**ARTIFICIAL INTELLIGENCE**

**(Real-World Problem Solving Using Artificial Intelligence Approaches)**

**Project on:**

**“Thyroid Cancer Risk and Diagnosis Prediction”**

**Submitted By**

Chaudhry Mustafa Mudasser(S5730857)

**Submitted On**

15/05/2025

**Table of Contents**

Abstract ……………………………………………………………………………………………….3

1. Introduction ………………………………………………………………………………………3
2. The Real-World Problem ……………………………………………………………………….4
3. Project Aim and Objectives …………………………………………………………………….4
4. Adopted Artificial Intelligence Approach ………………………………………………………5
5. Artificial Intelligence Approach Implementation ……………………………………………...5
6. Evaluation, Results and Discussions …………………………………………………………7
7. Conclusion and Future Work …………………………………………………………………..8
8. References ……………………………………………………………………………………….9

## Abstract

Over the past three decades, the incidence of thyroid cancer has risen substantially, making it the fastest-growing cancer among women (American Thyroid Association, 2017). Although the underlying causes remain uncertain, data indicate that women aged 40 to 44 exhibit the highest risk, whereas men face an increased risk between the ages of 70 and 74. Ongoing research explores various potential contributing factors such as age, obesity, family history, etc. (Cancer Research UK, 2023). With the advent of advanced imaging technologies, some experts now suggest that thyroid cancer may be overdiagnosed (American Thyroid Association, 2017). With a growing global population, healthcare resources are becoming increasingly limited. Accurately identifying benign tumors can significantly reduce unnecessary medical procedures and associated costs. Leveraging AI models to assist in precise prediction can help conserve time and healthcare resources. Moreover, early and accurate detection is critical, as it facilitates timely surgical intervention with minimal complications (Scientific Reports, 2022). This study employs an artificial intelligence approach using a publicly available dataset from Kaggle. This dataset was selected for two main purposes: first, to predict an individual’s risk of developing thyroid cancer; and secondly, to get an accurate diagnosis of the tumor. Multiple AI models, including Random Forest classifier, Bagging classifier, XGBBoost, etc., have been used on this dataset, each yielding varying levels of predictive accuracy. These models were then evaluated based on the time run and their accuracy. The main idea behind this was to build a model to predict the risk and diagnosis for patients.

## Introduction

The thyroid is a butterfly-shaped gland that rests at the base of the neck, right below the Adam's apple. The main purpose of thyroid cancer is to produce thyroxine(T4) and Triiodothyronine(T3) hormones. These hormones help the human body with multiple important tasks, such as regulating blood pressure, heart rate, body temperature, and weight(Mayo Clinic, 2024). The increase in thyroid cancer reports has been seen in the past three decades(American Thyroid Association, 2017). Thyroid glands not producing sufficient hormones or producing them with some abnormalities can lead to multiple issues within the body; two of the most common ones are hyperthyroidism and hypothyroidism. Due to the advancements in data processing and computing power, many new approaches have been discovered and implemented. My approach consisted mainly of applying a few approaches established before this paper and some basic models to see how good they would work for my dataset.

## The Real-World Problem

Thyroid Cancer reports are increasing, and it is estimated that in 2025, there will be 44,020 new cases (National Cancer Institute, 2025). Thyroid cancer treatment, on average, costs about $1425 to $17,000 (National Library of Medicine, 2022). If caught in the early stages, thyroid cancer can be treated with medication. With the help of a working AI model, which could predict the risk and diagnosis, we could catch the cancer early on and treat it with medication. Furthermore, the hospitals would also cut costs by reducing the diagnosis time.

## Project Aim and Objectives

This report aims to develop multiple AI models with the intent to accurately predict the cancer risk and diagnosis based on the publicly available dataset and evaluate the best model. Hence, contributing to a timely and accurate fix for this problem. The objectives for the project are as follows:

* To explore and preprocess the dataset:

Clean and preprocess the Kaggle dataset, i.e., handle missing values, normalization, and encoding categorical fields

* To implement multiple AI models:

As per the popular AI models and a couple of past approaches, train multiple AI models

* To compare the AI models:

Compare all models trained based on train and test accuracy, time used to train, and F1 score

* To evaluate the selected model:

Evaluate the selected model based on the classification report, Confusion matrix, precision, and recall.

* Feature importance analysis:

Based on the selected mode,l a feature analysis would also be done to determine which features have the most importance in the result

* To document findings:

Form a report to document all findings and see how a basic version of past approaches would work on the newly available dataset.

## Adopted Artificial Intelligence Approach

Classification problems often require supervised machine learning approaches when labelled data is in question. Similarly, to solve this problem, supervised machine learning has been used.

In the past LDA model with 6-fold cross-validation has been used on a different thyroid dataset to get a 99.62% accuracy(ResearchGate, 2016). A few other models were also considered based on their past performances in multiple real-world problems. These included XGBoost Classifier, Decision Tree, Decision Tree with bagging classifier, AdaBoost, Logistic Regression, and Random Forest classifier. Some of these models supported ensemble learning, and they are known for their high accuracy, capability to handle imbalanced datasets, and resilience to noise and overfitting. The end goal was to automate and enhance the accuracy of diagnosis by learning data patterns using AI. Due to computational limitations, certain models were not selected, particularly Multilayer Perceptrons (MLPs), a type of artificial neural network, because they require extensive hyperparameter tuning and are significantly more resource-intensive. In general, many neural network-based approaches were avoided for this project because of their high computational demands and longer training times. Thus, majorly ensemble methods were adopted for this medical classification problem.

Imbalanced datasets are very common in fields like these, and in this specific example, the dataset was indeed imbalanced, so an oversampler was needed. Random Oversampler, which duplicates minority entries, was used for this purpose. Data scaling was done using RobustScaler because, unlike MinMaxScaler and StandardScaler it is designed to be resistant to the influence of outliers, and in medical datasets, outliers can often be found(Medium, 2023). Furthermore, for every model, cross-validation was also used, mostly set to a 6-fold cross-validation.

## Artificial Intelligence Approach Implementation

The challenge that presented itself with this problem was to correctly predict a thyroid cancer patient's risk and diagnosis. The publicly available dataset has over two hundred thousand entries, which have been collected over many years. The records consist of many demographic and medical features such as Obesity, Smoking, TSH levels, Age, Gender, Country, etc. As per common stereotypes, this medical data was also imbalanced. Oversampling methods were used to fix that. The hypothesis behind this project was that if supervised machine learning models were used to train on this dataset, then they could accurately predict the risk and diagnosis, preventing unnecessary medical costs and leading to an early diagnosis and effective treatment plan. To achieve this and successfully deploy the models, the following steps were implemented in Python on Google Colab:

1. Data preprocessing:
   1. Missing values were dealt with
   2. Unnecessary columns like patientID were dropped from the dataset because they couldn’t possibly affect the output.
   3. A label encoder was used to transform categorical text labels into integer values for the model to train easily
   4. One-hot encoding was performed to convert fields like Gender, Country, etc., into their separate binary columns.
   5. RobustScaler was used to scale the numerical values. The reason behind choosing RobustScaler was that it is more effective for datasets with outliers.
2. Correlation:
   1. Correlation analysis was performed on the dataset to identify features most related to the target. However, it was found to be unreliable, as several features known to affect thyroid cancer showed weak correlation. This suggests that correlation alone might not reflect actual medical significance. Therefore, except for one model, correlation was not used further.
3. Data Splitting, Balancing, and Validation:

Apart from a select few models, all models shared the same basic approach

* 1. The original data was copied, and using train\_test\_split, it was split into 2 parts: 80% was given to train, and 20% to test.
  2. RandomOverSampler was used on the training data to reduce class imbalance.
  3. Aside from a few, 6-fold cross-validation was also used on almost all models.

1. Model Training:

All models were trained in two stages. The first model was trained to predict the thyroid cancer risk. In the second step, those predictions were appended back into the dataset, and then with their help, another model was trained to predict the cancer diagnosis, i.e., whether it was benign or malignant.

All models requiring a random state to replicate results were set to 42. The following models were trained:

* 1. LogisticRegression with max iterations set at 2000 without RandomOversampler and cross-validation.
  2. RandomForestClassifier without RandomOversampler and cross-validation.
  3. AdaBoostClassifier without RandomOversampler and cross-validation.
  4. DecisionTreeClassifier without RandomOversampler and cross-validation.
  5. Basic LDA model
  6. LDA model with Stratified KFold cross-validation
  7. XGBBoost with 6-fold cross-validation and RandomOverSampler
  8. Random Forest Classifier with RandomOverSampler and 6-fold cross-validation
  9. Random Forest Classifier with RandomOverSampler and 10-fold cross-validation
  10. Decision Tree with Bagging Classifier
  11. Random Forest Classifier with attribute selection, 6-fold cross-validation, and RandomOverSampler

1. Model Evaluation

All models were evaluated based on test accuracy, test F1 score, and total time taken to train

1. Model Selection

The model was selected based on multiple factors, mainly the highest accuracy and relatively less time taken to train when compared to other models.

1. Performance evaluation:

The model selected had the highest accuracy and relatively less time

Random Forest classifier with 6-fold cross-validation and random oversampler with the following scores:

Thyroid cancer risk:

1. Cross-validation Accuracy (Train): 1.0
2. Cross-validation Accuracy (Validation): 0.808
3. Cross-validation F1-score (Train): 1.0
4. Cross-validation F1-score (Validation): 0.8047
5. Test Accuracy: 0.6179
6. Test F1-score: 0.5853
7. Training Time: 344.37 seconds
8. Precision(weighted avg): 66%
9. Recall(weighted avg): 62%

Diagnosis:

1. Cross-validation Accuracy (Train): 1.0
2. Cross-validation Accuracy (Validation): 0.9606
3. Cross-validation F1-score (Train): 1.0
4. Cross-validation F1-score (Validation): 0.9605
5. Test Accuracy: 0.8286
6. Test F1-score: 0.8144
7. Training Time: 379.92 seconds
8. Precision(weighted avg): 82%
9. Recall(weighted avg): 83%

The obtained results prove that ensemble learning outperforms other AI models. However, the scores we obtained need improvement before the model can be successfully deployed.

## Evaluation, Results, and Discussions

The obtained results demonstrate that the Random Forest classifier with a random oversampler and a 6-fold cross-validation has the best accuracy. Increasing the cross-validation fold number and performing feature selection has had a minor effect. Other models used did use less time, but with their prediction, a major difference could be observed. The selected solution had two models. The first model predicting the cancer risk had a test accuracy of 61.79% and took 351.99 seconds to train, while the model predicting Diagnosis had a test accuracy of 82.86% and took 379.37 seconds to train. Also, with the help of feature selection, it was determined which features are the most important for both models.

Overall, the lower test accuracy of the first model shows that predicting risk is a complex task. It could require a more refined approach, including feature engineering or maybe a deeper model architecture.

## Conclusion and Future Work

The project successfully demonstrated the use of Artificial Intelligence to solve a real-world healthcare classification problem that is predicting thyroid cancer risk and diagnosis. By implementing multiple Artificial Intelligence concepts and models, a presentable accuracy was achieved, especially for the thyroid cancer diagnosis classification. Throughout the process, multiple artificial intelligence key concepts were applied, for example, supervised learning, cross-validation, and feature importance analysis. In the future, there is a lot of room for improvement, due to computational constraints, complex models were not considered, but in the future, more complex models like MLP could be considered. On these more complex models, hyperparameter optimization could be done to further fine-tune the model. More complex and advanced datasets could be used. In my observation, this publicly available dataset could not be the best fit for thyroid risk prediction. Improvements could be made to the dataset. I believe all these improvements could significantly improve the accuracy and real-world applicability of the solution.

## Refereneces

American Thyroid Association, 2017. Available from: https://www.thyroid.org/patient-thyroid-information/ct-for-patients/february-2017/vol-10-issue-2-p-9/ [Accessed 15 May 2025].

Cancer Research UK, 2023. Available from: https://www.cancerresearchuk.org/about-cancer/thyroid-cancer/causes-risks [Accessed 15 May 2025].

Mayo Clinic, 2024. Available from: https://www.mayoclinic.org/diseases-conditions/thyroid-cancer/symptoms-causes/syc-20354161 [Accessed 15 May 2025].

Scientific Reports, 2022. Available from: https://www.nature.com/articles/s41598-022-15342-z [Accessed 15 May 2025].

National Cancer Institute, 2025. Available from: https://seer.cancer.gov/statfacts/html/thyro.html [Accessed 15 May 2025].

National Library of Medicine, 2022. Available from: https://pubmed.ncbi.nlm.nih.gov/35511135/ [Accessed 15 May 2025].

ResearchGate, 2016. Available from: https://www.researchgate.net/publication/306007162\_Machine\_Learning\_Techniques\_for\_Thyroid\_Disease\_Diagnosis\_-\_A\_Review?enrichId=rgreq-7bd10dc5ced597bdfc701b654419a30f-XXX&enrichSource=Y292ZXJQYWdlOzMwNjAwNzE2MjtBUzo1OTgyMjA3MTU1OTM3MjhAMTUxOTYzODM0MTE2Nw%3D%3D&el=1\_x\_3&\_esc=publicationCoverPdf [Accessed 15 May 2025].

Medium 2023. Available from: https://medium.com/@hhuseyincosgun/which-data-scaling-technique-should-i-use-a1615292061e [Accessed 15 May 2025].