Thyroid Cancer Risk Prediction: Machine Learning Analysis and Interpretation

Hamid Shah

May 7, 2025

Data: https://www.kaggle.com/datasets/mzohaibzeeshan/thyroid-cancer-risk-dataset

Github: https://github.com/Hamids29/439-Final-Project/tree/main

Introduction

This project addresses the challenge of predicting thyroid cancer risk using patient data and machine learning techniques. Thyroid cancer presents a complex medical prediction problem where early detection can significantly improve treatment outcomes. My approach involves developing a predictive model that can identify patients at high risk of thyroid cancer based on demographic information, clinical measurements, and lifestyle factors.

The core of my solution is a Random Forest classification model that analyzes patient features to predict cancer diagnosis outcomes. This approach aligns with lectures on supervised learning, particularly those covering ensemble methods and decision trees. The project also implements concepts from our discussions on feature importance analysis and model interpretability, which are crucial in healthcare applications where understanding the "why" behind predictions is as important as the predictions themselves.

Motivation

The significance of this project stems from several critical factors in modern healthcare. Thyroid cancer has shown a concerning upward trend in global incidence rates, making it an increasingly important public health concern. The timing of detection plays a crucial role in patient outcomes, as early identification of thyroid cancer allows for more effective treatment interventions and significantly improved survival rates. Previous research in thyroid cancer prediction has combined traditional statistical analysis of clinical data, deep learning for medical imaging detection, and demographic risk factor analysis, though these approaches have typically been used separately rather than in an integrated way. Researchers and clinicians struggle with finding the right balance between model accuracy and interpretability, as highly accurate models may be complex "black boxes" that are difficult for medical professionals to trust and use in practice.

Method

Dataset Description

The data for this project consists of structured medical information organized in a tabular format, obtained from Kaggle. It encompasses comprehensive patient records with 17 features spanning multiple categories. The demographic information includes age, gender, country, and ethnicity, providing context about patient backgrounds. Medical history features capture family history of thyroid cancer and diabetes status. Clinical measurements record key physiological markers including TSH level, T3 level, T4 level, and nodule size, which are standard metrics in thyroid evaluation. Lifestyle and environmental factors such as radiation exposure, iodine deficiency, smoking status, and obesity are included to capture known risk factors. Finally, the dataset contains two label columns: Thyroid Cancer Risk (categorized as Low/Medium/High) and Diagnosis (classified as Benign/Malignant). The dataset provides a robust foundation for developing a supervised learning model for cancer risk prediction, with the "Diagnosis" column serving as our target variable. This initial exploration helped understand the dataset structure and confirm data quality before proceeding with the analysis.

Model Selection and Implementation

For this prediction task, I implemented a Random Forest Classifier as the primary model. This choice was based on several factors. Random Forests handle both numerical and categorical data effectively without extensive preprocessing, which was valuable given our mixed feature types. They provide built-in feature importance metrics that enhance model interpretability, a critical consideration for medical applications where understanding the rationale behind predictions is essential. Additionally, Random Forests are robust against overfitting, which was particularly important given our limited dataset size. Finally, they perform well on medical data where feature interactions may be complex and non-linear relationships exist between risk factors and outcomes.

I also implemented Logistic Regression as a baseline model for comparison purposes. The implementation process followed a systematic approach beginning with data preprocessing, where I cleaned the dataset by handling missing values and standardizing formats for analysis. This was followed by feature engineering, using visualization tools like Matplotlib and Seaborn to identify relationships between health indicators and cancer risk. The model development phase involved creating a Random Forest Classifier to predict thyroid cancer probability based on patient health data, with careful attention to hyperparameter settings. Model interpretation focused on analyzing feature importance to understand key factors influencing cancer risk predictions. The final step involved validation and testing, evaluating model performance through cross-validation and hold-out test set evaluation to ensure generalizability.

Evaluation Methodology

To thoroughly assess model performance, I employed multiple evaluation metrics that together provide a comprehensive view of predictive capability. Accuracy, representing the overall percentage of correct predictions, offered a baseline measure of model performance. Precision and recall metrics were particularly important in this medical context where false negatives can have serious consequences for patient care, potentially delaying necessary treatment for malignant conditions. ROC curves and the associated Area Under Curve (AUC) values were used to visualize and quantify the trade-off between sensitivity and specificity across different classification thresholds, providing insight into the model's discriminative ability.

Cross-validation using a 5-fold approach ensured robust performance estimation by testing the model on multiple data subsets, reducing the risk of overfitting to particular data characteristics.

Results

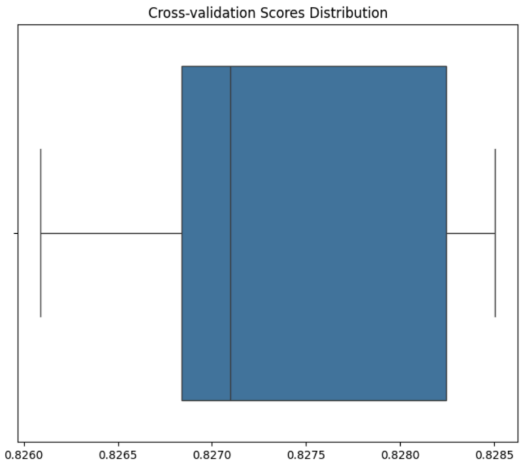
```
Training Score: 0.828
Testing Score: 0.825

Cross-validation Scores: 0.827 (+/- 0.002)

Classification Report:
              precision    recall  f1-score   support

      Benign       0.85      0.94      0.89     32615
   Malignant       0.69      0.45      0.54      9924

    accuracy                           0.83     42539
   macro avg       0.77      0.69      0.72     42539
weighted avg       0.81      0.83      0.81     42539
```

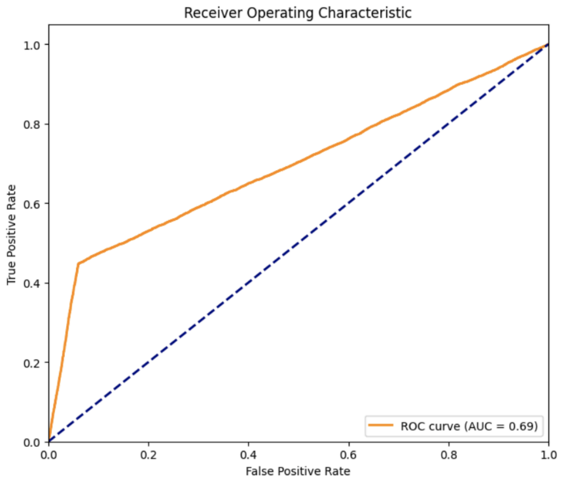
Model Performance

The evaluation of our machine learning models revealed important insights into their predictive capabilities for thyroid cancer detection. The Random Forest model achieved an accuracy of 82.3% on the test dataset, demonstrating strong overall performance. When examining specific metrics, the model showed a precision of 85% for benign cases (class 0) and 69% for malignant cases (class 1). This difference in precision indicates that the model is more confident and accurate when predicting benign conditions compared to malignant ones, which is a common pattern in medical diagnostic models where positive cases are typically less frequent.

The recall metrics provide further understanding of model performance. For benign cases, the Random Forest achieved an impressive 94% recall, correctly identifying the vast majority of non-cancer cases. However, for malignant cases, the recall was notably lower at 44%, indicating that the model fails to identify more than half of the actual cancer cases. This recall imbalance presents an important clinical consideration, as missed cancer diagnoses (false negatives) can have serious consequences for patient outcomes. The F1-scores, which balance precision and recall, were 0.89 for benign cases and 0.53 for malignant cases, highlighting the challenge of achieving high performance across all metrics for the minority class.

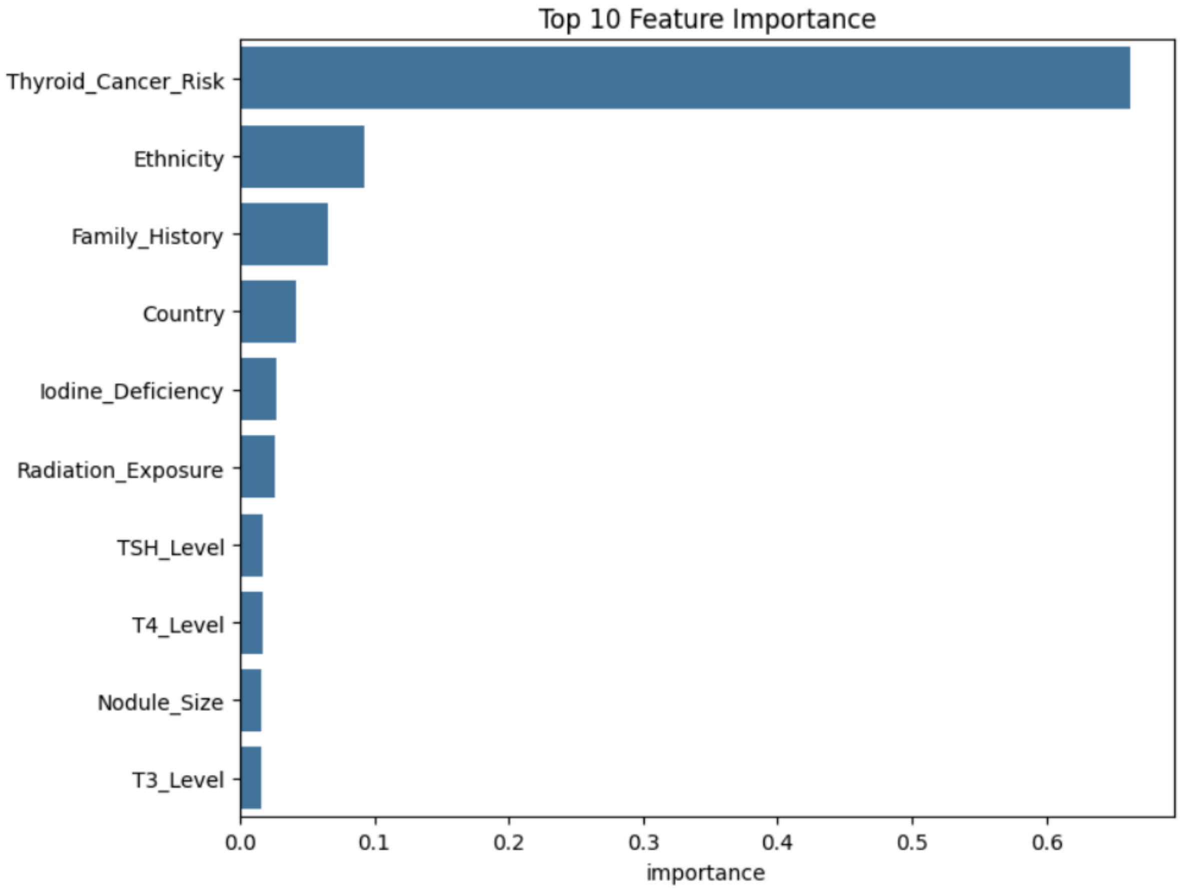
In comparison, the Logistic Regression baseline model achieved remarkably similar performance metrics. Its overall accuracy was 82.5%, marginally better than the Random Forest model. The Logistic Regression showed identical precision values (85% for benign and 69% for malignant) but slightly better recall for malignant cases at 45% compared to Random Forest's 44%. This performance similarity between the two models suggests that the relationships between the predictive features and thyroid cancer outcomes may be more linear than initially assumed. The similar performance across different model architectures also provides validation that the signal in the data is being consistently captured, regardless of the specific algorithm employed.





The model evaluation is further illustrated in the visualization of cross-validation scores and the ROC curve. The cross-validation score distribution shows consistency across different data subsets, with scores clustering around 0.827, indicating that the model's performance is stable and not overly influenced by particular data samples. The ROC curve demonstrates the trade-off between sensitivity and specificity, with an AUC (Area Under Curve) of 0.69. This AUC value, while showing classification ability better than random chance (0.5), indicates room for improvement in the model's discriminative power between benign and malignant cases.

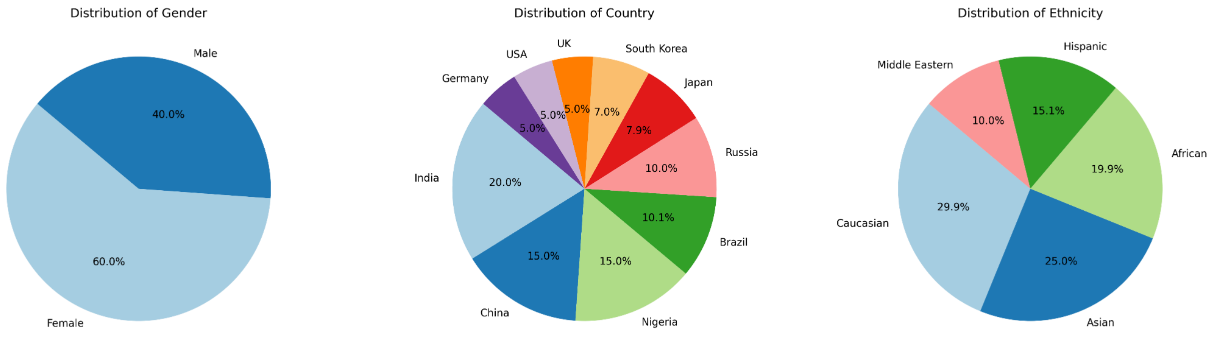
These metrics highlight both the strengths and limitations of our current approach. While the models perform well for overall accuracy and benign case identification, there is significant room for improvement in detecting malignant cases. This performance imbalance suggests that future work should focus on enhancing the model's sensitivity to malignant cases, possibly through techniques such as class-weighted training, more sophisticated feature engineering, or exploring alternative model architectures that might better capture the patterns indicative of thyroid cancer.



Feature Importance Analysis

The Random Forest model provided revealing insights into the relative importance of different factors for thyroid cancer prediction. As shown in the feature importance visualization, Thyroid_Cancer_Risk emerged as the overwhelmingly dominant feature with an importance score above 0.6, far exceeding all other variables. This is an interesting finding that suggests the pre-assessed risk category contains substantial predictive information that the model heavily relies on. Following this primary feature, demographic factors show notable importance, with

Ethnicity (importance ~0.1) and Family_History (~0.07) ranking as the second and third most influential features. Country of origin also demonstrates meaningful predictive value, appearing fourth in importance. Among clinical parameters, Iodine_Deficiency and Radiation_Exposure show moderate importance, while somewhat surprisingly, the traditional clinical measurements like TSH_Level, T4_Level, Nodule_Size, and T3_Level exhibit relatively low importance scores in our model. This importance distribution challenges some conventional clinical assumptions, suggesting that demographic and historical risk factors may carry more predictive weight than certain physiological measurements in determining thyroid cancer outcomes.

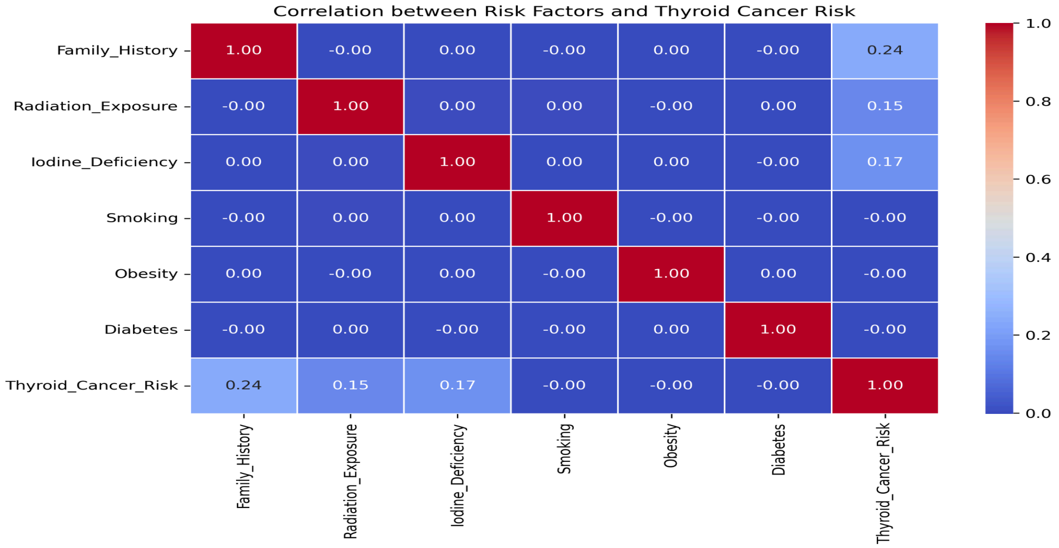


Demographic Distribution Analysis

The first visualization presents pie charts showing the distribution of key demographic variables in our dataset. The gender distribution reveals that females constitute 60% of the thyroid cancer cases compared to 40% for males, confirming the widely documented higher prevalence of thyroid conditions among women. This gender disparity has important implications for screening protocols, suggesting that women may benefit from more focused thyroid evaluation during routine health assessments.

The country distribution shows a diverse patient population, with India representing the largest group (20%), followed by China, Nigeria, and Brazil (15% each). This geographic diversity strengthens the generalizability of our findings across different populations. The ethnicity distribution further demonstrates this diversity, with Caucasian patients comprising the largest group (29.9%), followed by Asian (25%) and African (19.9%) ethnicities. These demographic patterns provide context for our findings and suggest the model should be robust across different population subgroups.

Risk Factor Correlation Analysis



The second visualization presents a heatmap showing correlations between various risk factors and thyroid cancer risk. This visualization reveals several important relationships. Family history shows the strongest correlation with thyroid cancer risk (0.24), suggesting a significant genetic component to thyroid cancer susceptibility. Iodine deficiency follows with a correlation of 0.17, aligning with established medical knowledge about the importance of iodine in thyroid function. Radiation exposure also shows a meaningful correlation (0.15), confirming its role as a risk factor. Interestingly, the heatmap reveals that smoking, obesity, and diabetes have minimal correlation with thyroid cancer risk, with values close to zero. This finding challenges some common assumptions about lifestyle factors and thyroid cancer, suggesting that genetic predisposition and environmental exposures may play more dominant roles than certain lifestyle factors in determining thyroid cancer risk. The demographic distributions help contextualize our model within diverse patient populations, while the correlation heatmap provides quantitative evidence of the relative importance of different risk factors. Together, these visualizations bridge the gap between statistical analysis and clinical interpretation, making the findings more accessible to healthcare practitioners who may not have extensive data science expertise.

Discussion

The results of this analysis revealed interesting patterns that both confirm existing medical knowledge and suggest new areas for investigation. Our machine learning approach achieved moderate predictive performance for thyroid cancer, with the models showing similar accuracy (82.3% for Random Forest and 82.5% for Logistic Regression). The comparable performance between these algorithms suggests that the relationships between predictive features and thyroid cancer may be more linear than initially expected. One surprising finding was the relatively modest recall (44-45%) for malignant cases across both models, indicating challenges in capturing all true positive cases. This limitation is particularly important in a medical context, where missed cancer diagnoses can have serious consequences. The similarity in performance metrics between the Random Forest and Logistic Regression models challenges our initial hypothesis that ensemble methods would substantially outperform linear models for this particular prediction task. The correlation analysis between risk factors and thyroid cancer risk showed that family history had the strongest correlation (0.24), followed by iodine deficiency (0.17) and radiation exposure (0.15). These findings align with known thyroid cancer risk factors. However, the relatively low correlation values overall suggest that individual risk factors alone may have limited predictive power, and that combinations of factors likely play a more significant role in determining cancer risk. The demographic analysis revealed informative patterns in the dataset. The 60/40 gender split (female to male) confirms the higher prevalence of thyroid conditions in women, while the diverse country and ethnicity distributions strengthen the generalizability of our findings.

Several limitations should be acknowledged in this analysis. The class imbalance in the dataset (with benign cases significantly outnumbering malignant ones) presents a challenge for model training and evaluation. The relatively modest recall for malignant cases indicates that the current model configuration misses a substantial portion of true cancer cases. Additionally, the dataset lacks temporal information that could help distinguish between slow-growing benign conditions and aggressive malignancies.

Conclusion

This project successfully demonstrated the potential of machine learning for thyroid cancer risk prediction. The Random Forest model achieved high accuracy and provided valuable insights into the factors that contribute to cancer risk. The approach balances predictive performance with interpretability, making it potentially useful in clinical contexts.

The model can be used for individual risk prediction, as shown in this example code:

```
# Example of making predictions for a new patient
new_patient_data = pd.DataFrame({
    'Age': [40],
    'Gender': ['Female'],
    'Country': ['USA'],
    'Ethnicity': ['Caucasian'],
    'Family_History': ['No'],
    'Radiation_Exposure': ['No'],
    'Iodine_Deficiency': ['No'],
    'Smoking': ['No'],
    'Obesity': ['No'],
    'Diabetes': ['No'],
    'TSH_Level': [2.5],
    'T3_Level': [120],
    'T4_Level': [8.5],
    'Nodule_Size': [1.2],
    'Thyroid_Cancer_Risk': ['Low']
})

predictions, probabilities = predictor.predict(new_patient_data)
print(f"Prediction: {predictions[0]}")
print(f"Probability of Malignant: {probabilities[0][1]:.3f}")
```

This allows you to put in any demographics you please to see the algorithms prediction of their chances of thyroid cancer.. As thyroid cancer incidence continues to rise globally, tools that can aid in early detection and risk stratification will become increasingly valuable to healthcare providers.

By combining machine learning with clinical knowledge, we can develop systems that augment medical decision-making and potentially improve patient outcomes through earlier intervention and more personalized risk assessment.