PREDICTING THYROID CANCER RISK USING MACHINE LEARNING

By Group 2

Tishani Wijekoon(S16379), Chami Sewwandi(S16028), W.K.Hiruni Hasara(S16210), S.Luxan(s16329)

ABSTRACT

Thyroid cancer is the most prevalent endocrine malignancy, and early detection plays a critical role in improving patient outcomes. This study explores the use of machine learning techniques for non-invasive, preoperative risk stratification of thyroid malignancy using basic demographic and clinical information. A comprehensive dataset of over 210,000 patient records was analyzed, containing both categorical and numerical variables related to thyroid cancer risk factors. After data cleaning, encoding, and standardization, various feature selection methods—such as Lasso regularization, statistical tests, and dimensionality reduction (FAMD)—were employed to identify influential predictors.

The study addressed class imbalance through SMOTE and trained multiple classification models, including Logistic Regression, XGBoost, and others. The Lasso-based logistic regression model identified *Country*, *Ethnicity*, *Family History*, *Radiation Exposure*, and *Iodine Deficiency* as the most predictive features. Among models tested, moderate classification performance was observed, with the highest AUC reaching 0.6699. Hierarchical clustering using FAMD-reduced data revealed complex patterns in patient profiles, hinting at latent structures in risk factor interactions.

While the models show promise in assisting early diagnostic decisions, the results also underscore the complexity of thyroid cancer prediction based solely on basic patient information. Future research may benefit from integrating additional clinical biomarkers or imaging data to enhance predictive power. Overall, this work highlights the potential of machine learning as a supportive tool in thyroid cancer risk stratification, aiming to reduce diagnostic delays and improve healthcare resource allocation.

Table of Contents

List of Figures

LIST OF TABLES

INTRODUCTION

Thyroid cancer is the most common malignancy of the endocrine system, with incidence rates steadily increasing worldwide. While the disease generally has a favorable prognosis when detected early, delays or inaccuracies in diagnosis can significantly impact treatment outcomes and long-term survival. Various demographic, environmental, and genetic factors contribute to the complexity of thyroid cancer, making early and accurate identification of malignant cases a critical public health concern.

The current clinical approach to diagnosing thyroid cancer involves a combination of physical examinations, thyroid function tests, imaging modalities such as ultrasound, and ultimately, fine-needle aspiration biopsy (FNAB). FNAB remains the gold standard for definitive diagnosis, providing cytological confirmation of malignancy. However, this method is invasive, costly, and can occasionally produce indeterminate or false-negative results. Moreover, the increasing number of thyroid nodules detected through imaging has led to concerns about overtreatment and unnecessary procedures, particularly when many nodules turn out to be benign.

Given these limitations, there is growing interest in developing predictive models that can assist clinicians in identifying high-risk patients using non-invasive, preoperative information. Machine learning techniques offer promising tools to analyze large, complex datasets and uncover patterns that may not be evident through traditional methods. By building predictive models based on routinely collected data, it is possible to support early risk assessment, reduce reliance on invasive procedures, and prioritize diagnostic resources more effectively.

THE QUESTION WE ARE GOING TO ANSWER

“Can we predict whether a thyroid tumor is cancerous or not using basic patient information?”

Right now, doctors use tests like scans and biopsies to diagnose thyroid cancer, but these can be expensive, take time, and may not always give clear answers. Our goal is to investigate whether machine learning models trained on patient data, can help doctors make faster and more accurate decisions, and maybe even reduce the number of tests some patients need.

DATA SET

Table 1- Data set description

Categorical	
Variable	*Description*
Patient_ID	Unique identifier for each patient
Gender	Patient’s gender (Male/Female)
Country	Country of residence.
Ethnicity	Patient’s ethnic background.
Family_History	Whether the patient has a family history of thyroid cancer (Yes/No)
Radiation_Exposure	Presence of iodine deficiency (Yes/No)
Iodine_Deficiency	Presence of iodine deficiency (Yes/No)
Smoking	Whether the patient smokes (Yes/No)
Obesity	Whether the patient is obese (Yes/No)
Diabetes	Whether the patient has diabetes (Yes/No)
Thyroid_Cancer_Risk	Estimated risk of thyroid cancer (Low/Medium/High
Diagnosis	Final diagnosis (Benign/Malignant)

Numerical	
Variable	*Description*
Age	Age of the patient
TSH_level	Thyroid-Stimulating Hormone level (µIU/mL)
T3_level	Triiodothyronine level (ng/dL)
T4_level	Thyroxine level (µg/dL)
Nodule_Size	Size of thyroid nodules (cm)

DATA PRE-PROCESSING AND FEATURE SELECTION

- Missing Value Handling - The dataset was free of missing values
- Categorical Encoding:
  - One-Hot Encoding - Applied to multi-class categorical features such as Country and Ethnicity.
  - Binary Mapping:
    - Gender: Male = 1, Female = 0
    - Family_History, Radiation_Exposure, Iodine_Deficiency, Smoking, Obesity, Diabetes: Yes = 1, No = 0
    - Diagnosis (target): Malignant = 1, Benign = 0
- Column Removal:
    - Patient_ID: Removed as it is an identifier with no predictive value.
    - Thyroid_Cancer_Risk: Excluded to prevent data leakage, as it may contain direct information about the target variable.

- Feature Scaling - All continuous and encoded numerical features were standardized using Z-score normalization.
- Train-Test Split: The dataset was partitioned into:
  - Training Set: 170,152 samples (80%)
  - Test Set: 42,539 samples (20%)
- Applied SMOTE (Synthetic Minority Over-sampling Technique) to the training set to handle imbalanced classification targets.

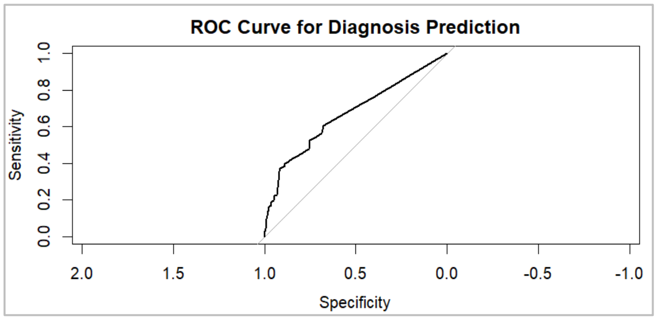

Figure 1-Feature selection evaluation (ROC curve)

- **<u>FEATURE SELECTION:</u>**

— Lasso Regularization (L1 penalty):
A logistic regression model with Lasso regularization was fitted using 10-fold cross-validation to determine the optimal lambda penalty value (0.0084).

— Selected Features:
Country, Ethnicity, Family_History, Radiation_Exposure, Iodine_Deficiency. These features were identified as the most strongly associated with the likelihood of thyroid cancer malignancy and were included in subsequent model training.

— Evaluation:

I. ROC Curve and AUC:
The model achieved an AUC of 0.6699, indicating a moderate ability to distinguish between malignant and benign cases.

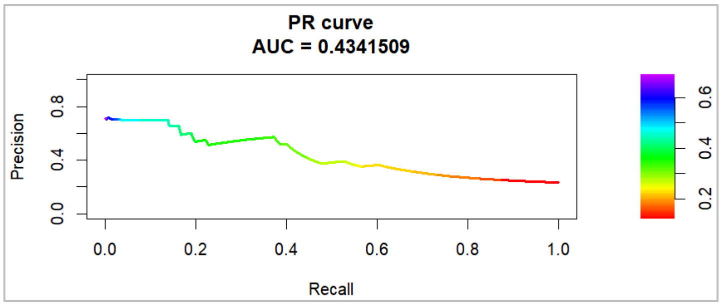

Figure 2-Feature selection evaluation (PR curve)

II. Mann-Whitney U test for continuous variables.

Table 2-Feature selection evaluation (Mann-Whitney U test)

Feature	p-value
Age	0.9485
TSH_Level	0.2050
T4_Level	0.3644
T3_Level	0.2720
Nodule_Size	0.4168

III. Chi-squared test exact test for categorical variables.

Table 3-Feature selection evaluation (Chi-squared test)

Feature	p-value	Feature	p-value
Gender	0.2233	Radiation_Exposure	0.0000
Country	0.0000	Iodine_Deficiency	0.0000
Ethnicity	0.0000	Smoking	0.7272
Family_History	0.0000	Obesity	0.4285

- Despite feature selection, all variables were retained for model fitting to ensure no valuable information was discarded.

- **FAMD ANALYSIS;**

— Since PCA is not ideal for categorical variables, FAMD (Factor Analysis of Mixed Data) was used to reduce dimensionality while preserving both categorical and continuous variable information.

— Due to computational limitations, only 5 components were retained.

— Hierarchical clustering was performed on 10,000 randomly selected samples using these components.

— The first five FAMD components each contributed similarly, with Component 1 explaining ~4.32%.

— Together, they explained 21.22% of the total variance.

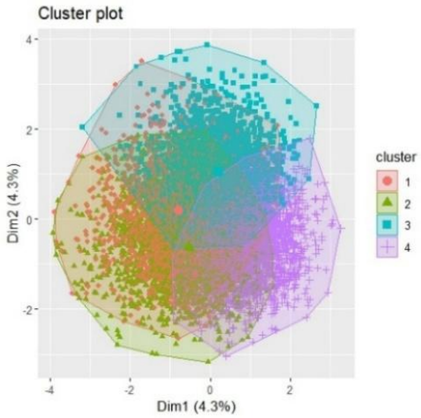

Figure 3-Cluster plot

- **CLUSTERING**

Table 4-Summary of clusters

Cluster	Age Group	Gender Dominance	Major Countries	Dominant Ethnicities	Key Health Traits
1	Older	Males (55.8%)	South Korea, India, UK, China	Hispanic, Middle Eastern, Asian (84.8%)	Highest smoking rate (28%)
2	Younger	Females (59.9%)	Nigeria, Germany, China, India	Asian, Caucasian, African (86.8%)	Highest iodine deficiency (32.6%) Highest diabetes rate (43.5%)
3	Older	Females (70.9%)	China, Japan, India, USA	Caucasian, African, Asian (86.5%)	Highest thyroid cancer family history (47%) High radiation exposure (25.5%)
4	Younger to Middle-aged	Females (60.8%)	India, Russia, Brazil, China	Caucasian, Asian, African (91.7%)	Highest obesity rate (42.8%) High radiation exposure (14.5%)

- Hierarchical clustering on the FAMD-reduced data produced four overlapping clusters, suggesting complexity in variable interactions and a potential need for more data for better separation.

IMPORTANT RESULTS OF THE EDA

1. DATA QUALITY

Initial inspection of the dataset revealed high data quality. There were no missing or duplicate values, and outlier detection using the interquartile range (IQR) method did not identify any significant outliers in the numeric features.

2. CLASS DISTRIBUTION

The response variable, *Diagnosis* (Benign vs. Malignant), exhibited a notable class imbalance. Specifically, 76.7% of the cases were benign, while only 23.3% were malignant.

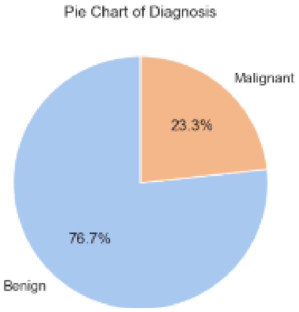

Figure 4-Pie chart of diagnosis

3. FEATURE DISTRIBUTIONS

Boxplots and descriptive statistics indicated that all numeric features were symmetrically distributed without significant skewness or extreme values. This suggests that the dataset is well-behaved and conducive to most statistical modeling techniques. And there is a moderate gender imbalance, with 60.1% of the cases being female and 39.9% male.

4. BIVARIATE ASSOCIATIONS WITH DIAGNOSIS

Several features demonstrated meaningful associations with the target variable:

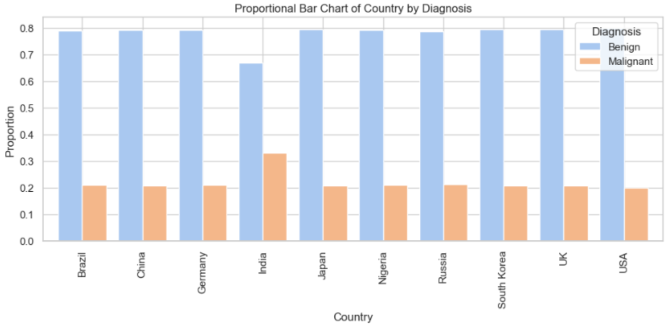

Figure 5-Proportional bar chart of country by diagnosis

— T4 Level: Malignant cases tended to have slightly lower median T4 levels compared to benign cases.

— Country: A higher proportion of malignant cases was observed among patients from India, indicating potential environmental or systemic factors.

— Ethnicity: Malignancy rates were highest among Asian individuals, pointing toward potential genetic or cultural influences.

— Family History: A positive family history showed a strong association with malignancy.

— Iodine Deficiency & Radiation Exposure: Both risk factors were more prevalent in malignant cases, aligning with established thyroid cancer risk literature.

5. CORRELATION ANALYSIS

Correlation analysis among the numeric features revealed generally weak linear relationships, indicating low multicollinearity. This is advantageous for model interpretability and suggests that the features may contribute independently to the prediction task.

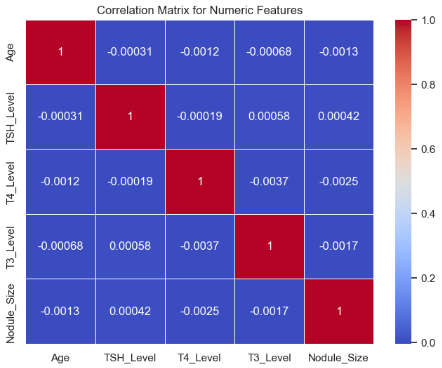

Figure 6-Correlation matrix

IMPORTANT RESULTS OF THE ADVANCED ANALYSIS

1. MODEL PERFORMANCE COMPARISON

Table 5-Model performance comparison

Model	Dataset	Precision (Malignant)	Recall (Malignant)	F1-Score (Malignant)	Accuracy
Logistic Regression	Train set	0.73	0.68	0.70	0.71
	Test set	0.33	0.41	0.36	0.67
Random Forest	Train set	1.00	1.00	1.00	1.00
	Test set	0.54	0.44	0.48	0.78
XGBoost	Train set	0.87	0.69	0.77	0.79
	Test set	**0.56**	**0.47**	**0.51**	**0.79**
SVM	Train set	0.73	0.68	0.70	0.71
	Test set	0.33	0.40	0.36	0.67
Naive Bayes	Train set	0.60	0.63	0.62	0.62
	Test set	0.66	0.31	0.42	0.58
Decision Tree	Train set	1.00	1.00	1.00	1.00
	Test set	0.34	0.47	0.39	0.66

2. MODEL-SPECIFIC ANALYSIS

To establish a reliable and interpretable thyroid cancer prediction model, multiple supervised learning algorithms were implemented and rigorously evaluated. Logistic regression was first employed as a baseline model. Although its performance on the training set was modest (F1-score = 0.70), it significantly underperformed on the test set (F1-score = 0.36), particularly in identifying malignant cases, indicating limitations in capturing complex, non-linear relationships within the data.

Random Forest (RF) and XGBoost models were subsequently tuned using GridSearchCV to optimize performance. The RF model achieved an excellent training performance with an F1-score of 1.00, but its test F1-score dropped to 0.48, revealing overfitting concerns. In contrast, the XGBoost model, with best-tuned parameters, exhibited better generalization, achieving a balanced F1-score of 0.51 on the test set and was deemed the most effective ensemble model.

The Support Vector Machine (SVM) model with a linear kernel was trained on a representative subset of 5,000 observations to manage computational complexity. This shows moderate performance with an accuracy of 0.67, with low values of precision (0.33) and F1 score(0.36) suggesting its struggle with the minority class.

Naïve Bayes classifiers were also evaluated, particularly after applying ROSE resampling to address class imbalance. The tuned full model, utilizing kernel density estimation, demonstrated improved performance over the base model, achieving a test F1-score of 0.43.

Feature importance analysis using the XGBoost model identified several variables as key predictors of thyroid malignancy. Country and Ethnicity emerged as significant demographic factors, potentially reflecting geographical disparities in diagnosis influenced by environmental exposure, healthcare access, or dietary habits. Additionally, Family History, Iodine Deficiency, and Diabetes were among the most critical clinical predictors, aligning well with established medical knowledge regarding genetic predisposition, micronutrient imbalance, and metabolic disorders.

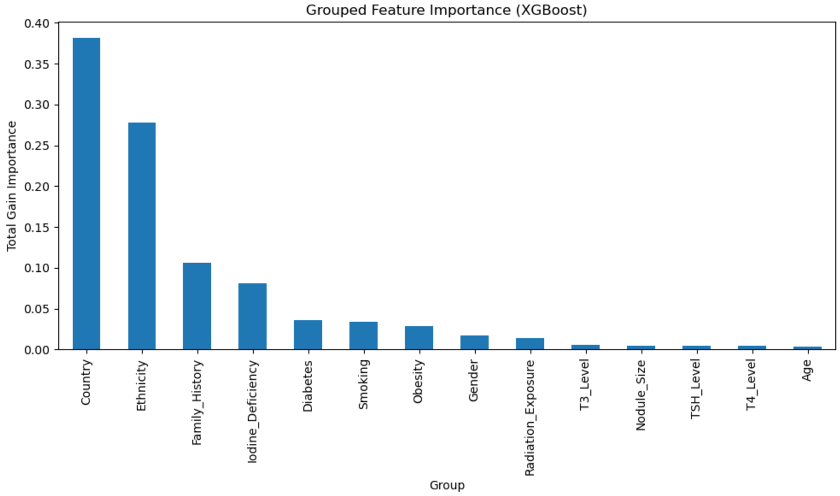

Figure 7- Feature importance (XGBoost)

DISCUSSION AND CONCLUSIONS

This study evaluated multiple supervised learning models to predict thyroid cancer using demographic and clinical variables. Simpler models like Logistic Regression and SVM performed poorly on the test set, especially in identifying malignant cases. In contrast, XGBoost outperformed other models by achieving a better balance between sensitivity and generalization, making it the most effective model overall. Although Random Forest performed well on the training data, its test performance dropped, indicating overfitting.

Key predictors identified by XGBoost included **Country**, **Ethnicity**, **Family History**, **Iodine Deficiency**, and **Diabetes**—all medically relevant risk factors. Naïve Bayes showed slight improvement after handling class imbalance with ROSE sampling but still struggled with accuracy.

To explore the data structure, FAMD was applied for dimensionality reduction across mixed data types. The first five components explained about 21% of the variance. Hierarchical clustering based on these components revealed possible hidden patterns or subgroups within the data.

In conclusion, ensemble models, especially XGBoost offer promising results for thyroid cancer prediction. Further improvements could come from better exploring clusters identified through unsupervised techniques.

APPENDIX

R Code

BIBLIOGRAPHY

1. Setiawan, K. E. (2024). Predicting Recurrence in Differentiated Thyroid Cancer: A Comparative Analysis of Various Machine Learning Models Including Ensemble Methods with Chi-Squared Feature Selection. *Commun. Math. Biol. Neurosci.*, 2024, 2024:55. https://doi.org/10.28919/cmbn/8506. ISSN: 2052-2541.

2. 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, 557–570. https://doi.org/10.3390/knowledge4040029.

3. Salman, K., & Sonuç, E. (2021). Comparative Analysis of Machine Learning Models for Thyroid Cancer Recurrence Prediction. *J. Phys.: Conf. Ser.*, 1963, 012140.

4. Iqbal, & Shahzad. (2024). *PJNMed*, 49–55. DOI: 10.24911/PJNMed.175-1721068107.

5. Aggarwal, A., Gotti, M., Kaur, E., & Lu, S. (2024). Comparative Analysis of Machine Learning Models for Thyroid Cancer Recurrence Prediction. *Indian Journal of Public Health Research & Development*, 11(03), March 2020.

6. https://www.mdpi.com/2673-9585/4/4/29

7. https://www.thyroid.org/patient-thyroid-information/ct-for-patients/april-2024/vol-17-issue-4-p-11-12/

8. https://www.sciencedirect.com/science/article/pii/S0720048X25001354

9. https://www.researchgate.net/publication/385996218_Predictive_Analytics_for_Thyroid_Cancer_Recurrence_A_Machine_Learning_Approach

10. https://medium.com/@miramnair/feature-selection-mutual-information-a0def943e1ed

11. https://cancer.ca/en/cancer-information/cancer-types/thyroid/risks#:~:text=A%20history%20of%20non%2Dcancerous,the%20thyroid%20(called%20thyroiditis).
12. https://www.r-bloggers.com/2019/06/running-umap-for-data-visualisation-in-r/
13. https://rpubs.com/pjmurphy/758265
14. https://rpubs.com/nchelaru/famd