**Thyroid Disease Classification**

A Multiclass Classification Approach

Revision Number: 1.0

Last date of revision: 04/05/2023

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 04/05/2023 | 1 | Initial HLD — V1.0 | M. Junaid |
|  |  |  |  |

## Contents

[Document Version Control. 2](#_TOC_250021)

[Abstract. 4](#_TOC_250020)

1. [Introduction 5](#_TOC_250019)
   1. [Why this High-Level Design Document?. 5](#_TOC_250018)
   2. [Scope. 5](#_TOC_250017)
   3. [Definitions 5](#_TOC_250016)
2. [General Description. 6](#_TOC_250015)
   1. [Product Perspective 6](#_TOC_250014)
   2. [Problem statement 6](#_TOC_250013)
   3. [PROPOSED SOLUTION 6](#_TOC_250012)
   4. [FURTHER IMPROVEMENTS 6](#_TOC_250011)
   5. [Technical Requirements. 6](#_TOC_250010)
   6. [Data Requirements 7](#_TOC_250009)
   7. [Tools used. 8](#_TOC_250008)
      1. Hardware Requirements. 8
      2. ROS(Robotic Operating System) 9
   8. [Constraints 9](#_TOC_250007)
   9. [Assumptions. 9](#_TOC_250006)
3. [Design Details 10](#_TOC_250005)
   1. [Process Flow. 10](#_TOC_250004)
      1. [Model Training and Evaluation 10](#_TOC_250003)
      2. [Deployment Process 11](#_TOC_250002)
   2. [Event log 11](#_TOC_250001)
   3. [Error Handling 11](#_TOC_250000)
   4. Performance. 12
   5. Reusability. 12
   6. Application Compatibility 12
   7. Resource Utilization 12
   8. Deployment. 12
4. Dashboards. 13
   1. KPIs (Key Performance Indicators) 13
5. Conclusion 14

## Abstract

Thyroid diseases, such as hyperthyroidism and hypothyroidism, affect a significant portion of the population and can lead to serious health complications. Accurate diagnosis and early intervention are crucial for effective treatment. In this project, we aim to develop a machine learning-based approach for the early detection and classification of thyroid disease. Our approach involves the use of multiclass classification models to predict the likelihood of a patient having a diseased state of the thyroid. We utilize various machine-learning techniques and feature engineering methods to achieve accurate and reliable predictions. Our system also includes data preprocessing, model training and evaluation, and automated reporting of data drift and model performance to ensure the model remains robust and effective. Overall, our project aims to provide a reliable and efficient tool for healthcare professionals in the diagnosis and treatment of thyroid disease.

### Introduction

* 1. Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all of the design aspects and define them in detail.
* Describe the user interface being implemented.
* Describe the hardware and software interfaces.
* Describe the performance requirements.
* Include design features and the architecture of the project.
* List and describe the non-functional attributes like:
  + Security
  + Reliability
  + Maintainability
  + Portability
  + Reusability
  + Application compatibility
  + Resource utilization
  + Serviceability

###### Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

##### Definitions

|  |  |
| --- | --- |
| **TERM** | **DESCRIPTION** |
|  |  |
| *Database* | A database is a collection of organized data that can be easily accessed, managed, and updated. |
| *IDE* | Integrated Development Environment |
|  |  |

#### General Description

##### Product Perspective

Random Forest based machine learning model is a Thyroid Disease Classifier which can help medical experts to detect different categories of thyroid disease and in early prediction of the likelihood for the disease development.

* 1. **Problem statement**

The main goal is to predict the estimated risk on a patient's chance of obtaining thyroid disease or not.

* 1. PROPOSED SOLUTION

In the provided dataset, target variable have 15 class labels however the share of some class label is negligible close to 0.001%. So decided to group the class labels into 8 class labels. This categorization is based on domain knowledge. Oversampling of minority class was also tested however that approach didn’t perform well on test data while considering it performance on train data

* 1. FURTHER IMPROVEMENTS

Performance can be improved by performing more the extensive hyperparameter tuning and by using more samples for minority class labels.

##### Data Requirements

Data requirement completely depend on our problem statement. We need thyroid diagnosis data for each class:

* Age
* Sex
* Sick or not
* Pregnant or not
* Thyroid Surgery or not
* Goitre or not
* Tumor or not
* Hypopituitary or not
* Hypothyroid or not
* Hyperthyroid or not
* Psych
* Thyroxine therapy or not
* Antithyroid Medication or not
* I131 treatment or not
* TSH Level
* T3 level
* T4 level
* T4U level
* TT4 level
* FTI level

###### Tools used

A picture containing graphics, graphic design, font, logo

Description automatically generatedPython programming language and frameworks such as NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, Plotly, Evidently, PyCharm, VSCode, Git, Azure, Render, S3 Bucket are used to build the whole model.

A picture containing traffic sign, symbol, sign

Description automatically generatedA picture containing font, electric blue, logo, graphics

Description automatically generated



[This Photo](https://lthub.ubc.ca/guides/github-instructor-guide/) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

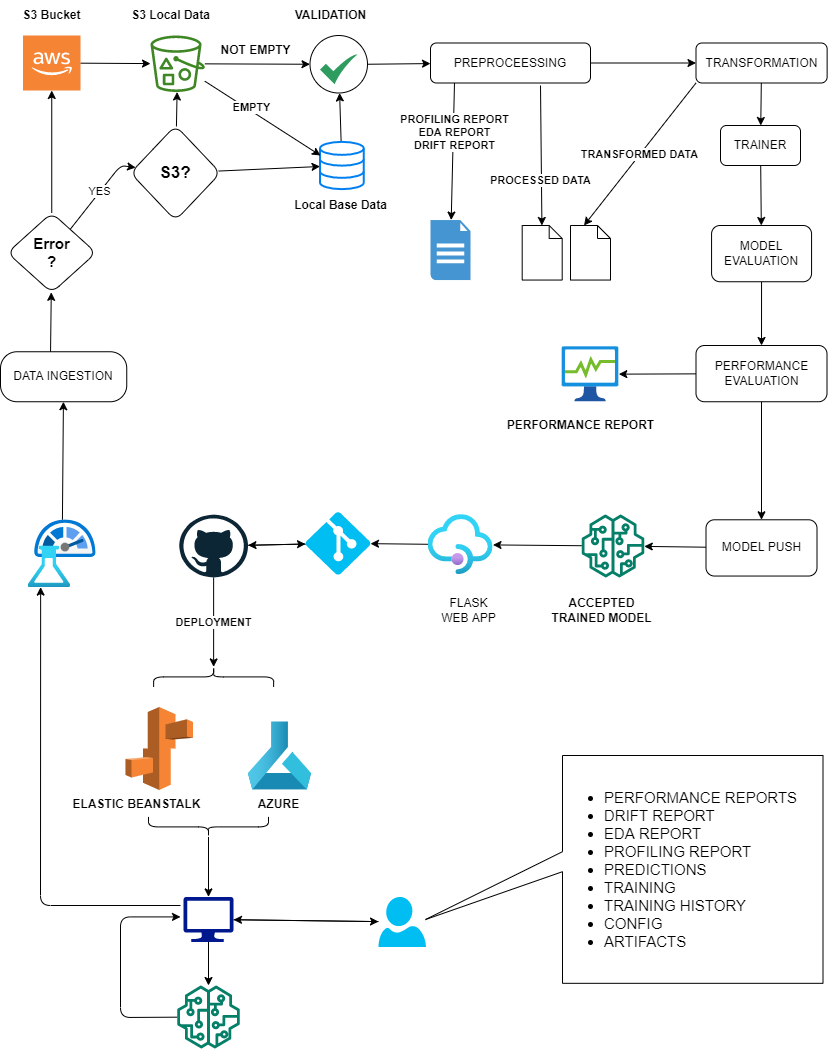
* PyCharm is used as IDE.
* For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
* Azure is used for deployment of the model.
* S3 Bucket for Data storage
* Front end development is done using HTML/CSS
* Python Flask is used for backend development.
* GitHub is used as version control system.
* GitHub Actions is used for CI / CD Pipeline

##### Assumptions

The main objective of the project is to implement the use cases as previously mentioned (2.2 Problem Statement) for new dataset that comes through diagnosis of thyroid. Machine Learning based thyroid disease classification model is used for detecting the above-mentioned use cases based on the input data. It is also assumed that all aspects of this project have the ability to work together in the way the designer is expecting.

## Design Details

###### Process Flow



###### Event log

The system should log every event so that the user will know what process is running internally.

Initial Step-By-Step Description:

1. The System identifies at what step logging required
2. The System should be able to log each and every system flow.
3. Developer can choose logging method. You can choose database logging/ File logging as well.
4. System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

##### Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong. An error will be defined as anything that falls outside the normal and intended usage.

1. Performance

Th model can be used for classification of 8 different thyroid diseased states. So we have paid more attention on ROC and F1 balance accuracy and log loss to prevent type 2 error. Also, model retraining is very important to improve performance.

##### Reusability

The code written and the components used should have the ability to be reused with no problems.

##### Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure the proper transfer of information.

##### Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

* 1. **Deployment**

A picture containing font, electric blue, logo, graphics

Description automatically generated

## Dashboards

Flask based web dashboard will display certain KPIs for the deployed model. All the reports, logs artifacts can be accessed directly from dashboard

##### KPls (Key Performance Indicators)

1. F1 weighted score for train and test
2. ROC AUC Ovr Weighted score for train and test
3. Balanced accuracy score for train and test
4. Log loss score for train and test

## Conclusion

Developed model is able to make prediction with F1\_weighted: 0.75, ROC AUC OR weighted: 0.86, balanced accuracy: 0.85 and log loss: 1.013

## References