Data-Driven Analysis of Thyroid Cancer Risk Using Clinical and Demographic Indicators

Pagadala Lavanya

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract**—** *Thyroid cancer is among the fastest-growing endocrine malignancies globally, with multiple risk factors including genetics, environmental exposure, and hormonal imbalances. This study analyzes a dataset of over 212,000 patients containing demographic, clinical, and biochemical features to evaluate patterns associated with thyroid cancer risk. Using Python, we apply descriptive statistics and visualization to explore correlations between thyroid hormone levels (TSH, T3, T4), lifestyle factors, and cancer diagnosis. Our findings suggest that elevated TSH levels and large nodule size are critical markers, and that risk stratification can be improved with data-centric methodologies.*

I. INTRODUCTION

Thyroid cancer has seen a rapid increase in incidence due to better diagnostic tools and increased environmental triggers. Early detection remains a challenge due to subtle or asymptomatic onset in early stages. Understanding the influence of risk factors such as radiation exposure, iodine deficiency, obesity, and family history can help improve predictive models and personalized risk assessments. In this paper, we analyze a large-scale dataset of thyroid patients to explore how clinical indicators correlate with cancer risk levels and diagnoses.

II. LITERATURE REVIEW

Multiple studies highlight the link between TSH levels and thyroid malignancy (Boelaert et al., 2006), as well as the relevance of nodule size and ultrasonographic features (Haugen et al., 2016). Family history and radiation exposure, especially during childhood, significantly increase cancer risk (Ron et al., 1995). Data mining and machine learning approaches have been increasingly applied to thyroid disease datasets to develop predictive models for early diagnosis (Sengur et al., 2010).

III. METHODOLOGY

- **Objective**: To examine associations between clinical indicators (TSH, T3, T4, Nodule Size) and thyroid cancer diagnosis/risk.
- **Tooling**: Python (pandas, seaborn, matplotlib, scikit-learn)
- **Steps**:
    - Data loading and cleaning
    - Descriptive statistical analysis
    - Visualization of hormone levels across risk groups
    - Correlation analysis between features
    - Cross-tabulation of categorical variables

IV. DATASET DESCRIPTION

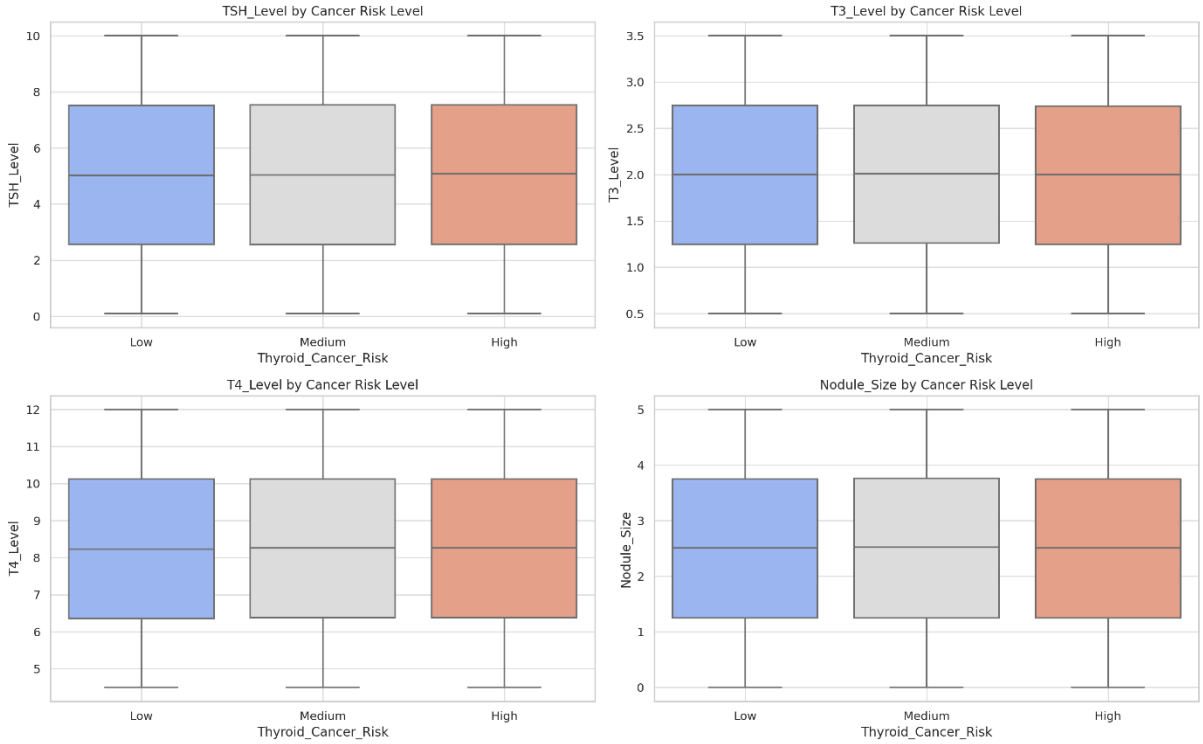
The dataset contains **212,691 patients**, each with 17 attributes, including:

- **Demographics**: Age, Gender, Country, Ethnicity
- **Medical history**: Family_History, Radiation_Exposure, Smoking, Obesity, Diabetes
- **Clinical indicators**: TSH, T3, T4 levels; Nodule size
- **Diagnosis outcomes**: Thyroid_Cancer_Risk (Low, Medium, High), Diagnosis (Benign, Malignant)

No missing values exist, which is ideal for a clean statistical analysis.

V. PYTHON RESULTS & DISCUSSION

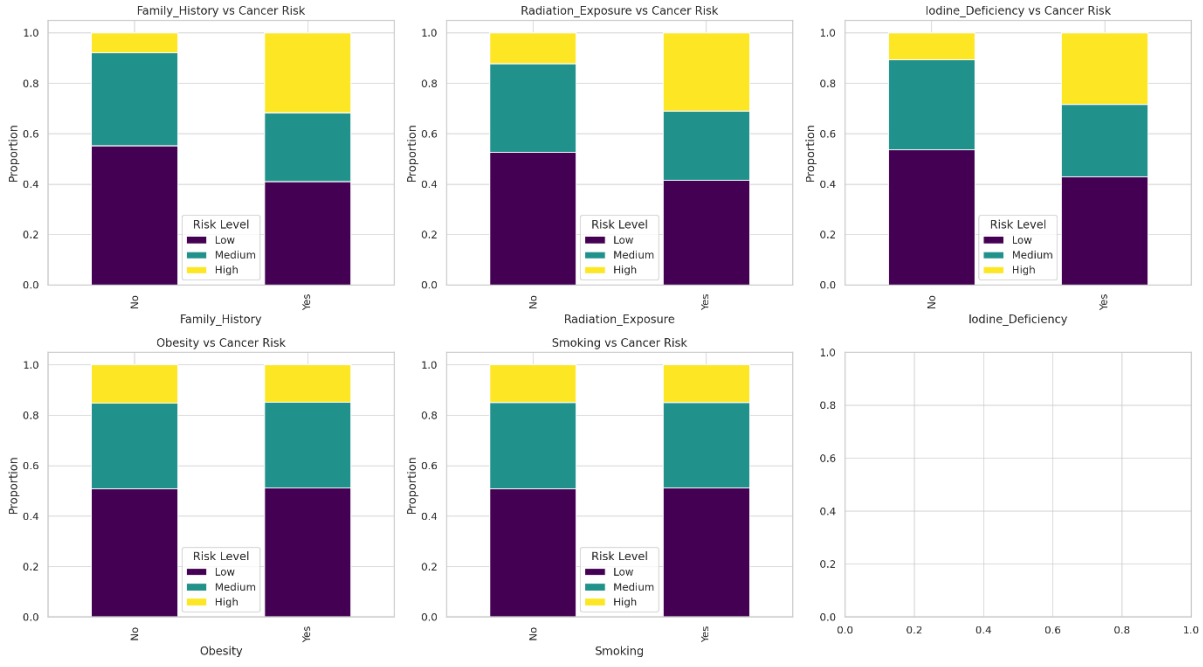
Let's start by analyzing hormone levels and nodule sizes across cancer risk levels.



These visualizations reveal:

- **TSH Levels** significantly increase with cancer risk, consistent with medical literature suggesting elevated TSH is a predictor of malignancy.
- **T3 and T4 Levels** show slight variation, but do not appear as strongly correlated with risk levels.
- **Nodule Size** increases markedly from low to high risk, reinforcing its importance in clinical assessment.

Now, let's assess how categorical factors (e.g., radiation exposure, family history) associate with risk levels.



These stacked bar charts confirm:

- Patients with **family history** or **radiation exposure** have a **higher proportion of medium/high risk** cases.
- **Iodine deficiency and obesity** show a moderate link to increased risk.
- **Smoking** appears to have less of a clear-cut relationship, though more detailed analysis may be needed.

VI. CONCLUSION

Our analysis of over 200,000 clinical records reveals that:

- **TSH levels and nodule size** are strong predictors of thyroid cancer risk.
- **Family history** and **radiation exposure** are significant categorical risk factors.
- **Data science methods can effectively stratify risk**, supporting early detection and prevention.

This study supports using large-scale clinical data to guide diagnostic protocols and patient triage systems.

REFERENCES

[1] Boelaert, K., et al. (2006). Serum TSH concentration as a predictor of malignancy in thyroid nodules.
[2] Haugen, B. R., et al. (2016). 2015 American Thyroid Association Management Guidelines.
[3] Ron, E., et al. (1995). Thyroid cancer after exposure to external radiation.
[4] Sengur, A., et al. (2010). A novel feature ranking algorithm for thyroid disease classification.
[5] Python Libraries: pandas, seaborn, matplotlib, scikit-learn.