodine_Deficiency: Factor, nominal binary categorical variable (Yes and No)
- Smoking: Factor, nominal binary categorical variable (Yes and No)
- Obesity: Factor, nominal binary categorical variable (Yes and No)
- Diabetes Factor, nominal binary categorical variable (Yes and No)
- TSH_Level: num, numerical variable in decimals
- T3_Level: num, numerical variable in decimals
- T4_Level: num, numerical variable in decimals
- Nodule_Size: num, numerical variable in decimals
- Thyroid_Cancer_Risk: Factor, ordinal multi-categorical variable (Low, Medium, and High)
- Diagnosis : Factor, nominal binary categorical variable (Benign and Malignant)

2) Dataset Summary

Table I.
Summary of Numerical Variables

	Age	TSH_Level	T3_Level	T4_Level	Nodule_Size
Min.	15.00	0.100	0.500	4.500	0.000
1st Qu.	33.00	2.570	1.250	6.370	1.250
Median	52.00	5.040	2.000	8.240	2.510
Mean	51.92	5.045	2.002	8.246	2.503
3rd Qu.	71.00	7.520	2.750	10.120	3.760
Max.	89.00	10.000	3.500	12.000	5.000

As shown in Table I, numerical variables such as Age, TSH, T3, T4, and Nodule Size exhibit a wide range. The average age is 51.92 years. Hormone levels and nodule sizes show moderate variability, supporting the dataset's clinical relevance.

Table II.
Summary of Categorical Variables (Gender, Country, and Ethnicity)

Gender		Country		Ethnicity	
Female	127527	India	42496	Afrian	42414
Male	85164	China	31978	Asian	53261
		Nigeria	31918	Caucasian	63669
		Brazil	21413	Hispanic	32012
		Russia	21297	Middle Eastern	21335
		Japan	16867		
		(Other)	46722		

Table II highlights a larger female population (127,527 vs. 85,164 males), with patients from countries such as India, China, and Nigeria. Ethnic groups include African, Asian, Caucasian, Hispanic, and Middle Eastern populations.

Table III.
Summary of Categorical Variables (F_H, Rad. Smoke, Obesity, and Diabetes)

	F_H	Rad.	Iodine	Smoke	Obesity	Diabetes
No	148866	180831	159673	170260	148805	170098
Yes	63825	31860	53018	42431	63886	42693

According to Table III, lifestyle and medical history variables. Most patients reported no family history, radiation exposure, or smoking. However, a significant portion had obesity (63,886) and diabetes (42,693).

Table IV.
Summary of Categorical Variables (T_Cancer_Risk, Diagnosis)

T_Cancer_Risk		Diagnosis	
Low	108388	Benign	163196
Medium	72400	Malignant	49495
High	31903		

Table IV summarizes most cases were low-risk (108,388), with 49,495 diagnosed as malignant. These labels support downstream predictive modeling.

3) Size of Dataset
Number of rows: 212691
Number of columns: 16

4) Targeted Variable
Dependent Variable: *Thyroid_Cancer_Risk*
Benefit:

- *Thyroid_Cancer_Risk* has multiple categories (Low, Medium, High) that allows for a more detailed classification model.
- Predicting risk level helps clinicians or healthcare systems screen or prioritize patients before final diagnosis. Aligns with preventive medicine, identifying high-risk individuals early for further testing.
- These models provide probabilities per class, making the output interpretable and suitable for decision thresholds.

B. *Data Prepocessing*

1) Drop Variables
Dropping Patient_ID variable since it's only used to identify each patient uniquely and only need to keep the variables that may impact the response variable.

2) Check and Handle Missing Values
After checking, it turns out that the dataset has no missing value. For this case, no imputation or further steps are needed.

3) Convert Categorical Variables
Converting all categorical variables in the data, such as *Gender, Country, Ethnicity, Family_History, Radiation_Exposure, Iodine_Deficiency, Smoking, Obesity, Diabetes, Thyroid_Cancer_Risk, and Diagnosis* into a factor.

C. *Exploratory Data Analysis*

1) Distribution of Thyroid Cancer Risk

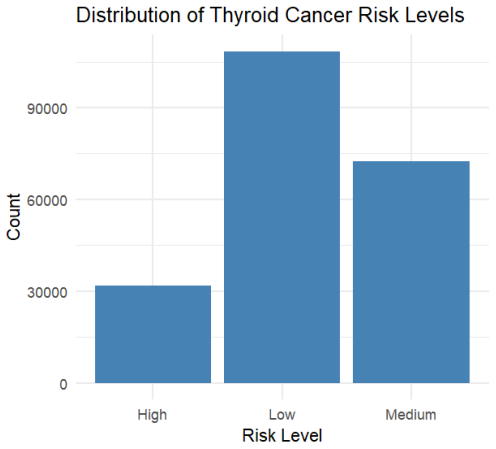

Fig. 1. Distribution of Thyroid Cancer Risk

This bar plot shows the three levels of thyroid risk levels. The categories are low, which dominate the distribution, medium, representing a smaller proportion, and the smallest proportion of high level.

2) Age Distribution by Thyroid Cancer Risk

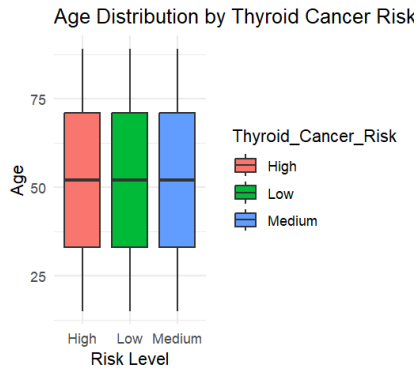

Fig. 2. Age Distribution by Thyroid Cancer Risk

This plot shows that the distribution of age between thyroid cancer risk's levels are almost the same.

3) TSH Level Distribution by Thyroid Cancer Risk

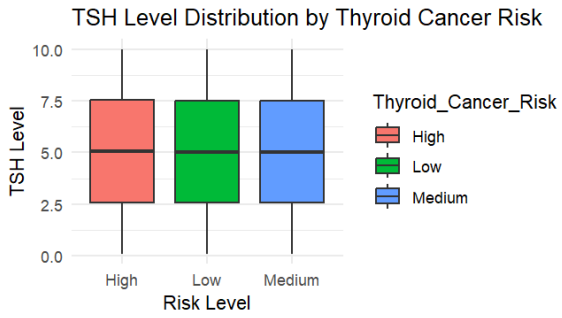

Fig. 3. TSH Level Distribution by Thyroid Cancer Risk

Same as Age, this plot shows that the distribution of TSH Level between thyroid cancer risk's levels is almost the same.

4) T3 Level Distribution by Thyroid Cancer Risk

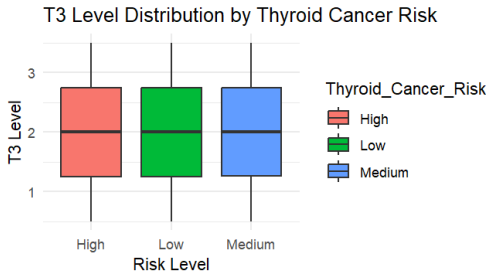

Fig. 4. T3 Level Distribution by Thyroid Cancer Risk

This plot shows that the distribution of T3 Level between thyroid cancer risk's levels are also almost the same.

5) T4 Level Distribution by Thyroid Cancer Risk

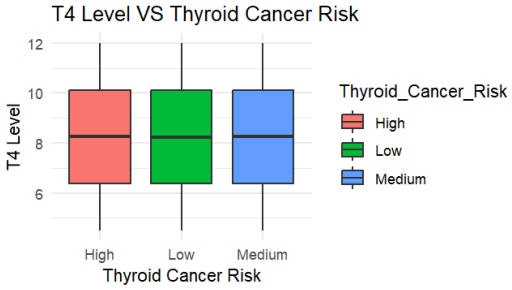

Fig. 5. T4 Level Distribution by Thyroid Cancer Risk

Same with before, this plot shows that the distribution of T4 Level between thyroid cancer risk's levels are also almost the same.

6) Nodule Size Distribution by Thyroid Cancer Risk

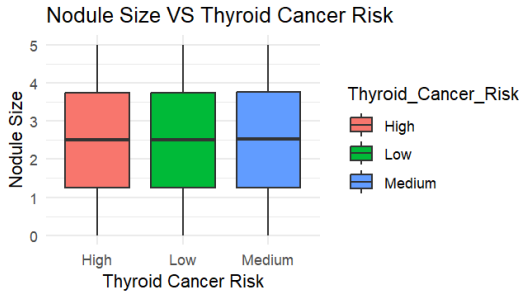

Fig. 6. Nodule Size Distribution by Thyroid Cancer Risk

This plot shows that the distribution of nodule size between thyroid cancer risk's levels are almost the same.

7) Spearman Correlation Test between Thyroid Cancer Risk and Numerical Variables

To examine the monotonic relationship between thyroid cancer risk level (Risk_Numeric) and numerical variables, Spearman's rank correlation tests were conducted. The results are as follows:

- Age: $\rho = 0.0033$, $p = 0.133$
- TSH Level: $\rho = 0.0023$, $p = 0.292$
- T3 Level: $\rho = 0.0002$, $p = 0.927$
- T4 Level: $\rho = 0.0030$, $p = 0.161$
- Nodule Size: $\rho = -0.0002$, $p = 0.917$

None of the correlations were statistically significant ($p > 0.05$), indicating no evidence of a monotonic relationship between the risk level and any of the numerical variables.

8) Chi-square Test between Thyroid Cancer Risk and Categorical Variables

Table V.
Chi-Square Test Between Thyroid Cancer Risk and Categorical Variables

Variable	χ^2	df	*p*-value	Significance
Gender	5.2414	2	0.07275	Not Significant
Country	13307	18	$< 2.2\times10^{-16}$	Significant
Ethnicity	24049	8	$< 2.2\times10^{-16}$	Significant

Family History	19927	2	$< 2.2\times10^{-16}$	Significant
Radiation Exposure	7505.2	2	$< 2.2\times10^{-16}$	Significant
Iodine Deficiency	9946.1	2	$< 2.2\times10^{-16}$	Significant
Smoking	1.7412	2	0.4187	Not Significant
Obesity	2.1119	2	0.3479	Not Significant
Diabetes	1.2533	2	0.5344	Not Significant

The Pearson's Chi-square test of independence was employed to assess the association between thyroid cancer risk and several categorical variables. The results are presented in Table I. Significant associations were found with Country, Ethnicity, Family History, Radiation Exposure, and Iodine Deficiency ($p < 0.05$). In contrast, no statistically significant relationship was observed for Gender, Smoking, Obesity, or Diabetes.

9) Proportion of Thyroid Cancer Risk by Country

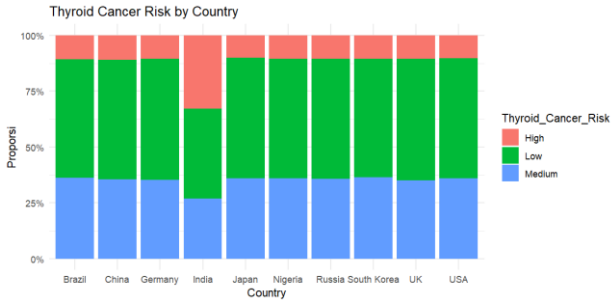

Fig. 7. Proportion of Thyroid Cancer Risk by Country

The distribution of thyroid cancer risk levels is generally consistent across countries, with most showing similar proportions of low, medium, and high risk. However, India stands out with a noticeably higher proportion of individuals categorized as high risk.

10) Proportion of Thyroid Cancer Risk by Ethnicity

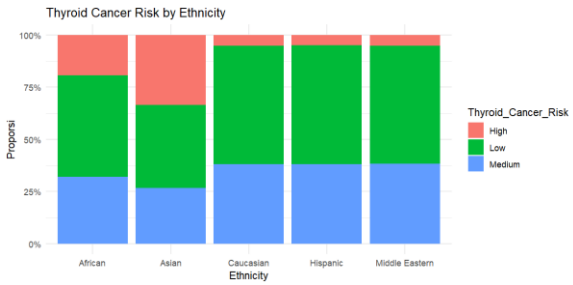

Fig. 8. Proportion of Thyroid Cancer Risk by Ethnicity

The visualizations indicate a clear disparity in thyroid cancer risk levels among different ethnic groups. Individuals of Asian ethnicity exhibit the highest proportion of high-risk classifications, followed by those of African ethnicity. In contrast, the proportions of individuals classified as high risk are notably lower among Caucasian, Hispanic, and Middle Eastern groups, with these groups showing a greater tendency toward low-risk classifications.

11) Proportion of Thyroid Cancer Risk by Family History

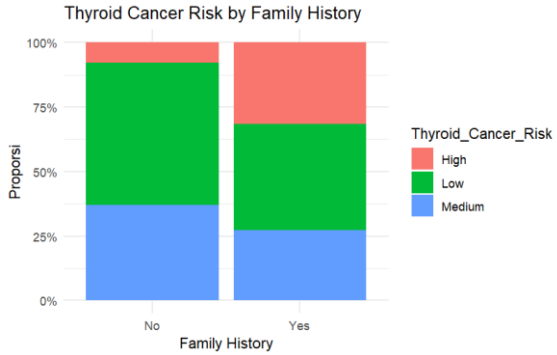

Fig. 9. Proportion of Thyroid Cancer Risk by Family History

The stacked bar chart shows the individuals with a family history of thyroid cancer show a higher proportion of high-risk classifications compared to those without such a history.

12) Proportion of Thyroid Cancer Risk by Radiation Exposure

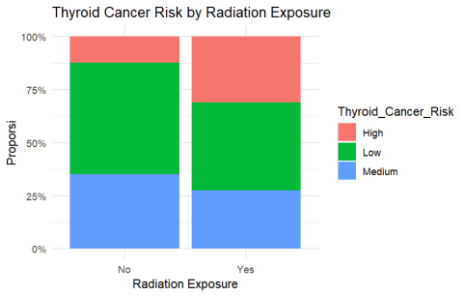

Fig. 10. Proportion of Thyroid Cancer Risk by Radiation Exposure

This plot indicates that individuals with a history of radiation exposure exhibit a higher proportion of high-risk classifications for thyroid cancer compared to those without such exposure.

13) Proportion of Thyroid Cancer Risk by Iodine Deficiency

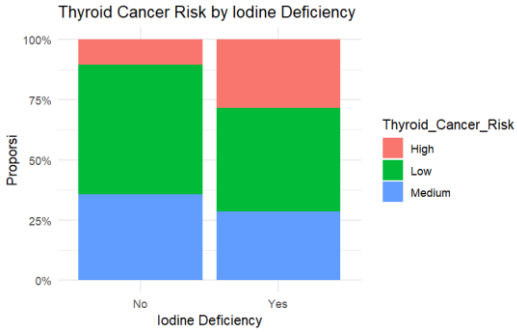

Fig. 11. Proportion of Thyroid Cancer Risk by Iodine Deficiency

The visualization indicates that individuals with iodine deficiency have a higher proportion of high-risk classifications for thyroid cancer compared to those without iodine deficiency

D. Preparation Before Building Models

1) Import Libraries

The used library for building models is MASS. The MASS library provides access to various statistical functions, including as.factor(), which is used to convert a variable into a factor type. In this study, the MASS package plays a key role in model building, particularly in the implementation of ordinal logistic regression through the polr() function.

2) Converting the Target Variable to a Factor

The target variable Thyroid_Cancer_Risk is explicitly converted to a factor to ensure compatibility with classification models. This step is essential because many machine learning algorithms in R, including ordinal logistic regression, require the dependent variable to be a factor when dealing with categorical outcomes.

E. Building Categorical Analysis Model

1) Model Building

In this study, three ordinal logistic regression models were developed to predict thyroid cancer risk based on different subsets of features.

Model a) utilized all available predictor variables, both categorical and numerical. Model b) was trained using only numerical variables, while Model c) employed only categorical variables.

a) Ordinal Logistic Regression Using All Variables

Among the variables provided in the dataset, the Diagnosis variable was excluded from model training. This variable indicates whether a clinical diagnosis related to thyroid cancer risk has already been made. The model is intended to assist in early risk prediction before any clinical diagnosis is made. Therefore, using the Diagnosis variable would result in artificially inflated model performance and would not generalize to unseen cases. To ensure a valid and unbiased predictive model, only pre-diagnostic features such as hormone levels, age, lifestyle factors, and family history were included in the training process.

Table VI.
Result of Ordinal Logistic Regression Model Using All Variables

Variable	Estimate	Std.Error	z_value	Pr(>|z|)
Age	0.00029	0.00020	1.44927	0.14726
GenderMale	-0.01421	0.00882	-1.61181	0.10700
CountryChina	-0.02799	0.01761	-1.58991	0.11186
CountryGermany	-0.04367	0.02375	-1.83888	0.06593
CountryIndia	0.94967	0.01673	56.75757	0.00000
CountryJapan	-0.04261	0.02054	-2.07450	0.03803
CountryNigeria	-0.01977	0.01760	-1.12365	0.26116
CountryRussia	-0.03473	0.01930	-1.79913	0.07200
CountrySouth Korea	-0.01095	0.02121	-0.51656	0.60546
CountryUK	-0.06605	0.02374	-2.78148	0.00541
CountryUSA	-0.02845	0.02372	-1.19981	0.23021
Ethnicity Asian	0.63862	0.01283	49.76682	0.00000
Ethnicity Caucasian	-0.61983	0.01256	-49.35527	0.00000
Ethnicity Hispanic	-0.63005	0.01488	-42.33552	0.00000
Ethnicity Middle Eastern	-0.60564	0.01686	-35.93032	0.00000
Family_HistoryYes	1.06129	0.00950	111.71713	0.00000
Radiation_ExposureYes	0.80121	0.01225	65.37992	0.00000
Iodine_DeficiencyYes	0.78915	0.00999	78.96219	0.00000
SmokingYes	-0.01859	0.01082	-1.71836	0.08573
ObesityYes	-0.01058	0.00942	-1.12334	0.26129
DiabetesYes	-0.00474	0.01080	-0.43892	0.66072
TSH_Level	0.00138	0.00151	0.91651	0.35940
T3_Level	0.00313	0.00499	0.62714	0.53057
T4_Level	0.00241	0.00200	1.20476	0.22830
Nodule_Size	-0.00007	0.00299	-0.02508	0.97999
Low|Medium	0.58569	0.03007	19.47952	0.00000
Medium|High	2.58684	0.03072	84.21312	0.00000

The model's coefficients and corresponding p-values were computed to assess the statistical significance of each predictor. Several predictors were found to be highly significant ($p < 0.05$), including:

- CountryIndia, EthnicityAsian, EthnicityCaucasian, EthnicityHispanic, and EthnicityMiddle Eastern
- Family_HistoryYes, Radiation_ExposureYes, and Iodine_DeficiencyYes

These variables demonstrated strong associations with thyroid cancer risk levels. Other variables, such as Age, TSH_Level, T3_Level, T4_Level, Nodule_Size, and lifestyle factors like Smoking, Obesity, and Diabetes did not show statistically significant associations at $\alpha = 0.05$.

The confusion matrix for the predictions compared to actual thyroid cancer risk levels is summarized in Table I.

The model correctly predicted the risk level in approximately 51.352% of cases.

Table VII.
Confusion Matrix of Ordinal Logistic Regression Model Using All Variables

Predicted vs Actual	Low	Medium	High
Low	78174	52391	2759
Medium	29661	9665	17761
High	553	344	11383

PARAMETER ESTIMATION AND HYPOTHESIS TESTING FOR MODEL A

The ordinal logistic regression model, based on the proportional odds assumption, estimates parameters using maximum likelihood estimation (MLE). This means that the model assumes the log-odds of being at or below a particular category are a linear function of the independent variables. Formally, the model can be written as:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \theta_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where:

- j=1, 2 corresponds to the ordinal outcome levels (Low, Medium, High),
- θ_j represents the threshold (intercept) for category j,
- $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients for the predictors $X_1, X_2, \ldots, X_k$

Each β coefficient reflects the effect of a one-unit increase in the corresponding predictor on the log-odds of being in a higher risk category.

The coefficients for each predictor were estimated using MLE, and the results are summarized in Table VI, including standard errors, z-values, and p-values.

To test whether each predictor significantly contributes to the model, hypothesis testing was conducted using the following framework:

To evaluate the statistical significance of each predictor, hypothesis testing is conducted using the following:

- Null hypothesis (H_0): The predictor variable has no effect on thyroid cancer risk.
- Alternative hypothesis (H_1): The predictor variable has a significant effect on thyroid cancer risk.

Variables with p-values < 0.05 are considered statistically significant and likely to contribute to the prediction of higher cancer risk levels. These include:

- Family_HistoryYes ($p < 0.05$)
- Radiation_ExposureYes ($p < 0.05$)
- Iodine_DeficiencyYes ($p < 0.05$)
- CountryIndia, EthnicityAsian, EthnicityCaucasian, EthnicityHispanic, and EthnicityMiddle Eastern

On the other hand, predictors such as Age, TSH_Level, T3_Level, T4_Level, Nodule_Size, Smoking, Obesity, and Diabetes have p-values greater than 0.05 and are thus not statistically significant in this model.

The magnitude and sign of each β coefficient indicate the direction and strength of the variable's contribution to thyroid cancer risk classification. For instance, a positive β for Family_HistoryYes suggests that having a family history increases the odds of being in a higher-risk category

b) Ordinal Logistic Regression Using Only Numerical Variables

Model B was trained using only numerical predictors such as age, hormone levels (TSH, T3, T4), and nodule size, excluding any categorical features. The goal was to evaluate how well the numerical features alone can predict thyroid cancer risk.

Table VIII.
Result of Ordinal Logistic Regression Model Using Only Numerical Variables

Variable	Estimate	Std. Error	t-value	p-value
Age	0.00029	0.00019	1.50454	0.13244
TSH_Level	0.00153	0.00145	1.05801	0.29005
T3_Level	0.00047	0.00478	0.09777	0.92212
T4_Level	0.00268	0.00191	1.40398	0.16032
Nodule_Size	-0.00029	0.00286	-0.09983	0.92048
Low	Medium	0.08342	0.02376	3.51097
Medium	High	1.77966	0.02415	73.68649

None of the numerical predictors were statistically significant at the 0.05 level, except for the intercept thresholds separating the risk categories (Low|Medium and Medium|High), indicating limited predictive power from numerical variables alone.

The confusion matrix (Table V) showed the model predicted all cases as belonging to the "Low" risk category, resulting in an overall accuracy of approximately 50.96%.

Table IX.
Confusion Matrix of Ordinal Logistic Regression Model Using Only Numerical Variables

Predicted \ Actual	Low	Medium	High
Low	108388	72400	31903
Medium	0	0	0
High	0	0	0

PARAMETER ESTIMATION AND HYPOTHESIS TESTING FOR MODEL B

Model B applies the proportional odds logistic regression model using only numerical predictors such as Age, TSH_Level, T3_Level, T4_Level, and Nodule_Size. The formulation is:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \theta_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where j=1,2 represents the thresholds for the three-category outcome (Low, Medium, High). Each β indicates the effect of a one-unit increase in the respective numerical variable on the log-odds of being in a higher risk category, and predictors X include Age, TSH_Level, T3_Level, T4_Level, and Nodule_Size.

To evaluate the significance of each numerical predictor in Model B, the following hypotheses were tested:

- Null hypothesis (H_0): The numerical predictor has no effect on thyroid cancer risk.
- Alternative hypothesis (H_1): The numerical predictor has a significant effect on thyroid cancer risk.

Model B was constructed using only numerical predictors, including age, TSH, T3, T4, and nodule size. The

parameter estimates obtained through MLE showed that none of these variables had statistically significant p-values (all $p > 0.05$), suggesting a lack of meaningful association between these numerical indicators and thyroid cancer risk levels.

- Age: p = 0.13244
- T3_Level: p = 0.92212
- T4_Level: p = 0.16032
- Nodule_Size: p = 0.92048

This reinforces the conclusion that numerical variables alone do not provide adequate discriminatory power in stratifying thyroid cancer risk using ordinal logistic regression.

c) Ordinal Logistic Regression Using Only Categorical Variables

Model c) was trained using only categorical predictors such as gender, country, ethnicity, family history, radiation exposure, iodine deficiency, smoking, obesity, and diabetes status. Numerical variables were excluded to evaluate the predictive power of categorical features alone.

Table X.
Result of Ordinal Logistic Regression Model Using Only Categorical Variables

Variable	Estimate	Std. Error	t-value	p-value
GenderMale	-0.01423	0.00882	-1.61391	0.10655
Country China	-0.02807	0.01761	-1.59459	0.11080
Country Germany	-0.04387	0.02375	-1.84726	0.06471
CountryIndia	0.94956	0.01673	56.75218	0.00000
CountryJapan	-0.04262	0.02054	-2.07511	0.03798
Country Nigeria	-0.01974	0.01760	-1.12196	0.26188
Country Russia	-0.03480	0.01930	-1.80291	0.07140
CountrySouth Korea	-0.01099	0.02120	-0.51819	0.60433
Country UK	-0.06611	0.02374	-2.78414	0.00537
Country USA	-0.02846	0.02371	-1.20026	0.23004
Ethnicity Asian	0.63851	0.01283	49.75936	0.00000
Ethnicity Caucasian	-0.61990	0.01256	-49.36141	0.00000
Ethnicity Hispanic	-0.63007	0.01488	-42.33768	0.00000
Ethnicity Middle Eastern	-0.60581	0.01686	-35.94109	0.00000
Family_HistoryYes	1.06136	0.00950	111.72710	0.00000
Radiation_ExposureYes	0.80127	0.01225	65.38611	0.00000
Iodine_Deficiency Yes	0.78911	0.00999	78.95973	0.00000
Smoking Yes	-0.01858	0.01082	-1.71748	0.08589
ObesityYes	-0.01051	0.00942	-1.11545	0.26466
Diabetes Yes	-0.00474	0.01080	-0.43901	0.66065
Low|Medium	0.53768	0.01749	30.73584	0.00000
Medium|High	2.53879	0.01856	136.76321	0.00000

Several categorical predictors were found to be statistically significant at the $\alpha = 0.05$ level, including:

- CountryIndia, CountryJapan, and CountryUK
- EthnicityAsian, EthnicityCaucasian, EthnicityHispanic, and EthnicityMiddle Eastern
- Family_HistoryYes, Radiation_ExposureYes, and Iodine_DeficiencyYes

These predictors demonstrated strong associations with thyroid cancer risk levels. Other categorical variables such as Gender, Smoking, Obesity, and Diabetes were not statistically significant.

Notably, positive coefficients for Family_HistoryYes (1.06136), Radiation_ExposureYes (0.80127), and Iodine_DeficiencyYes (0.78911) indicate an increased risk of thyroid cancer for individuals with these factors. Negative coefficients observed for CountryUK (-0.06611) and some ethnicity categories suggest a comparatively lower risk relative to the baseline.

The model achieved a prediction accuracy of approximately 51.356%, indicating moderate ability to classify thyroid cancer risk levels using only categorical variables.

Table XI.
Confusion Matrix of Ordinal Logistic Regression Model Using Only Categorical Variables

Predicted \ Actual	Low	Medium	High
Low	78,173	52,389	2,759
Medium	29,666	19,674	17,761
High	549	337	11,383

PARAMETER ESTIMATION AND HYPOTHESIS TESTING FOR MODEL C

Model C applies the same ordinal logistic regression form but restricts the predictors to categorical variables. As with the previous models, the log-odds of being in a higher category are modeled as a linear function of the dummy-coded categorical variables:

$$\log\left(\frac{P(Y \le j)}{P(Y > j)}\right) = \theta_j + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where j=1,2 represents the thresholds for the three-category outcome (Low, Medium, High). Each X_k in this model represents a level of a categorical variable relative to a reference group. The estimated coefficients β_k indicate how belonging to a particular category affects the odds of being at a higher risk level compared to the reference.

Like the previous models, Model C employed maximum likelihood estimation to calculate parameter estimates for each categorical variable. These coefficients reflect the change in log-odds of being in a higher thyroid cancer risk category relative to the reference group for each predictor level. Hypothesis tests were conducted using z-values derived from the ratio of each estimate to its standard error. To assess the statistical relevance of the categorical predictors in Model C, the following hypotheses were evaluated for each variable:

To assess the statistical relevance of the categorical predictors in Model C, the following hypotheses were evaluated for each variable:

- Null hypothesis (H_0): The categorical predictor has no effect on thyroid cancer risk.
- Alternative hypothesis (H_1): The categorical predictor has a significant effect on thyroid cancer risk.

Model C focused exclusively on categorical predictors. The estimation process yielded several significant variables ($p < 0.05$), consistent with the results from Model A.
Significant predictors included:

- Family_HistoryYes ($p < 0.05$)
- Radiation_ExposureYes ($p < 0.05$)
- Iodine_DeficiencyYes ($p < 0.05$)
- CountryIndia ($p < 0.05$)
- EthnicityAsian ($p < 0.05$)

Non-significant variables included Gender, Smoking, Obesity, and Diabetes.

These results confirm the robustness of categorical predictors across models and further support their use in simplified, interpretable risk assessment frameworks.

2) Model Comparison

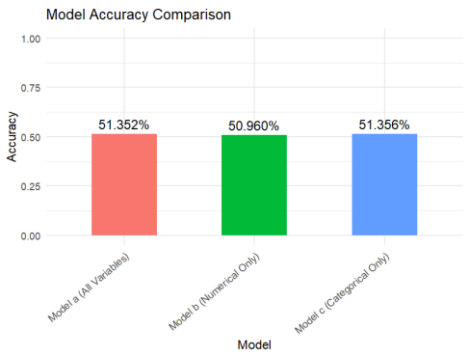

Fig. 12. Model Accuracy Comparison

The performance comparison of the three ordinal logistic regression models indicates that Model C, which utilizes only categorical predictor variables, achieves a comparable accuracy of 51.356% relative to Model A 51.352% that employs all available variables. Model B, based solely on numerical variables, yielded a slightly lower accuracy of 50.96%.

Importantly, Model C attains this level of accuracy while relying on fewer predictor variables, thereby reducing model complexity and enhancing interpretability. This parsimonious approach facilitates easier implementation and may offer practical advantages in clinical screening environments where categorical information such as demographics and exposure history is readily available.

Consequently, Model C is recommended as the preferred model for early thyroid cancer risk prediction, as it effectively balances predictive performance with simplicity and operational efficiency.

III. CLOSING

A. Summary

This study presents a comprehensive analysis of thyroid cancer risk using an ordinal logistic regression model implemented in R. The objective was to classify individuals into low, medium, or high-risk categories based on a combination of demographic, lifestyle, and physiological factors. The dataset, sourced from Kaggle, contains 212,691 entries and 16 variables, including patient age, gender, ethnicity, country, family history, exposure to radiation, iodine deficiency, and thyroid hormone levels (TSH, T3, T4), as well as nodule size.

Three versions of the ordinal logistic regression model were developed:

- Model A: Used all available variables (categorical and numerical)
- Model B: Used only numerical predictors (e.g., hormone levels and age)
- Model C: Used only categorical predictors (e.g., ethnicity, country, lifestyle factors)

Statistical testing through Chi-square and Spearman correlation revealed that categorical variables, particularly family history, radiation exposure, iodine deficiency, country, and ethnicity, had significant associations with thyroid cancer risk levels. In contrast, physiological measurements such as TSH, T3, and T4 levels showed no statistically significant correlation with risk.

Model performance was evaluated using confusion matrices and classification accuracy. Model C, which used only categorical variables, achieved an accuracy of 51.356%, comparable to Model A with the accuracy of 51.352% that used all predictors. This indicates that categorical features alone provide substantial predictive power, reducing model complexity while maintaining interpretability and performance.
Key insights from the model include:

- Individuals with a family history of thyroid disease, radiation exposure, or iodine deficiency are more likely to fall into the high-risk group.
- Asian ethnicity and origin from India were positively associated with higher risk levels.
- Other ethnic groups such as Caucasian, Hispanic, and Middle Eastern were more likely to be classified as low risk.
- Age and hormone levels had limited predictive value in distinguishing between risk categories.

B. Conclusion

The application of an ordinal logistic regression model to thyroid cancer risk classification has demonstrated that categorical variables, especially family history, radiation exposure, iodine deficiency, ethnicity, and country, are strong

indicators of increased cancer risk. The model successfully categorized patients into low, medium, and high-risk groups, providing useful insights for early screening and population-level risk assessment.

Statistical testing confirmed that variables such as family history, radiation exposure, iodine deficiency, and ethnicity are significantly associated with higher thyroid cancer risk levels ($p < 0.05$), while age and hormone levels were not predictive in this context.

Among the models tested, the one utilizing only categorical variables offered the best balance between simplicity, interpretability, and predictive performance, achieving a classification accuracy of 51.356%. The findings suggest that a categorical variable based approach can be efficiently integrated into public health strategies, particularly in regions with limited access to detailed physiological data or diagnostic tools.

However, the modest accuracy levels highlight the challenges posed by overlapping risk factor distributions and imbalanced class sizes in the dataset. This indicates the need for future research to explore more sophisticated machine learning models (e.g., Random Forest, XGBoost), perform feature selection or engineering, and apply resampling techniques to improve accuracy, especially in distinguishing between medium and high-risk cases.

Ultimately, this research contributes to the growing field of predictive healthcare analytics and supports the development of data-driven tools for early cancer risk identification, which can inform both individual-level clinical decisions and broader health policy interventions.

REFERENCES

[1] Penn Medicine, “Thyroid Cancer,” PennMedicine.org, 2024. [Online]. Available: https://www.pennmedicine.org/conditions/thyroid-cancer

[2] American Cancer Society, "Thyroid Cancer: Risk Factors," 2023. [Online]. Available: https://www.cancer.org/cancer/types/thyroid-cancer/causes-risks-prevention/risk-factors.html

[3] Canadian Cancer Society, "Thyroid cancer risk factors," 2023. [Online]. Available: https://cancer.ca/en/cancer-information/cancer-types/thyroid/risks

[4] H. L. Fontham, H. Sung, J. Ferlay, M. Laversanne, R. L. Siegel, and F. Bray, “Global cancer statistics 2022: GLOBOCAN estimates of cancer incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, vol. 74, no. 2, pp. 139–165, 2024, doi:10.3322/caac.21839. [Online]. Available: https://doi.org/10.3322/caac.21839

[5] SpringerLink, "Advances in Thyroid Cancer Diagnostics and Imaging," 2024. [Online]. Available: https://link.springer.com/article/10.1007/s12672-024-01017-w

[6] BMC Public Health, "Thyroid cancer burden and healthcare disparities in Indonesia," 2025. [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-025-21960-9

[7] Cancer Network, "Taking a Multidisciplinary Approach to Thyroid Cancer Treatment," 2024. [Online]. Available: https://www.cancernetwork.com/view/taking-a-multidisciplinary-approach-to-thyroid-cancer-treatment

[8] A. Forma, M. Di Donato, C. De Meis, G. Senese, G. Teti, A. Albasini, and A. Panno, “Thyroid cancer: Epidemiology, classification, risk factors, diagnostic and prognostic markers, and current treatment strategies,” *International Journal of Molecular Sciences*, vol. 26, no. 11, p. 5173, 2023. [Online]. Available: https://www.mdpi.com/1422-0067/26/11/5173

[9] Frontiers in Endocrinology, "Application of machine learning and statistical models for thyroid cancer risk stratification," 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2024.1366687/full

[10] Z. Wang, X. Wang, T. Wang, J. Qiu, and W. Lu, “Localization and Risk Stratification of Thyroid Nodules in Ultrasound Images Through Deep Learning”, Ultrasound in Medicine and Biology, vol. 50, no. 6, p882-887, Jun. 2024. [Online]. Available: https://www.umbjournal.org/article/S0301-5629(24)00112-1/abstract

[11] H. A. A. S. Husna, “Ordinal Logistic Regression Analysis on Factors Affecting Thyroid Cancer Stages,” Bachelor's thesis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, 2020. [Online]. Available: https://repository.its.ac.id/81028/

[12] F. M. Girardi, L. M. da Silva, and C. D. Flores, “A predictive model to distinguish malignant and benign thyroid nodules based on age, gender and ultrasonographic features,” National Library of Medicine, Nov. 2017, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9442819/

[13] M. D. Tronko et al., "A cohort study of thyroid cancer and other thyroid diseases after the chornobyl accident: thyroid cancer in Ukraine detected during first screening," vol. 98, no. 13, pp. 897-903, Jul. 2006. https://pubmed.ncbi.nlm.nih.gov/16818853/

[14] E. Kebebew., “Hereditary non-medullary thyroid cancer: Genetic background and clinical implications,” Endocrine-Related Cancer, vol. 32, no. 5, pp. 678-82, 2008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18058169/

[15] C. M. Kitahara et al., “Cigarette smoking, alcohol intake and thyroid cancer risk: a pooled analysis of prospective studies,” PubMed Central, vol. 23, no. 10, pp. 1615–1624, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3511822/

[16] J. Ma et al., “Obesity and Risk of Thyroid Cancer: Evidence from a Meta-Analysis of 21 Observational Studies,” PubMed Central, vol. 30, no. 10, pp. 1385–1396, Oct. 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4315628/

[17] E. Fiore, “Lower levels of TSH are associated with a lower risk of papillary thyroid cancer in patients with thyroid nodular disease: Thyroid autonomy may play a protective role,” National Library of Medicine, vol. 16, no. 4, pp. 1251–60, Dec. 2009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19528244/

[18] E. Sunyi, “A Comparative Analysis of Machine Learning Algorithms for Thyroid Nodule Malignancy Prediction,” Journal of Thyroid Disorders & Therapy, vol. 13, no. 2, 2024. [Online]. Available: https://www.longdom.org/open-access/a-comparative-analysis-of-machine-learning-algorithms-for-thyroid-nodule-malignancy-prediction-108505.html

[19] S. Khodabandelu, et al., “Development of a Machine Learning-Based Screening Method for Thyroid Nodules Classification by Solving the Imbalance Challenge in Thyroid Nodules Data,” National Library of Medicine, vol. 22, no. 3, Oct. 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36511373/

[20] M. Z. Zeeshan, "Thyroid Cancer Risk Dataset," Kaggle, Feb. 2025. [Online]. Available: https://www.kaggle.com/datasets/mzohaibzeeshan/thyroid-cancer-risk-dataset

[21] P. Li, et al., “Prediction models constructed for Hashimoto’s thyroiditis risk based on clinical and laboratory factors,” Frontiers in Endocrinology, vol. 13, Aug. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2022.886953/full

[22] X. Zhang, Y. Wang, J. Wang, Z. Zhang, and Y. Li, “Machine learning–based dynamic prediction of lateral lymph node metastasis in patients with papillary thyroid cancer,” Frontiers in Endocrinology, vol. 13, Oct. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fendo.2022.1019037/full