



# **Hybrid Machine Learning Framework for Multi-Class Thyroid Risk Assessment**

Prepared By

Name: Ahmed Abdullah

ID No: 253000598

Submitted in fulfillment of the requirements for the degree Master of  
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

**Supervisor Name: Dr. Amjad Qtaish**

Collage of Informatics, Midocean University, Comoros

Email:.....

## Acknowledgements

First and foremost, I would like to express my gratitude to Dr. Amjad Qataish for his guidance, constructive criticism, and suggestions throughout this process of research. The contribution of Dr. Qataish has been significant in ensuring the success of this research.

Secondly, I would like to thank the teaching faculty of the College of Informatics at Midocean University for providing me the most conducive learning environment and knowledge that has helped me in completing my master's studies.

Thirdly, I would like to thank my family members and friends for supporting and encouraging me all through my master's program. Their belief in me has been the most important factor behind all my achievements.

Finally, I would like to thank everyone who has directly or indirectly contributed to the completion of this research.

# Abstract

Diseases of the thyroid gland represent one of the most common types of endocrine diseases, which require prompt and accurate diagnosis in order to prevent possible complications for patients. In this paper, an AI-powered Clinical Decision Support System (CDSS) for Multi-Class Classification of the Risk Factors for Thyroid Diseases based on machine learning algorithms is proposed. For the research, the UCI ANN Thyroid Disease dataset was used, with the accent being made on five laboratory biomarkers that are widely used in medicine: Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), Total Thyroxine (TT4), Thyroxine Uptake (T4U), and the Free Thyroxine Index (FTI).

The Extreme Gradient Boosting (XGBoost) classifier was developed for predicting three classes of patients: Normal, Hypothyroidism, and Hyperthyroidism. The achieved results of the classification model represent accuracy equal to 97.23%, while the weighted precision, recall and F1 score have values close to 97%. This study shows the excellent classification performance of the model that was built using the combination of only laboratory biomarkers. The proposed classifier was implemented in the Streamlit-based CDSS, which enables to predict the disease, receive the confidence score, probability distribution, laboratory reference range, and clinical recommendations.

**Keywords:** Thyroid Disorders, Machine Learning, XGBoost, Clinical Decision Support System, Artificial Intelligence, Healthcare Analytics, Thyroid Hormones.

## List of Publications

NONE

# Table of Contents

|                                                        |      |
|--------------------------------------------------------|------|
| List of Abbreviations .....                            | viii |
| List of Figures .....                                  | ix   |
| List of Tables .....                                   | x    |
| 3.1 Introduction.....                                  | 23   |
| 3.2 Research Methodology .....                         | 25   |
| 3.3 Proposed System Architecture.....                  | 28   |
| 3.4 Dataset Description.....                           | 30   |
| 3.5 Data Preprocessing .....                           | 34   |
| 3.6 Machine Learning Model Development.....            | 38   |
| 3.7 Model Evaluation.....                              | 42   |
| 3.8 Clinical Decision Support System Development ..... | 47   |
| 3.9 System Workflow .....                              | 52   |
| 3.10 Chapter Summary .....                             | 55   |
| Chapter 4: Experimental Results and Discussion.....    | 56   |
| 4.1 Introduction.....                                  | 56   |
| 4.2 System Implementation .....                        | 57   |
| 4.3 Experimental Setup.....                            | 61   |
| 4.4 Model Training and Performance Evaluation .....    | 64   |
| 4.5 Clinical Decision Support System Prototype .....   | 69   |

|                                                 |    |
|-------------------------------------------------|----|
| 4.6 System Validation Using Test Cases.....     | 73 |
| 4.7 Discussion.....                             | 78 |
| 4.8 Limitations of the Proposed Framework ..... | 79 |
| 4.9 Chapter Summary .....                       | 80 |
| 5.1 Introduction.....                           | 82 |
| 5.2 Summary of the Research.....                | 82 |
| 5.3 Research Objectives Achievement.....        | 84 |
| 5.4 Main Research Contributions .....           | 86 |
| 5.5 Limitations of the Study .....              | 88 |
| 5.6 Future Work.....                            | 90 |
| 5.7 Final Conclusion.....                       | 93 |

# List of Abbreviations

- **AI:** Artificial Intelligence
- **ANN:** Artificial Neural Network
- **CDSS:** Clinical Decision Support System
- **DT:** Decision Tree
- **EHR:** Electronic Health Record
- **FN:** False Negative
- **FP:** False Positive
- **FTI:** Free Thyroxine Index
- **HPT axis:** Hypothalamic-Pituitary-Thyroid (axis)
- **LIME:** Local Interpretable Model-agnostic Explanations
- **ML:** Machine Learning
- **SHAP:** SHapley Additive exPlanations
- **SVM:** Support Vector Machine
- **T4U:** Thyroxine Uptake
- **TP:** True Positive
- **TRH:** Thyrotropin Releasing Hormone
- **TSH:** Thyroid Stimulating Hormone
- **TT4:** Total Thyroxine
- **UCI:** University of California, Irvine (repository of data sets)
- **XAI:** Explainable Artificial Intelligence
- **XGBoost:** Extreme Gradient Boosting

## List of Figures

|                                                            |    |
|------------------------------------------------------------|----|
| Figure 3.1 Overall Research Methodology.....               | 34 |
| Figure 3.2 Proposed System Architecture.....               | 39 |
| Figure 3.3 Data Processing Workflow.....                   | 46 |
| Figure 3.4 XGBoost Model Development Process.....          | 51 |
| Figure 3.5 Model Evaluation Workflow.....                  | 55 |
| Figure 3.6 Clinical Decision Support System Workflow.....  | 60 |
| Figure 3.7 Overall System Workflow.....                    | 63 |
| Figure 4.1 system Implementation Workflow.....             | 69 |
| Figure 4.2 Software Architecture.....                      | 70 |
| Figure 4.3 Experimental Workflow.....                      | 73 |
| Figure 4.4 Confusion Matrix of the XGBoost Classifier..... | 75 |
| Figure 4.5 Laboratory Data Entry Module.....               | 79 |
| Figure 4.6 AI Prediction Module.....                       | 80 |
| Figure 4.7 Validation Workflow.....                        | 86 |

## List of Tables

|                                                                        |     |
|------------------------------------------------------------------------|-----|
| Table 2.1 Summary of Related Work.....                                 | 29  |
| Table 3.1 XGBoost Hyperparameter Configuration.....                    | 49  |
| Table 3.2 Main Components of the Clinical Decision Support System..... | 63  |
| Table 4.1 Development Environment.....                                 | 68  |
| Table 4.2 Main Software Components.....                                | 69  |
| Table 4.3 XGBoost Configuration.....                                   | 72  |
| Table 4.4 Summary of Model Performance.....                            | 77  |
| Table 4.5 Functional Testing Summary.....                              | 82  |
| Table 4.6 Functional Validation Results.....                           | 84  |
| Table 4.7 Example Validation Using UCI Dataset Records.....            | 85  |
| Table 4.8 Clinical Validation Test Cases.....                          | 85  |
| Table 5.1 Achievement of Research Objectives.....                      | 94  |
| Table 5.2 Summary of Research Contributions.....                       | 96  |
| Table 5.3 Summary of Study Limitations.....                            | 98  |
| Table 5.4 Recommendations for Future Research.....                     | 101 |

# Chapter 1 Introduction

## 1.1 Overview

Thyroid issues are one of the most common conditions worldwide and are known to affect a wide range of physiological processes within human beings. The thyroid gland has a critical direct impact on regulating metabolism, cardiovascular function, and body temperature. Any disturbance in thyroid hormone production may lead to significant complications, including hypothyroidism and hyperthyroidism. Due to the prevalence of these conditions and their impact on human health, accurate and timely diagnosis remains a critical clinical point (Taylor et al., 2022).

Usually, thyroid issues are evaluated through lab testing of markers, such as thyroid-stimulating hormone (TSH), triiodothyronine (T3), and thyroxine (T4). These numbers are common in providing insights in clinical practice (Biondi & Cappola, 2021). This lab metric does not always provide a direct insight, especially when numbers fall near diagnostic thresholds or when a lot of indicators must be taken into consideration at the same time.

Recent practices done by technology / medical members in illness identification and classification. These practices provide support in analyzing medical data and lab results. In our case in this paper, many studies introduced the benefits of implementing machine learning models in the identification of conditions in the clinics based on lab numbers (Rajkomar et al., 2022). Many

good results have been proven through adopting these techniques in terms of accuracy.

Based on the facts mentioned, there is a gap that needs to be filled to not only provide accurate classification of thyroid disorders using lab results but also present the results in a clear, integrated method that can support clinical decisions. In this study we will address this need through implementing a machine learning framework that focuses on lab results while stressing interpretability and usability.

## **1.2 Problem Definition**

Machine learning methods have been applied to prediction thyroid disease; most studies have mainly focused on improving classification using lab biomarkers such as TSH, T3, and T4. While these techniques have proven effective results, their adoption in real clinical practices remains limited (Biondi & Cappola, 2021). Frequently, models are implemented and adopted based on metrics alone, without sufficient consideration of how the results can be integrated and implemented by medical practitioners.

In clinical practice, diagnostic decisions are based on metrics and understanding of how different metrics contribute to the final consideration. However, many machine learning frameworks operate as mysterious systems, where predictions are generated without detailed and clear explanation. The absence of transparency reduces trust in these kinds of frameworks and limits their implementation in medical environments. This can result in even highly accurate models not providing meaningful support for clinical practices.

Also, the absence of datasets that have accurate data impacted the adoption of more accurate methodologies. These issues are presented many times by authors in their papers in chapter 2 (Literature Review); new papers still present their work focusing only on lab testing results. Through this, a gap was found between the actual performance and technology in supporting medical people on their daily activities.

By looking into my statements above, I can provide an urge for a framework that does not only focus on providing medical people a classification of illnesses but also provides an enhancement of using the lab test results. The future framework is expected to provide a better understanding and provide medical teams structured results that match clinical results.

### **1.3 Thesis Objectives**

I have a lot of goals that I look forward to introducing in this paper, but the most important goal that I would like to introduce in this paper is a machine learning technique for thyroid problems with UCI laboratory datasets.

The research will be guided by the following guidelines to help us meet our objective:

- Use the UCI Thyroid data set and regard it as the major data set for our study.
- Adopt decision trees and random forests as machine learning methods.
- Test our findings/outputs using numbers and make sure they are accurate.
- Create a framework that provides results to medical practitioners in a way that they understand.

- Reveal the limitations and challenges encountered during this study and provide suggestions on how to improve it in the future.

By doing these things, I would offer a framework that will provide the combination of high performance medical practitioners can rely on and interpretability for advanced machine learning.

## **1.4 Thesis Contribution**

First, the study will pinpoint weaknesses in existing studies concerning the status of thyroid prediction at present. The above will be done by emphasizing the need to enhance not only the performance of the models but also their interpretability and practical implementation.

Decision trees and random forests will be used in the analysis of thyroid prediction by means of the UCI Thyroid dataset and precision, recall, and F1 scores. The above-mentioned conceptual model is proposed so that findings can be communicated in an easy-to-understand way for medical practitioners.

Ultimately, issues with laboratory metric-based models will be determined, along with recommendations for the future.

## **1.5 Thesis Organization**

### **Chapter 1**

Chapter 1 will have an introduction to the problem and background information of the problem, and objectives of the work will be discussed in this chapter.

### **Chapter 2**

In chapter 2, a literature review will be done and comparisons of the results will be done using similar papers. In the last part of this chapter, comparison of the similar paper work will be done.

### **Chapter 3**

Chapter 3 will discuss the methodology that includes selection of the dataset and selecting appropriate methodology to accomplish the objective.

### **Chapter 4**

Chapter 4 will discuss the results obtained from experimental work along with discussions on it. It will be done through presentation of the results using model evaluation measures of different classification models.

### **Chapter 5**

Conclusion will be presented in the last chapter of the thesis.

## **Chapter 2: Background and Related Work**

Thyroid disorder is considered an important disease from the medical aspect because of its high prevalence among humans and impacts on the body processes, such as metabolic rate, heart functions, and temperature regulation. In order to determine the efficiency of functioning of the thyroid gland, there is a need to evaluate different biochemical indices and their complex interconnection with the symptoms. This makes the computer-based application quite necessary because this system will be capable of analyzing clinical indices. Machine learning techniques have been proven to be rather useful in contemporary medicine because they allow for detecting regularities in health records that can be used for diagnosing different ailments. For this reason, many studies have utilized machine learning techniques for predicting or classifying diseases on the basis of different laboratory indices. Therefore, applying medical records of thyroid disorders with the use of machine learning techniques is rather required.

### **2.1 Thyroid Gland and Thyroid Disorders**

There exist many individuals globally who are experiencing health problems related to the thyroid. The thyroid gland is responsible for regulating various metabolic activities within the body. The thyroid affects a number of biological activities such as metabolism rates, cardiovascular actions, thermoregulation, and energy levels, among others. This implies that any form of dysfunction in the levels of thyroid hormone production may affect various organs and systems in the body (Biondi & Cappola, 2021).

The thyroid gland is in the front part of the neck and produces two primary hormones: thyroxine (T4) and triiodothyronine (T3). These hormones are responsible for adjusting metabolism. The secretion of thyroid hormones is controlled by the hypothalamic–pituitary–thyroid (HPT) axis. In this scenario, the hypothalamus releases thyrotropin-releasing hormone (TRH), which increases the pituitary gland to produce thyroid-stimulating hormone (TSH). TSH then increases the thyroid gland to release T3 and T4. The levels of these hormones are regulated through a feedback mechanism that helps maintain hormonal balance (Ross et al., 2016).

Thyroid dysfunction happens when the hormonal system does not function as usual. The most common forms of thyroid dysfunction are hypothyroidism and hyperthyroidism. Hypothyroidism happens when the thyroid gland produces not enough levels of thyroid hormones. Patients with hypothyroidism may experience symptoms such as fatigue, weight gain, cold intolerance, and dry skin. Hyperthyroidism, on the other hand, occurs when the thyroid gland produces excessive thyroid hormones and may lead to symptoms such as weight loss, palpitations, anxiety, and heat intolerance (Biondi & Cappola, 2021; Chiovato et al., 2023).

To identify thyroid disorders, medical practitioners involve lab results of several biochemical indicators. Measurements of thyroid-stimulating hormone (TSH), triiodothyronine (T3), and thyroxine (T4) are commonly used to evaluate thyroid function. In some cases, additional derived indicators such as the Free Thyroxine Index (FTI) are used to improve the outcome understanding. Because thyroid hormones operate within a feedback system, involving these biomarkers

requires evaluating their relationships rather than examining each measurement alone (Ross et al., 2016).

Symptoms, in addition to lab indicators, also play an important role in the evaluation process. Yet, there is a lack of specificity with respect to symptoms since there are many symptoms that are not unique to thyroid disorders. These include fatigue, changes in weight, feeling hot or cold, and palpitations, among others. This causes clinicians to rely on a combination of laboratory indicators and evaluations done in the clinics when evaluating thyroid disorders (Chiovato et al., 2023).

The integration of various biochemical markers along with symptoms in a patient can sometimes be difficult, especially when lab values come very close to the cut-off values for diagnosis or there is a lack of definition in the symptoms. This has driven the need for computer-based algorithms capable of analyzing multiple parameters at once. Recent developments in AI and machine learning have revealed promising results for the analysis of health care data and predicting diseases (Khan et al., 2024).

It is crucial to have an understanding of the medical background regarding the physiology of the thyroid, its biomarkers, and symptoms before discussing the application of computational methods in predicting thyroid diseases. The knowledge acquired forms the basis for analyzing machine learning patterns that help in assessing thyroid data.

## **2.2 Machine Learning in Healthcare**

Machine learning technology has become the new trend in contemporary analysis software applications in the health care sector. It involves software programs that allow machines to identify patterns in data and make predictions or classifications without requiring substantial coding for each task performed by the algorithm. Machine learning allows an algorithm to process big data and establish correlations among different variables that cannot be easily established through conventional statistics (Esteva et al., 2021).

In the health care industry, machine learning technology has increasingly been applied as the new trend in diagnosing diseases, assessing risk factors, and assisting decision-making processes. Medical data has several complicated relationships among patient data, laboratory reports, physical observations, and historical data. Machine learning technologies have been applied in processing such data and identifying patterns to assist physicians in their diagnoses and predictions (Topol, 2023).

Machine learning algorithms application in health care includes designing clinical decision support systems. The latter should provide clinical feedback through evaluation of data. The intention here is not to replace medical expertise but to complement the process through the provision of feedback which helps the practitioners understand information and detect disease patterns. Machine learning algorithms have found their way in many areas of medicine ranging from prediction of cardiovascular conditions, cancers diagnosis, and endocrine conditions analysis (Khan et al., 2024).

Through the machine learning patterns, the lab indicators and symptoms can easily be integrated and evaluated together. Diagnosing any thyroid condition requires evaluating several lab indicators and symptoms at the same time. This might be quite hard under traditional approaches. To overcome such challenges, computational models could be used to evaluate combinations of lab results and symptoms to establish probabilities of thyroid conditions.

For this case study, the purpose of using the proposed algorithm is to assist in evaluating the risks of developing thyroid conditions through integration of clinical symptoms and lab indicators.

### **2.3 Machine Learning for Thyroid Disease Prediction**

Nowadays, machine learning models have been adopted in healthcare domains to aid disease identification and classification. Using machine learning models in the field of healthcare allows one to analyze the data obtained during practical activities and discover any patterns which might be related to thyroid disease. The connection between lab results and symptoms in thyroid diseases is rather complicated, which makes machine learning essential.

A number of research works have used machine learning models for predicting thyroid diseases. Most of those works have used laboratory parameters, including thyroid stimulating hormone (TSH), triiodothyronine (T3), and thyroxine (T4) as predictor variables. Several supervised machine learning models, such as decision tree, support vector machine, random forest, and neural network, were suggested as ways to classify patients according to their thyroid diseases (normal thyroid function, hypothyroidism, and hyperthyroidism).

Recently, researchers have focused on improving prediction performance by comparing various machine learning models. Many machine learning algorithms have been examined using structured data such as the UCI Thyroid Disease Data Set. From the findings, it is evident that machine learning models can yield positive results in predicting thyroid disease (Khan et al., 2024).

While significant improvements have been made regarding the thyroid diseases prediction issue by machine learning algorithms, almost all research papers have considered only the results from laboratory tests without mentioning the signs of patients. When diagnosing thyroid disease in the context of medicine, physicians consider the lab test results as well as signs of patients.

In addition, a lot of literature talks about the preciseness of prediction without considering any other criteria, for instance, interpretability. Thus, if we talk about health issues, the machine learning algorithms should be precise and at the same time interpretable to make decisions based on their results.

Several disadvantages inherent to the existing approaches applied in assessment of thyroid data have been listed in this paper. Having considered the mentioned disadvantages, it was suggested to use the hybrid approach which combines symptom assessment and machine learning algorithms.

## **2.4 Comments of other researchers**

Thyroid disease has been studied by many researchers using different methods. Both machine learning laboratory methods and clinical assessment methods have been employed to determine diagnosis and/or prognosis. Much current research focuses on laboratory biomarkers as the input to the predictive model.

Along these lines, Islam et al. (2022) found that machine learning models can be highly effective at predicting hypothyroidism from biological indicators such as TSH, T3 and T4 levels during hospital training sessions. Similar outcomes were achieved by Alshamrani et al. (2023) by applying several machine learning techniques to organized thyroid datasets to achieve accurate classification results. While results from such data are significant and reliable, more recent analyses by Shen et al. (2023) indicate that many methods heavily rely on organized numerical data and focus on improving the accuracy of future predictions.

Several researchers explored in depth the clinical aspects of thyroid diseases, emphasizing the difficulty of the diagnosis and the limitations of an entirely laboratory-based approach. Biondi and Cappola (2021) highlighted the many links of thyroid dysfunction, and its correct interpretation requires a thoughtful clinical approach. Chiovato et al. (2023) stressed that, in the real-life clinical practice, the diagnosis of thyroid diseases mainly relies on the correct interpretation of the laboratory data in the frame of the whole clinical case. There has been growing interest in how AI can be utilized within clinical practice as part of AI in healthcare. In their review of decision-support systems, Sutton et al. (2020) suggested that output should be clear, transparent and appropriate.

Effectiveness and integration into clinical workflows (Jiang et al. 2022) are also key considerations. However, as noted by Sendak et al. (2021), many machine learning models fail to translate into clinical practice due to user unfriendliness and lack of trustworthiness. In addition, it is generally agreed that a model which achieves high performance in prediction must also have good interpretability and reproducibility (Miotto et al. 2021, Rajkomar et al. 2022).

Thyroid diseases in the lab have been much better characterized in recent years. However, a major limitation of current research is the lack of focus on the interpretability and applicability of model results. Even though good predictions can be made, the results are not interpretable enough and therefore not useful enough for clinical decision support.

### **2.4.1 Analytical Discussion of Previous Research**

It therefore becomes clear that existing literature offers a wealth of information but fails to fulfill the required criteria. Machine learning studies conducted in laboratories have shown remarkable prediction capabilities; nevertheless, their heavy reliance on numerical measures makes it harder to comprehend their clinical significance, as mentioned in various reviews, including Shen et al. (2023).

Unlike that, clinical studies like the one carried out by Biondi & Cappola (2021) and Chiovato et al. (2023) demonstrate how complicated it is to analyze thyroid disorders and the requirement for the interpretation of findings. Nonetheless, there are no computational methods suggested by previous studies for automatic classification.

Furthermore, the mentioned properties that an AI-based system needs to have for being considered successfully implemented in the medical environment include interpretability, reliability, and clinically oriented nature, according to the literature on clinical applications, such as Sutton et al. (2020), Sendak et al. (2021), and Jiang et al. (2022). However, although there is a lot of evidence for the need for the discussed attributes, most algorithms are oriented exclusively towards increasing the accuracy of their models, as was indicated, among others, by Rajkomar et al. (2022) and Miotto et al. (2021).

Taking into consideration the presented limitation, the current paperwork opts to use machine learning based on numerical laboratory values, focusing specifically on its interpretability.

The idea behind such a choice implies that the emphasis should not be on the predicted value itself but rather on explaining how the behavior of the created algorithm operates. According to this purpose, one may assume that the research will be valuable for the classification of the discussed health condition.

## 2.5 Research Gap

In this regard, it must be said that these articles show a shift in the field in their departure from diagnostic techniques and application of more sophisticated technology in practice, such as machine learning. At the same time, no matter what new developments may appear in this field, the fact still exists that physicians do not want to gather all data obtained during laboratory analysis and during clinical diagnosis into a single database.

It can be exemplified by the guidelines proposed by Ross et al. (2016). According to the recommendations provided by the authors, physicians should diagnose and treat patients with various conditions related to the functioning of the thyroid gland according to data obtained as a result of laboratory analysis of blood, i.e., TSH, T3, and T4. Recommendations are created on the basis of experience and acquired medical knowledge; hence, there is no need for any sort of data modeling and management in this context.

However, now consider the example of Dua and Graff (2019) and the thyroid data from the UCI Machine Learning Repository. This is an excellent resource for those who need to prepare their data, select features, and run ML models. However, the dataset is numeric only, meaning no records about patient symptoms or narratives of cases can be found there. According to Shen et al. (2023), such datasets overvalue laboratory parameters in the analysis and undervalue the importance of the whole clinical picture.

Sutton et al. (2020) concentrate on the usability of the systems being discussed from a wide range of perspectives. Specifically, the researchers indicate that such

systems must be utilized naturally and must have been developed in line with certain procedures in the industry. However, there is no list of specific requirements regarding the type of database or prediction tool that should be used in their study. Useful suggestions are provided; nevertheless, they do not have anything to do with thyroid diseases or machine learning algorithms.

Biondi and Cappola (2021) focused entirely on the medical domain. This article examined the difficulty of interpreting hormone profiles, which is a complicated issue. There is no mention of machine learning, nor any discussion of metrics for predictions or automation. These results can be used directly by physicians, yet that is where their practical application ends.

One can cite the work of Esteva et al. (2021). They stressed the potential benefits of deep learning technologies in medicine, particularly in the area of image processing. They were able to achieve very high detection rates thanks to large databases and neural networks. However, this study did not consider laboratory data that could be structured, such as the information related to thyroids.

Sendak et al. (2021) studied the factors required to introduce machine learning to regular healthcare processes. They identified various issues in implementation such as trustworthiness, usability, and adaptability of models to particular infrastructure. This research is not based on particular dataset and predictive model but rather discusses general challenges faced by scientists trying to make machine learning usable in practice.

In 2021, Miotto et al. emphasized the significance of combining various data sets to enhance prediction power in the domain of medicine. The authors of the study surveyed various datasets and models and found that combination is the answer.

Moving away from the model-oriented research, Islam et al. (2022) examined different machine learning techniques for diagnosing thyroid diseases with the help of the UCI database. These researchers covered everything from data preprocessing to training a model and obtaining various measures such as accuracy and F1-score. The outcome was promising, yet the research lacked practicality in terms of physicians being able to implement it in their everyday practice. Similarly, Rajkomar et al. (2022) noted that while many models are technically sound, no one understands the logic behind them. In other words, the technology is available, yet not ready for implementation.

Jiang et al. (2022) considered the ways in which artificial intelligence might integrate into clinical procedures. These authors' approach relied mostly on theoretical considerations rather than practical solutions to be adopted by practitioners.

Back in Rajkomar et al. (2022), the authors revisited another important topic, i.e., besides being good at prediction, models should be interpretable and transparent. The topic was covered through a variety of medical applications, but thyroid disease and guidelines were not included. Nevertheless, the main idea was crystal clear.

Alshamrani et al. (2023) once again compared machine learning models that demonstrated high prediction rates based on the structured laboratory dataset related to thyroid diseases. However, similar to other works, the models remained rather a black box since they worked accurately but lacked interpretability. The same problem was identified by Shen et al. (2023).

Chiovato et al. (2023) brought back the basics of clinical work through their research in laboratories of the thyroid gland. Unlike the other works discussed above, Chiovato et al. (2023) focused on the problem clinically and noted that accurate diagnosis is quite difficult and complicated, especially without the use of AI technology.

Shen et al. (2023) carried out a detailed literature review on machine learning in predicting outcomes for endocrines. According to the findings of these researchers, it can be noticed that most researchers focus only on structured data and precision. There is no interest whatsoever in usability and explainability. This is the reason why more research is necessary.

**Khan et al. (2024) showed once again that machine learning algorithms continue increasing the accuracy rate in diagnosing endocrines. As can be seen from other studies, researchers have not considered anything but structured data and traditional criteria.**

## 2.5.1 Summary of Related Work

| # | Study Name                                                  | Date | Author           | Research Focus                             | Method / Approach                                                                | Key Contribution                                                     | Identified Limitation                                                                                               |
|---|-------------------------------------------------------------|------|------------------|--------------------------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| 1 | American Thyroid Association Guidelines for Hyperthyroidism | 2016 | Ross et al.      | Clinical diagnosis of thyroid disorders    | Clinical guidelines based on laboratory biomarkers (TSH, T3, T4)                 | Established standardized diagnostic protocols for thyroid assessment | Does not include computational models or automated prediction                                                       |
| 2 | High-Performance Medicine: Human and AI                     | 2019 | Topol            | AI in healthcare                           | Conceptual framework (no dataset or model)                                       | Emphasized integration of human expertise with AI systems            | No disease-specific implementation or predictive model                                                              |
| 3 | UCI Machine Learning Repository                             | 2019 | Dua & Graff      | Dataset for ML research                    | Structured dataset for supervised learning (lab-based features)                  | Provides benchmark dataset for thyroid classification                | Limited to laboratory data; lacks clinical context (Shen et al., 2023)                                              |
| 4 | Clinical Decision Support Systems Review                    | 2020 | Sutton et al.    | Decision-support systems                   | Literature review (no specific dataset/model)                                    | Highlighted importance of interpretability and usability             | No thyroid-specific application or ML implementation                                                                |
| 5 | Subclinical Thyroid Disease                                 | 2021 | Biondi & Cappola | Thyroid physiology and clinical evaluation | Clinical review based on medical evidence                                        | Provided comprehensive understanding of thyroid dysfunction          | No computational or predictive modeling approach                                                                    |
| 6 | Deep Learning in Healthcare                                 | 2021 | Esteva et al.    | AI in healthcare                           | Deep learning models (primarily imaging datasets) with accuracy-based evaluation | Demonstrated strong performance of AI in disease detection           | Focuses on imaging; limited applicability to structured lab data; interpretability challenges (Miotto et al., 2021) |
| 7 | ML Challenges in Healthcare                                 | 2021 | Sendak et al.    | ML implementation in healthcare            | Analytical review of real-world deployment challenges                            | Identified issues of usability, reliability, and integration         | No disease-specific framework or predictive model                                                                   |
| 8 | Deep Learning Review in Healthcare                          | 2021 | Miotto et al.    | AI in healthcare                           | Literature review across multiple datasets and models                            | Emphasized importance of integrating multiple data sources           | No structured framework for integration; interpretability remains limited                                           |
| 9 | ML for Thyroid Disease Prediction                           | 2022 | Islam et al.     | Thyroid classification                     | UCI dataset, ML models (DT, SVM, etc.),                                          | Achieved strong classification                                       | Focuses mainly on accuracy; limited                                                                                 |

|    |                                     |      |                   |                                       |                                                                                |                                                                             |                                                                                |
|----|-------------------------------------|------|-------------------|---------------------------------------|--------------------------------------------------------------------------------|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------|
|    |                                     |      |                   |                                       | evaluation using Accuracy, Precision, Recall, F1-score                         | performance using lab data                                                  | interpretability (Rajkomar et al., 2022)                                       |
| 10 | AI in Healthcare Review             | 2022 | Jiang et al.      | AI applications in healthcare         | Review-based analysis (no dataset/model)                                       | Highlighted importance of clinical integration of AI                        | No disease-specific implementation or technical framework                      |
| 11 | Machine Learning in Medicine        | 2022 | Rajkomar et al.   | ML in clinical practice               | Review of ML applications in healthcare                                        | Emphasized need for interpretability and clinical trust                     | Does not provide implementation methods for improving interpretability         |
| 12 | ML Comparative Analysis for Thyroid | 2023 | Alshamrani et al. | Thyroid prediction                    | Structured dataset, multiple ML models, evaluated using classification metrics | Demonstrated effectiveness of ML algorithms for thyroid classification      | Relies on lab-based features; lacks interpretability (Shen et al., 2023)       |
| 13 | Advances in Thyroid Diagnosis       | 2023 | Chiovato et al.   | Clinical thyroid diagnosis            | Clinical review of thyroid disorders                                           | Highlighted importance of combining lab results with clinical understanding | No machine learning or predictive modeling approach                            |
| 14 | ML in Endocrine Diseases            | 2023 | Shen et al.       | ML applications in endocrine diseases | Systematic review of datasets, models, and evaluation methods                  | Identified trends in ML-based endocrine prediction                          | Most models rely on lab data and focus on accuracy; limited interpretability   |
| 15 | AI in Endocrine Disease Prediction  | 2024 | Khan et al.       | ML in endocrine diseases              | Comparative ML models with standard evaluation metrics                         | Showed potential of AI in improving prediction performance                  | Limited focus on interpretability and clinical usability (Sendak et al., 2021) |

**Table 2.1 Summary of Related Work**

## **Chapter 3: Proposed Approach (Methodology)**

### **3.1 Introduction**

This chapter details the methodology adopted for designing, developing and evaluating the proposed AI based CDSS for risk assessment in relation to multiple classes of thyroid disorder. The framework integrates machine learning and clinical decision support concepts such that clinicians can use routinely acquired laboratory test results to screen for thyroid disorders. Through the entire process, the aim was to design an intelligible and reliable screening system that aids in early detection while supplementing and not substituting clinician's judgement.

The development process went through a number of phases which included dataset acquisition, data pre-processing, feature selection, development of model, model evaluation and prototype development. The publicly available UCI ANN Thyroid Disease dataset was used as the main data source for this study since it has been used for a long time in benchmarking thyroid diseases classification and consists of routine lab tests. While the original dataset includes numerous clinical features, only five laboratory variables were selected for modelling which include Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), Total Thyroxine (TT4), Thyroxine Uptake (T4U) and Free Thyroxine Index (FTI).

A supervised learning model was then trained on the Extreme Gradient Boosting (XGBoost) algorithm. XGBoost was chosen for the task among other algorithms as it possesses great predictive power, robustness to overfitting, the ability to detect non-linear correlations and applicability to the structured tabular medical

data format. The classifier is a multi-class predictor which predicts one of the three thyroid states, namely Hyperthyroidism, Hypothyroidism and Normal thyroid, and its performance was estimated with such metrics as accuracy, precision, recall, F1-score and confusion matrix analysis.

In order to create a tool which could be used in clinical practice, the trained model was implemented into a fully functional CDSS that utilizes the Streamlit library for interaction purposes. With this application users can input lab values, get an automated risk assessment, estimate prediction certainty, check reference ranges and obtain meaningful recommendation from the model. In addition to the predictive component of the tool a clinical validation layer with reference ranges was introduced.

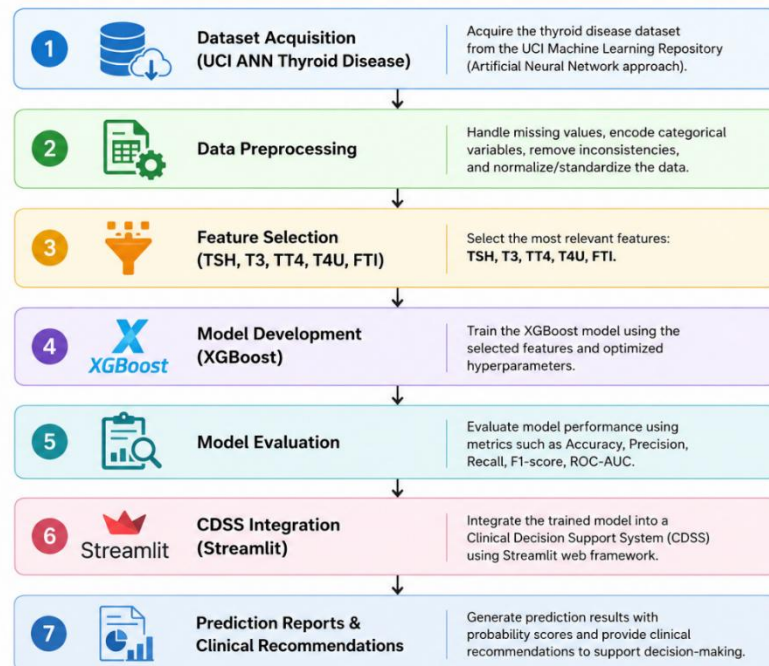
Figure 3.1 gives a brief description of the full flowchart of methodology. It starts from data collection and preparation, continues with feature selection, training of the XGBoost model, and evaluation of the results of its performance. The next step is the implementation of the proposed model into the CDSS and generation of the prediction report. All aforementioned steps represent a decision-making support system that is reliable, understandable, and medical-related.

The following sections of the chapter are organized in the following order. Description of the general research methodology can be found in Section 3.2. Section 3.3 gives information about the system architecture. Selection of the dataset is presented in Section 3.4. Data preprocessing is described in Section 3.5. Laboratory features used in the current study are considered in Section 3.6, where the XGBoost classifier is developed. Section 3.7 contains information on the model evaluation. The CDSS prototype is presented in Section 3.8. Workflow of the system is described in Section 3.9.

### 3.2 Research Methodology

Quantitative and experimental approach was utilized in the development and analysis of an AI-based CDSS for multiple classes risk assessment of thyroid diseases. The primary aim of the research was to develop a classification system which could assign the patient to be affected with Hyperthyroidism, Hypothyroidism or to be Normal according to lab results. Supervised Machine Learning approaches and clinical validation reasoning were applied to design a reliable, comprehensible and feasible decision support system.

The methodology has been carried out through a series of interrelated activities, starting from data gathering up to implementation of a system prototype, all of which were designed in order to provide the data quality, prediction accuracy, reliability and clinical relevance of the developed model. Figure 3.1 illustrates the research methodology.



**Figure 3.1 Overall Research Methodology**

Acquisition of the dataset was performed in the first stage, where the UCI ANN Thyroid Disease Dataset was acquired from the UCI Machine Learning Repository. The dataset is benchmark for thyroid disease prediction and it consists of laboratory data, which are applied in predicting the disease. However, since other attributes are present in the original data set, only the laboratory values, which were important for predicting thyroid disease were selected.

The next stage consisted of preprocessing of the acquired data. In the process of preprocessing, analysis of data structure was performed, relevant laboratory values were selected, data classes were identified and the categorical class labels were transformed to numeric class labels.

Next, feature selection procedure was conducted. According to the results of clinical practice as well as the literature review on the endocrinology field, five lab features were chosen as the predictors: TSH, T3, TT4, T4U, and FTI. These lab tests are designed to measure the overall activity of thyroid and contain necessary physiological data to identify normal thyroid activity as well as hypothyroidism and hyperthyroidism. In addition, five predictors are sufficient for good performance of the model and increase its interpretability.

After that, an XGBoost classifier was constructed using these predictors. XGBoost is the ensemble machine learning approach based on gradient-boosted decision trees and is characterized by high accuracy, computational efficiency, as well as the possibility of learning complex non-linear relationships between structured data. This algorithm builds the sequence of decision trees, where each new tree corrects mistakes of the previous trees and learns how to predict the most probable class of thyroid based on their results.

A separate validation data set helped during the learning process and offered an early hint at the level of generalization performance prior to testing the model against the test data set. Measures employed to test the performance of the model included accuracy, precision, recall, F1-Score, and confusion matrices. These are measures that give information regarding overall performance as well as class separation.

The classifier was employed through the use of the Streamlit library while developing an interface for CDSS. Here, users would be able to enter the lab values and get the predictions alongside the interpretability recommendations. This was intentionally done in the manner of a healthcare dashboard interface.

First, the proposed prototype incorporates a two-step prediction process which sets it apart from other existing thyroid prediction models. If normalized data, similar to those in the UCI dataset, are employed in the analysis, the prediction is made by the trained XGBoost classifier. If the data in the form of real clinical cases are used, there is one additional layer which involves clinical validation in the form of reference ranges. That guarantees the uniformity of processing for both research and clinical data as well as transparency of the output.

Finally, the whole model was verified with both unseen dataset entries and clinically realistic laboratory values. The output of the process includes predicted class, prediction confidence level, probabilities for all classes, reference range verification, clinical interpretation, and recommendation message. This output proves that clinical decision support and machine learning can be combined to aid physicians in the analysis of thyroid disorders.

That makes the developed model clinically and technically sound and implementable in the form of decision support software.

### 3.3 Proposed System Architecture

The CDSS proposed in this work was created as a modular system that uses the ideas of machine learning and clinical decision support systems in order to predict the existence of a thyroid disease. The structure of the system may be considered as layered, in which every element performs its own job and then passes it to another element. In this way, a prediction is obtained.

There are six layers in this architecture - Data Input Layer, Input Processing Layer, Prediction Engine, Clinical Validation Layer, Decision Support Layer and User Interface Layer. Together, they convert lab data into clinically usable recommendation.

Data Input Layer is the initial layer that accepts input from the user. The user should provide the demographic information about the patient - age and gender, and also laboratory data of five markers - TSH, T3, TT4, T4U and FTI. The latter markers were chosen due to their relevance to the clinical state and availability.

The Input Processing Layer checks whether the entered values are consistent and what kind of values they are. When the values are entered in the normalized form of the UCI dataset, they are passed to the machine learning engine; however, when the values are entered in the standard clinical units, the clinical validation module is automatically activated. Therefore, the prototype will be able to handle both the research data and realistic clinical scenario automatically.

The Prediction Engine acts as the main AI algorithm, and it uses an XGBoost classifier on five laboratory variables. The classifier learned the correlation between the values of laboratory variables and thyroid diseases from the labeled data during the training. In turn, when receiving the input feature vector, the classifier computes the probabilities for each class and selects the most probable

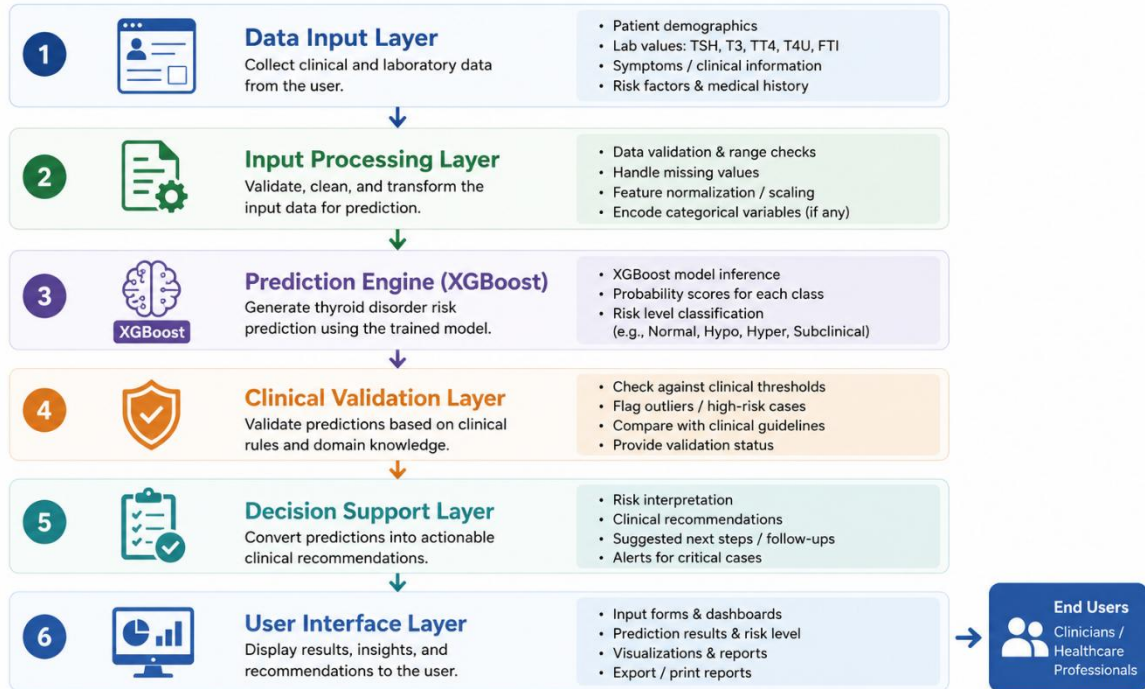
one. Additionally, the probabilities are used to calculate the confidence score presented to the user.

The Clinical Validation Layer improves the interpretability of the prediction through comparison of the values of laboratory variables with the known thyroid reference ranges. Instead of replacing the model, the clinical validation layer is supposed to add more interpretation to the prediction.

The outputs of the Prediction Engine and the Clinical Validation Layer are then taken to the Decision Support Layer where predictions, confidence, reference range, and validation come together to form one assessment. The Decision Support Layer gives an interpretable diagnosis, a confidence estimate, a risk category, and recommendation message based on the aforementioned components.

Lastly, the final layer is the User Interface Layer which was developed with Streamlit. With the help of this web-based user interface, users enter patient information, laboratory values, prediction confidence, reference ranges, and finally download the assessment report. The approach of this design makes the process transparent with respect to model predictions.

In terms of designing the software, we have several advantages that come up. First, we get modularity, which means that our system will be easier to maintain due to the ability of changing individual layers without affecting the others. In addition, we leave some room for potential improvements, such as integration of laboratory parameters, other machine learning algorithms, EHRs, etc.



**Figure 3.2 Proposed System Architecture**

### 3.4 Dataset Description

It holds true when considering a predictive model that is dependent on the training set of data while making decisions; hence choosing a suitable dataset is one of the key requirements of developing a successful classifier. In this case, the Artificial Neural Network Thyroid Disease dataset from the University of California Irvine Machine Learning Repository was used as the main dataset.

This dataset has been extensively used in the domain of machine learning and has been used as a standard for classifying diseases of the thyroid gland. The selected dataset contains information about patients obtained during an examination of their thyroid function and consists of demographic, laboratory test, and symptoms of the disease.

Many of the initial attributes can be seen as categorical, administrative or variables that add no additional information about thyroid classification. To make

the model interpretable and practical, this study uses only those laboratory parameters that are routinely collected as part of thyroid function tests.

The five selected variables will serve as the input data for the algorithm:

- Thyroid Stimulating Hormone (TSH)
- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)
- Free Thyroxine Index (FTI)

This set of variables represents a core set of biochemical tests conducted by physicians for evaluation of the thyroid gland status. This set of variables provides sufficient physiological information about patients' thyroid glands to identify whether a thyroid gland is healthy or not.

The UCI ANN Thyroid Disease dataset not only comprises laboratory bio-markers but also other demographic or clinical characteristics such as age, sex, drugs, pregnancy, surgery and various binary symptoms. But for this research only five laboratory markers: TSH, T3, TT4, T4U, FTI were selected since they are key biochemical markers that are usually used for diagnosing thyroid disorders.

The rest of the characteristics have been intentionally excluded for three reasons. Firstly, some of them are not the actual signs of thyroid dysfunction but are either indirect factors or just historical data. Secondly, some of them include missing values or are binary features extracted from questionnaires that will raise the data variability. Thirdly, the aim of the current research was to build a simple and efficient AI-based Clinical Decision Support System that is able to perform predictions solely on the basis of laboratory tests.

In addition to this, the usage of only five laboratory biomarkers made the model more concise yet did not affect its classification accuracy which remained at 97.23%, which means that all the necessary information is contained in these biomarkers in order to make an accurate classification of the thyroid diseases.

The predictive classification problem can be described as follows: the problem consists of classifying objects into three non-intersecting classes:

- Hyperthyroid
- Hypothyroid
- Normal

Every object belongs to one of these classes. In order to build the classifier for the supervised learning, diagnostic classes have been transformed into numeric values.

Before the classifier was trained, the data set has been split into the training and testing sets. The former is used to identify patterns between laboratory variables and the diagnosis while the latter is used for the evaluation of the model performance.

stored in the dataset's own preprocessed representation rather than in standard clinical reporting units, which means the model performs best on data resembling the original dataset. To address this, the prototype includes a clinical validation layer that supports interpretation of real clinical values alongside the model's predictions.

There are three principal grounds for the selection of the dataset. First, it is open source, so the methodology can be easily replicated by anyone. Secondly, it has been widely used in the prior research on thyroid prediction, so comparison can

be made with the existing findings in the literature. Finally, it contains measurements which were performed in practice and, therefore, make the model applicable to the healthcare context.

There are, however, several disadvantages of the dataset which should be noted. Firstly, it is an old dataset which was collected in certain clinical conditions which may not reflect the current patient population and laboratory measurements. Secondly, the laboratory measurements in the dataset are presented in its own internal format and not in clinical reporting format, which means that the model performs well only with the data of the same format. This is why the clinical validation component was added to the prototype.

Overall, the use of UCI ANN Thyroid Disease Dataset serves as a good ground for the development of the CDSS.

### **3.5 Data Preprocessing**

Preprocessing of data is a vital stage of machine learning process as the quality of the input data directly affects the quality of the resulting model. Data sets relevant to the medical domain possess redundant features, categorical variables, missing values, and inconsistencies that should be addressed before modelling. Therefore, the data preprocessing approach was applied to preprocess the UCI ANN Thyroid Disease data set.

The goal was to convert the initial data to clinically meaningful data for modelling. The flowchart of data preprocessing is shown on figure 3.4.

#### **3.5.1 Dataset Inspection**

The first step of data preprocessing was the analysis of the data set. The aim was to determine the data type, the clinical significance and suitability of the feature for the predictive model. The data set includes demographic data, laboratory results, Boolean values, referrals and diagnosis.

The initial statistical analysis provided information about the distribution, inconsistency and integrity of the data.

#### **3.5.2 Feature Extraction**

Laboratory variables linked directly to the functioning of the thyroid gland have been chosen, with the primary emphasis being on the biochemical parameters that are analyzed during the routine examination of the thyroid gland. The following laboratory parameters were selected:

- Thyroid Stimulating Hormone (TSH)
- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)

- Free Thyroxine Index (FTI)

These parameters are considered to be the most important ones for evaluation of the thyroid gland functionality, and their selection will not only allow for better interpretability but also reduce computational costs. Small number of five clinical variables will result in fast execution, easier understanding and better chances of practical implementation.

### 3.5.3 Target Variable Preparation

As the problem is to classify data into multiple classes, the diagnostic label column was chosen as a target one. The classes are the following:

- Hyperthyroid
- Hypothyroid
- Normal

Since machine learning algorithms require numeric values for target values, the classes were coded numerically.

| Original Class | Encoded Label |
|----------------|---------------|
| Hyperthyroid   | 0             |
| Hypothyroid    | 1             |
| Normal         | 2             |

The encoding was applied only for training purposes. The predicted numerical values within the CDSS are then encoded back to their corresponding clinical values.

### 3.5.4 Data Partitioning

To generalize the model, two distinct datasets were utilized: the training dataset, in which the XGBoost algorithm learned the correlation between the lab tests and

the diagnosis, and the testing dataset, which was not involved in the learning phase of the model but used to measure the performance of the model on new observations. The isolation of the testing data from the training dataset prevents an overly optimistic assessment of the model's performance.

### **3.5.5 Data Representation**

The values of the laboratory parameters have been preserved in their original form from the UCI ANN Thyroid Disease database, and therefore, they are different from the values provided in any clinical laboratory database. The values of TSH are normalized and not in clinical units, and thus, the learning of the model takes place based on such kind of data.

In order to increase the utility of the model, a clinical validation layer has been added to the CDSS. Thus, the clinical validation and interpretation along with prediction can be done by the user in case they give the input values in clinical units. Thus, it helps the model to learn using two processes together, increasing its utility without hampering its performance.

### **3.5.6 Preprocessed Dataset**

The final form of the dataset obtained after preprocessing contained five laboratory variables as the predictors and one target variable indicating the thyroid disease class. This dataset is used as an input to the XGBoost classifier.

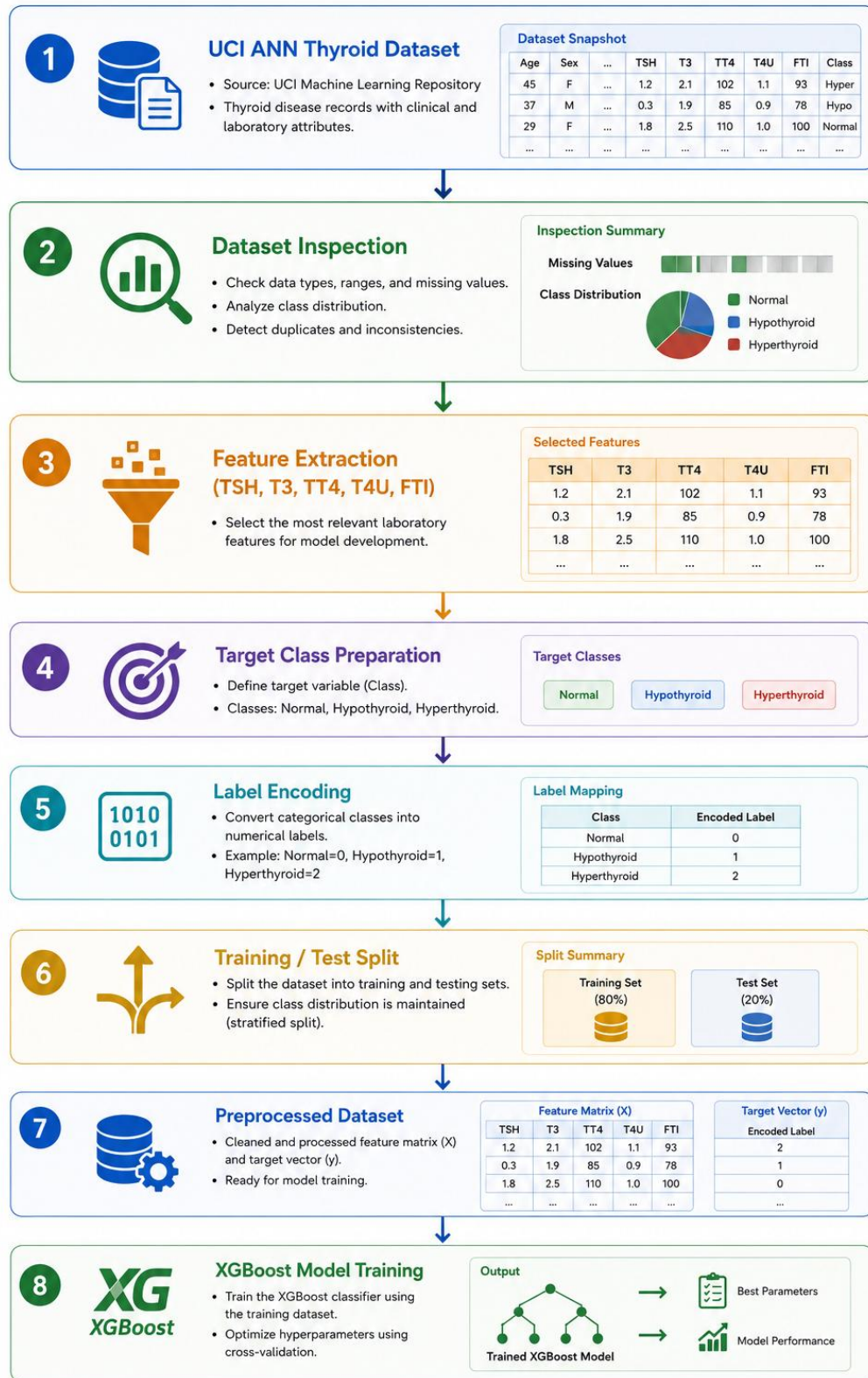


Figure 3.3 Data Preprocessing Workflow

### **3.6 Machine Learning Model Development**

The main intelligence of the CDSS is the supervised learning model which does the multi-class classification of the thyroid diseases. After the preprocessing step, the preprocessed data was used to develop an algorithm that finds the relationship between the laboratory test results and the class. A very crucial aspect is having a classifier that is able to classify the patients in one of the three classes which are; Hyperthyroidism, Hypothyroidism, and Normal Thyroid Function.

Some of the past researches aimed at developing a classifier for the classification of the thyroid gland using some algorithms like Decision Tree, SVM, Artificial Neural Networks, Random Forest and Gradient Boosting. Each algorithm has its own advantages, and the most appropriate algorithm will be the one that gives optimal accuracy, stability, efficiency and appropriateness for the dataset. From many algorithms which I have considered early in my research, I decided to use XGBoost as my classifier.

#### **3.6.1 Selection of the XGBoost Algorithm**

XGBoost is a complex version of gradient boosting, which improves the prediction performance and maintains computational efficiency. While using the only decision tree implies that the prediction comes from only one hierarchical tree structure, XGBoost employs the ensemble of trees, where each tree adjusts the mistakes of the previous ones in order to gradually enhance the prediction performance.

There are several factors that motivated the choice of the algorithm. Firstly, the XGBoost model works extremely well for structured tabular data such as UCI ANN Thyroid Disease Dataset since thyroid measurements include numerical values and thus, the tree-based model will be more effective than deep learning

methods. Secondly, the built-in regularizations provided by XGBoost will minimize the risk of overfitting in case of the medical data that can have quite complicated variable relations. Thirdly, the model captures the non-linear relationships between the biomarkers, which cannot be observed in linear models. Finally, the model predicts the classification labels and probabilities that can be used as the confidence scores in the CDSS.

### **3.6.2 Model Training Process**

The training procedure was performed according to the principle of supervised learning. Every case included five lab tests as input parameters and the disease diagnosis as output parameter:

- Thyroid Stimulating Hormone (TSH)
- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)
- Free Thyroxine Index (FTI)

The algorithm learns the rules for mapping inputs to outputs according to the labeled examples. Contrary to conventional programming, it was not necessary to explain to the algorithm the principles of diagnosing this disease – the model has been able to do that just relying on its experience. With every boost, the new decision tree is added to the model in order to correct the errors made by the previous one.

### **3.6.3 Hyperparameter Configuration**

XGBoost relies on many hyperparameters that determine the complexity of the tree, its learning rate, and the bias-variance trade-off. The final hyperparameters setting was determined through experimentation, taking into account the balance

between prediction power and generalization. The list of the main hyperparameters is provided in Table 3.4.

| Parameter                         | Description                | Value                |
|-----------------------------------|----------------------------|----------------------|
| Algorithm                         | Classification model       | XGBoost              |
| Objective function                | Multi-class classification | multi:softprob       |
| Number of classes                 | Output classes             | 3                    |
| Number of trees<br>(n_estimators) | Total boosting iterations  | 100                  |
| Maximum tree depth<br>(max_depth) | Maximum depth of each tree | 6                    |
| Learning rate (learning<br>rate)  | Step size during boosting  | 0.1                  |
| Evaluation metric                 | Performance measure        | Multi-class log loss |

**Table 3.1 XGBoost Hyperparameter Configuration**

### 3.6.4 Prediction Mechanism

assessments while keeping the confidence behind each prediction visible.

After being trained, the model is then used within the CDSS to conduct real-time inference. When laboratory data is inputted through the interface, the system creates the feature vector based on five variables and passes it to the model as the input. Prediction involves the following steps:

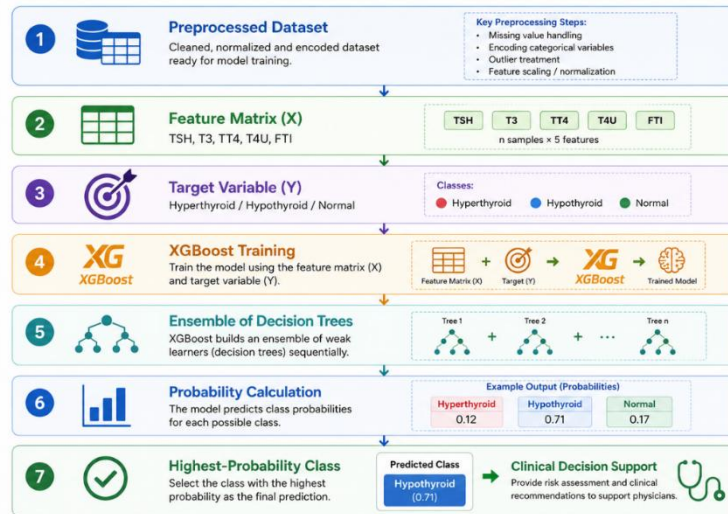
- Feature vector construction: input transformation into the numeric vector, which serves as the input of the model.
- Model inference: the model makes predictions using the ensemble, wherein each decision tree analyses the input based on the learned knowledge.

- Probability estimation: results of the decision trees are processed in order to determine the probabilities of classes, and the one that has the highest probability becomes the output of the prediction.
- Clinical interpretation: prediction, its confidence score, and probability distribution are passed to the Decision Support Layer, which transforms it into an interpretable report.

The described process ensures quick and consistent predictions with the clear level of confidence.

### **3.6.5 Integration into the Clinical Decision Support System**

The trained machine learning model is serialized and then added to the Streamlit application through the usage of the Joblib library. This will be done only once during initialization of the system hence no need to train again during the predictions. The prediction component works separately from the user interface component. While the user interface component gathers patient information and laboratory information, the prediction component does the classification. This separation makes it easier to make maintenance possible. Furthermore, it becomes possible to change or update the model without making any changes to the user interface component. The prediction component also works together with the clinical validation component which makes it possible to explain the measurement based on the existing references.



**Figure 3.4 XGBoost Model Development Process**

### 3.7 Model Evaluation

Nevertheless, the assessment of the model prior to applying it to the CDSS will be needed in order to ensure the quality of the model. Since our goal is a thyroid disorders classifier, its performance must be tested on the unseen dataset to verify whether it is able to generalize well.

For this reason, after training the XGBoost model, it was tested on the unseen test set which played no part in the process of creating our classifier. This way it is possible to avoid any bias during the process of assessing the model's performance, and receive accurate figures proving the generalizability of our model.

However, in our case, there is a multi-class classification problem, and one evaluation metric will not be enough to provide an answer on the performance of our classifier. That is why, five metrics have been used for the purpose of evaluation, namely, accuracy, precision, recall, F1-score, and confusion matrix. They are provided in Table 3.5 and Figure 3.6.

### 3.7.1 Classification Accuracy

Classification accuracy is the ratio of the number of records classified correctly to the total number of records that have been classified. Classification accuracy provides an overall measure of performance and is one of the most widely used measures in supervised learning. Classification accuracy is calculated as follows:

$$\text{Accuracy} = (\text{Number of Correct Predictions}) / (\text{Total Number of Predictions})$$

If the value of classification accuracy is high, it shows that the classifier classifies the records into their correct categories. But classification accuracy is not a reliable measure by itself.

### 3.7.2 Precision

Precision is the measure of the accuracy of the classifier's positive predictions, or how many of those objects predicted to belong to the class actually do. Precision is calculated as:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}),$$

where TP is true positive and FP is false positive. The higher the precision of the classifier is, the smaller its chances of creating any false positives are.

### 3.7.3 Recall

Recall is also called sensitivity or true positive rate. It characterizes the ability of the classifier to detect those objects that actually belong to a certain class. Recall is calculated as:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}),$$

where TP is true positive and FN is false negative. The importance of recall is especially relevant in the field of medicine due to potential serious consequences of a mistake.

#### **3.7.4 F1-Score**

The score takes account of both precision and recall and is good for multiclass problems because of its combination of accuracy and coverage. The following formula is used to calculate the F1-score:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}).$$

Unlike accuracy, the F1-score also takes account of both false negatives and false positives.

#### **3.7.5 Confusion Matrix**

The confusion matrix allows a more detailed assessment of the performance of the classifier through providing data about correct and incorrect classification results for each individual class. For a multiclass problem, the matrix provides information about how many times records of a certain class were assigned to some other class. This will allow identifying which class is harder to classify, and this gives additional insight into possible mistakes, which may be critical in the healthcare sector because of the clinical context of such misclassifications.

#### **3.7.6 Classification Report**

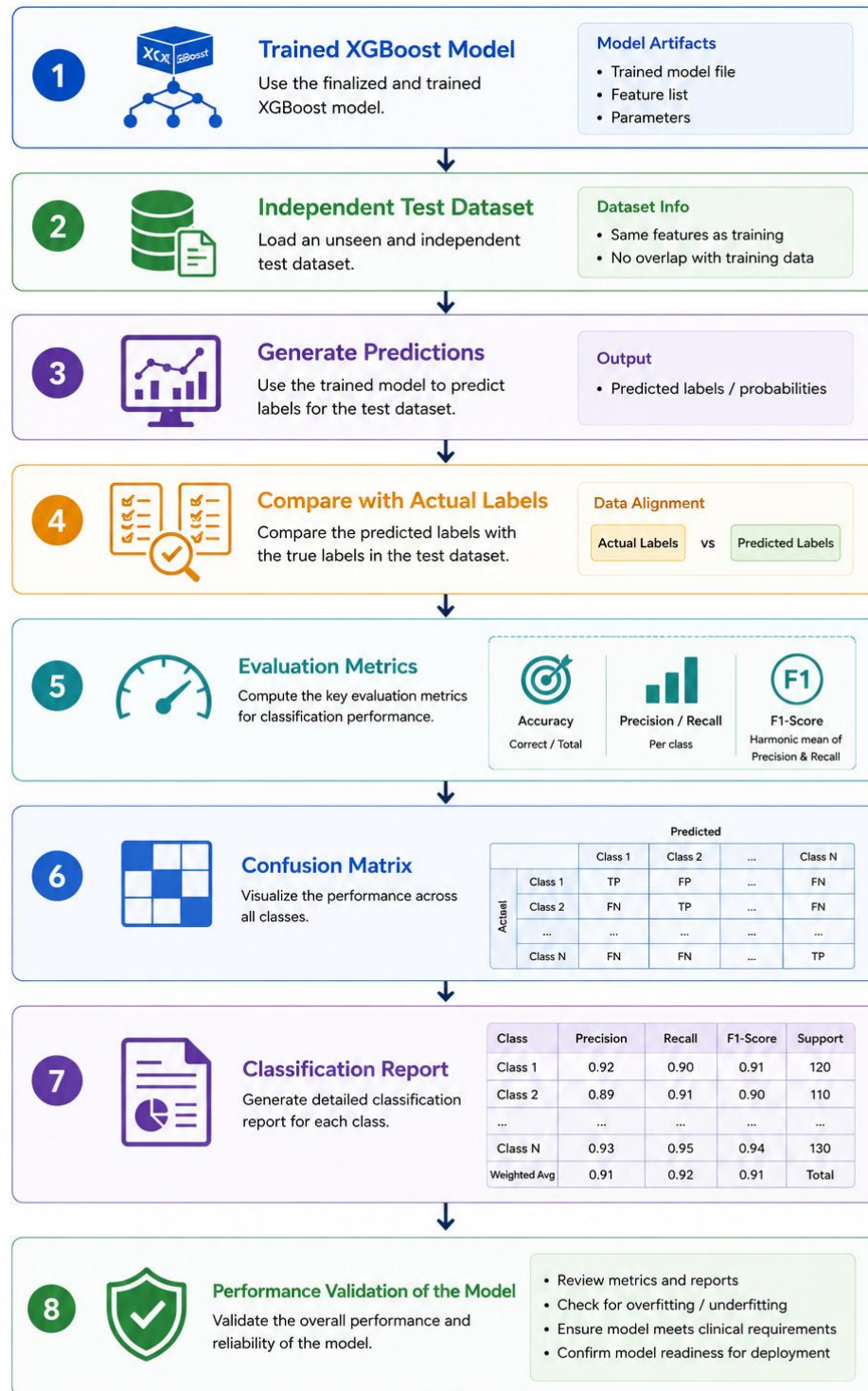
A classification report was also created. It displays information on precision, recall, f-measure and support for each class, which allows a better insight into how accurate the classifier is, information which the accuracy alone cannot provide. That allows a better understanding of the classifier's performance on predicting Hyperthyroidism, Hypothyroidism and Normal thyroid.

### 3.7.7 Integration of Evaluation Results

However, the results acquired through the metrics prove that the XGBoost model already achieved acceptable performance levels before being incorporated into the CDSS. The model was only exported and implemented into the Streamlit prototype when it successfully passed the test on an independent test set. In the application, the model does not only provide the classification result but also the probability of belonging to every possible category, and that is what determines the confidence score in the interface.

| Metric                | Purpose                                         | Importance                                               |
|-----------------------|-------------------------------------------------|----------------------------------------------------------|
| Accuracy              | Measures overall classification performance     | Provides an overall indication of prediction correctness |
| Precision             | Measures correctness of positive predictions    | Reduces false positive diagnoses                         |
| Recall                | Measures ability to detect actual disease cases | Reduces missed thyroid disorder cases                    |
| F1-score              | Balances precision and recall                   | Provides a comprehensive measure of performance          |
| Confusion Matrix      | Visualizes classification outcomes              | Identifies patterns of correct and incorrect predictions |
| Classification Report | Summarizes metrics for each class               | Enables detailed evaluation of individual classes        |

**Table 3.5 Performance Evaluation Metrics**



**Figure 3.5 Model Evaluation Workflow**

### **3.8 Clinical Decision Support System Development**

Lastly, the model was integrated into a CDSS with a user interface. Despite the fact that the XGBoost classifier is used for generating predictions, it is also necessary to design an interface that would allow for presenting the predictions in a clinically relevant manner to the clinicians. Therefore, an interactive CDSS was designed as a prototype in order to bridge the gap between the outputs of the model and the decision of the clinician.

For developing the CDSS, the Streamlit package, which is an open source and publicly available Python package for creating interactive data science and machine learning applications, was used. The reason behind using the Streamlit package lies in the possibility to use trained models with intuitive user interfaces and to deploy the application in real time, which are key aspects of a proof-of-concept application in the healthcare domain. In order not to write Python code to launch the application, the prototype allows the user to use a dashboard for manually entering laboratory values, and then the system will perform the assessment and present its output in the form of a report.

#### **3.8.1 User Interface Design**

The interface was developed based on the principles of simplicity and ease-of-use and includes sections which guide the user through prediction procedure. First of all, demographic information, including ID, age and gender of a patient is collected. These variables are not currently utilized as predictors in our system, although, they can be seen as an additional information for the report and make our prototype more realistic. Secondly, five laboratory measurements are introduced for prediction:

- Thyroid Stimulating Hormone (TSH)

- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)
- Free Thyroxine Index (FTI)

All these variables are entered using numeric fields and this method allows reducing number of errors during data entry. Before starting the prediction procedure, the interface displays automatically the type of entered data distinguishing between normalized UCI values and clinical units.

### **3.8.2 Prediction Engine Integration**

The prediction engine is the core AI component of the CDSS. The machine learning model, after training, is saved using Joblib and then loaded at the beginning of the application. When the evaluation process starts, the input values get transformed into a feature vector and evaluated using the prediction engine that computes the probabilities for all classes. The most probable class gets selected as the output prediction, while the whole probability distribution and maximal probability are returned together, which constitutes the confidence score. Maintaining the separation between the prediction algorithm and the graphical user interface enables us to keep the modularity of this CDSS system, allowing for the retraining or replacing the machine learning model.

### **3.8.3 Clinical Validation Layer**

The other unique feature of this CDSS system is the clinical validation layer. The training dataset used normalization based on UCI values, but clinicians use reference ranges to interpret results. In order to align the two approaches, the rule-based clinical validation module compares each measured value to thyroid reference ranges and returns information on whether the measure falls inside or

outside the range. The information provided by this module does not substitute the prediction made by the machine learning algorithm.

#### **3.8.4 Assessment Report Generation**

This turns numerical model output into clinically understandable information that supports thyroid assessment.

After the prediction stage, there will be a creation of a structured report based on the findings of the model and the validation phase in order to provide a complete overview of the assessment. The report will include the following information:

- Patient Data
- Laboratory Values
- Type of input recognized
- Engine of Prediction used
- Class of Thyroid predicted
- Confidence level of the prediction
- Probability Distribution of all classes
- Comparison of Laboratory values with reference range
- Clinical significance
- Recommendation message

#### **3.8.5 Explainability and User Interpretation**

Having interpretability in an AI system in application to healthcare is essential because medical professionals do not want to depend on predictions which cannot be interpreted. This is why CDSS contains some methods which help make the process more interpretable. Apart from the predicted class, there is the level of confidence of the prediction, the probability of each class, and the comparison of laboratory test results to the norms which means that not only the

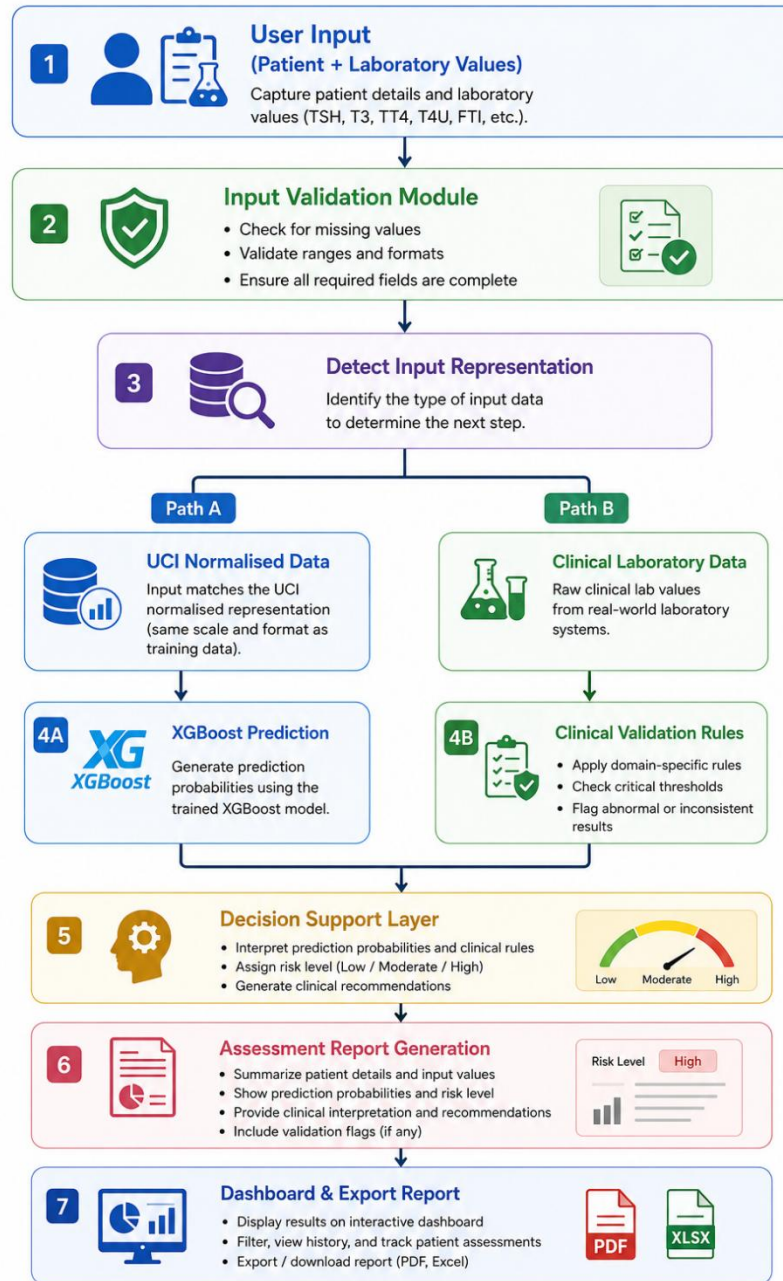
result but also the reasons behind it are available to the users. The facts which led to such a conclusion and the corresponding recommendation are described as well.

**3.8.6 Prototype Workflow**

Firstly, this process begins with the entry of the patient information and lab measurements data. The second step is where the validation part of the system determines whether the values entered are normalized data from research or normal clinical data. After validation of the two, the predictor uses them to come up with the output in terms of prediction and probabilities and at the same time validating the clinical validation module in terms of measuring the value against the reference ranges. In the end, this will form one report by the Decision Support Layer.

| Component                 | Function                                                 |
|---------------------------|----------------------------------------------------------|
| User Interface            | Collects patient information and laboratory measurements |
| Input Validation Module   | Verifies entered values and detects input type           |
| XGBoost Prediction Engine | Performs multi-class thyroid classification              |
| Clinical Validation Layer | Compares laboratory values with reference ranges         |
| Decision Support Layer    | Combines AI predictions with clinical interpretation     |
| Report Generation Module  | Produces a structured assessment report for the user     |

**Table 3.2 Main Components of the Clinical Decision Support System**



**Figure 3.6 Clinical Decision Support System Workflow**

The major elements of the system and their functions are listed in Table 3.6, whereas the workings of these elements during evaluation are described through Figure 3.7.

### 3.9 System Workflow

The suggested CDSS utilizes a systematic workflow where data acquisition, predictions, clinical validation, and output presentation are all integrated into one system. Such a design enables the user to make the assessment efficiently yet in a clear manner. This can be seen from Figure 3.8 below which presents the whole workflow of operation.

The process begins when the user accesses the Streamlit interface and proceeds to enter the patient information, which involves the identifier, age, and gender. Though the data is not considered during the prediction, it is included in the report for documentation purposes.

This is followed by the laboratory input, which includes the five predictive measurements:

- Thyroid Stimulating Hormone (TSH)
- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)
- Free Thyroxine Index (FTI)

These are the parameters used in order to generate the feature vector that will be analyzed by the prediction engine. Once all the values have been entered, the software checks to see whether the input values are complete and numeric, and whether they fit into the normalized UCI format or the clinical units format. The automatic check makes testing possible in both scenarios without having to manually toggle between the two modes.

In case the input values fit into the normalized UCI format, the feature vector is fed into the XGBoost classifier, where the measurements are analyzed, probabilities are calculated for each class and the one with the greatest probability

serves as the diagnosis while the probability is the confidence level. In case the input values are in clinical units format, then the Clinical Validation Layer is used instead.

The prediction itself and the output of the Clinical Validation Layer will be then transferred to the Decision Support Layer for combination with the other mentioned information: prediction, confidence score, probability distribution, reference comparison, and rule-based interpretation, transforming numerical prediction into clinically relevant information. The final step is the creation of Streamlit dashboard presenting the prediction of disease, confidence score, probability distribution, reference comparison, interpretation, and recommendation with the option to export the results into a report.

Thus, it can be seen how clinical decision support and AI work together in a synergetic manner. In contrast to making separate decisions, the proposed system operates as an intelligent assistant interpreting measurements based on the knowledge from history.

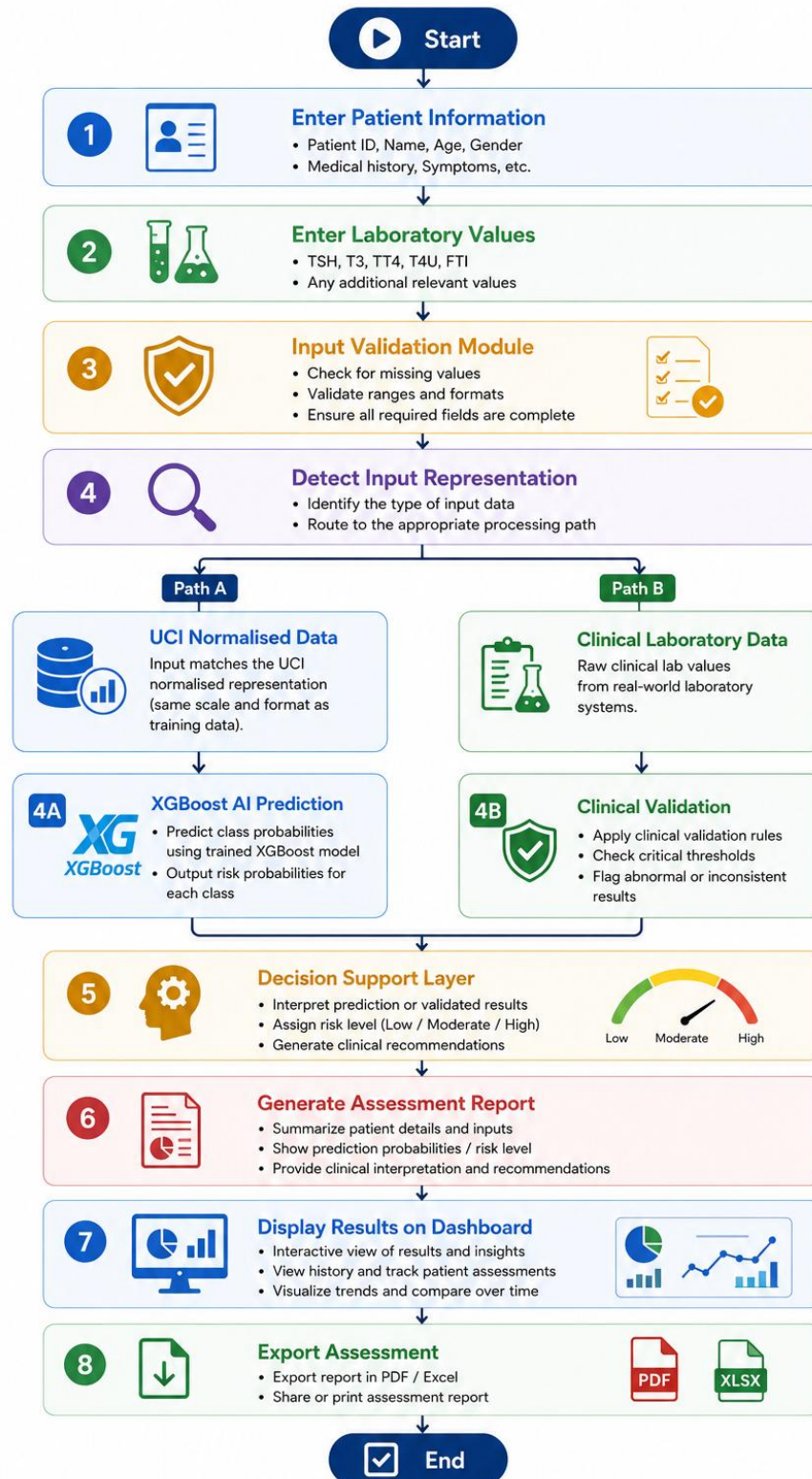


Figure 3.7 Overall System Workflow

### **3.10 Chapter Summary**

In this chapter, the methodology for developing the proposed AI-based CDSS for risk assessment of multiple classes of thyroid diseases has been elaborated. The methodology comprises of selection of dataset, preprocessing, feature extraction, model development, model evaluation, and development of a web-based prototype of the CDSS. The dataset for training purposes was chosen from UCI ANN Thyroid Disease. The number of predictors in this case would be five lab parameters. The multi-class classifier based on XGBoost algorithm has been developed and analyzed with respect to accuracy, precision, recall, F1-score, and confusion matrix.

In order to develop the CDSS, the model has been implemented in Streamlit framework which allows to provide input regarding the lab data, conduct risk assessment, calculate confidence, check references, and receive recommendations. Besides, a clinical validation layer has been developed which facilitates interpretation of results via comparison with reference values.

Thus, the above-mentioned methodology can be regarded as the technical part of the framework and serves for the evaluation of its performance. Further chapters will cover the implementation of the framework and analysis of its performance.

# **Chapter 4: Experimental Results and Discussion**

## **4.1 Introduction**

In this chapter, the realization, experimental evaluation and performance analysis of the AI-based CDSS for multi-class risk assessment of thyroid diseases is done. This chapter intends to illustrate the realization of the methodology discussed in chapter 3 and to evaluate the performance of the framework depending on the results acquired from the trained model and the developed prototype.

The chapter begins with a discussion on the experimental environment and software that has been used during the development process. The chapter discusses the implementation of the UCI ANN Thyroid Disease dataset and the development and training of the XGBoost classifier. The performance of the classifier is evaluated depending on the accuracy, precision, recall, F1 score and confusion matrix.

After evaluating the model, the chapter presents the realization of the web-based CDSS with Streamlit. The chapter explains how the predictions from the model are integrated in the healthcare application for assisting in the thyroid disease assessment by the clinicians.

Practicality was evaluated by testing the prototype with both standardized UCI entries and lab data that mimics real clinical data, proving that it can process varied types of input data while generating intelligible output. This chapter concludes with a discussion on the merits and weaknesses of the framework as well as potential future steps.

## 4.2 System Implementation

This chapter explains the process of developing the methodology presented in Chapter 3 as a prototype for assessing thyroid disease risks. The developed system integrates data processing, forecasting, validation, and user interface into one system. A modular architecture was utilized in order to increase maintainability, scalability, and flexibility of the software as new functionalities may be introduced in the future.

All the components of the system have been created in Python due to the great variety of libraries related to machine learning, data analysis, and web development. Data exploration, processing, and modeling were done using Jupyter notebook. Development was done using Visual Studio Code and Anaconda distribution. CDSS was implemented in Streamlit which allows deploying machine learning models as web applications with minimum configuration.

The machine learning library used was XGBoost. The trained model was saved using Joblib so that it can be serialized and used further without any re-training during each session. The following libraries were used for data processing and manipulation: pandas for data manipulation and preprocessing; numpy for mathematical operations and vectorization of features; scikit-learn for calculating performance metrics: accuracy, precision, recall, F1 score, confusion matrix, classification report. Polly created the interactive visualization of confidence scores and prediction probabilities on the dashboard.

The development process was composed of five main steps. The first step entailed the loading and pre-processing of the data. The second step involved the development and training of the XGBoost classifier on the selected features. The third step entailed the evaluation of the classifier using an independent test

dataset. The fourth step entailed the integration of the classifier into the Streamlit CDSS, whereas the fifth step entailed the validation of the whole prototype using the normalized UCI and clinical data. Figure 4.1 shows below.

The development environment can be seen in Table 4.1, whereas the major software components and their respective functionality are described in Table 4.2. The whole software architecture is given in Figure 4.2.

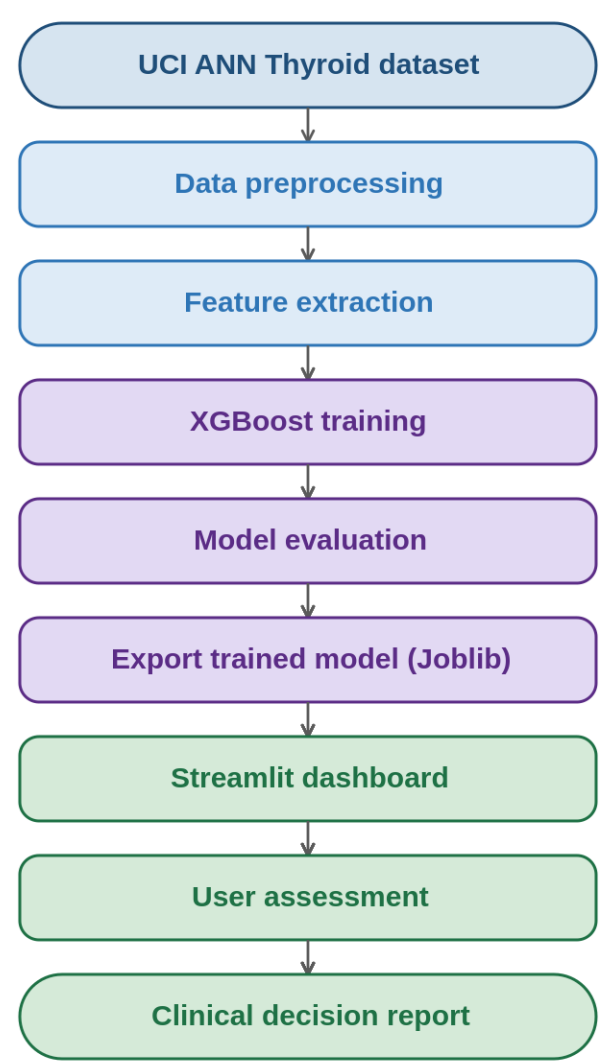
| Component                | Specification      |
|--------------------------|--------------------|
| Programming language     | Python 3.12.7      |
| Development environment  | Anaconda           |
| Notebook environment     | Jupyter Notebook   |
| Code editor              | Visual Studio Code |
| Machine learning library | XGBoost            |
| Data processing          | Pandas             |
| Numerical computing      | NumPy              |
| Model evaluation         | Scikit-learn       |
| User interface framework | Streamlit          |
| Data visualization       | Plotly             |
| Model serialization      | Joblib             |
| Operating system         | Windows 11         |

**Table 4.1 Development Environment**

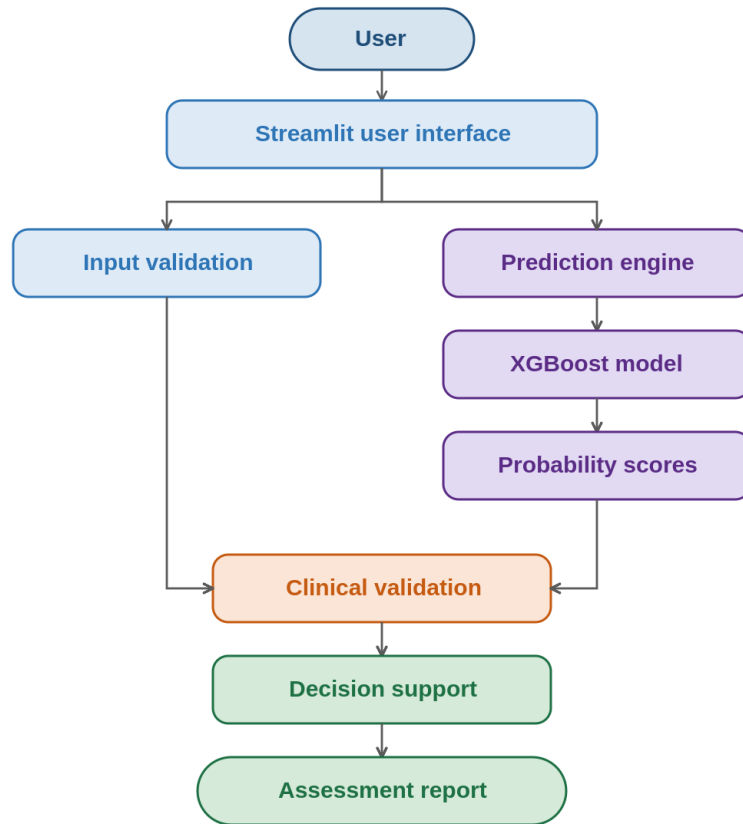
| Component                 | Function                                                                    |
|---------------------------|-----------------------------------------------------------------------------|
| UCI ANN Thyroid Dataset   | Provides historical patient records used for model development              |
| Data Preprocessing Module | Extracts selected laboratory features and prepares the dataset for training |
| XGBoost Classifier        | Learns relationships between laboratory measurements and thyroid disorders  |
| Prediction Engine         | Performs real-time multi-class classification                               |

| Component                  | Function                                                  |
|----------------------------|-----------------------------------------------------------|
| Clinical Validation Module | Compares laboratory values with clinical reference ranges |
| Decision Support Module    | Generates interpretation and recommendations              |
| Streamlit Dashboard        | Provides the graphical user interface                     |
| Report Generation Module   | Produces downloadable assessment reports                  |

**Table 4.2 Main Software Components**



**Figure 4.1 System Implementation Workflow**



**Figure 4.2 Software Architecture**

### 4.3 Experimental Setup

This subsection outlines the setting of development, training, and evaluation of the proposed CDSS model based on XGBoost. It was ensured that the model training was performed under controlled conditions and tested on a separate test set. All experiments have been carried out in the Anaconda environment running Python and using Jupyter Notebook for data pre-processing, development, and evaluation. The process included five steps sequentially: dataset preparation, feature extraction, model training, model evaluation, and integration into the prototype of CDSS, which is presented in Fig. 4.3.

#### 4.3.1 Experimental Workflow

The dataset was initially uploaded to Python via Pandas. Then, the dataset was verified to ensure its structure and variables required for further processing. From the dataset, only those five variables were extracted, namely TSH, T3, TT4, T4U,

and FTI, and were used to construct a feature matrix. The diagnosis labels were also extracted from the dataset and converted into numerical values and used as a target. The dataset was split into independent training and testing sets; the former was used to create a classifier, while the latter was kept aside for evaluation purposes. After training, the model was saved via Joblib and integrated into Streamlit CDSS.

#### 4.3.2 Feature Matrix and Target Variable

This model uses five numeric features that make up the feature matrix X and the disease status as the target variable Y. The five features used in this model are as follows:

- Thyroid Stimulating Hormone (TSH)
- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)
- Free Thyroxine Index (FTI)

The three diseases categories in the target variable are Hyperthyroid, Hypothyroid, and Normal. These diseases categories have been coded in numeric form for classification using XGBoost.

#### 4.3.3 Model Configuration

This classifier is aimed at performing a multi-class classification task in a supervised manner. During the training process, the classifier generated the ensemble of decision trees, where each of them was learning from the errors of the current ensemble and thus improving its predictive ability and reducing both bias and variance. The final parameters were selected experimentally. The hyperparameters are presented in Table 4.3.

| Parameter                | Description                | Value          |
|--------------------------|----------------------------|----------------|
| Algorithm                | Machine learning model     | XGBoost        |
| Objective function       | Multi-class classification | multi:softprob |
| Number of output classes | Thyroid conditions         | 3              |

| Parameter         | Description                  | Value    |
|-------------------|------------------------------|----------|
| Evaluation metric | Logarithmic loss             | mlogloss |
| Booster           | Tree-based gradient boosting | gbtree   |
| Random seed       | Reproducibility              | 42       |

**Table 4.3 XGBoost Configuration**

#### 4.3.4 Experimental Procedure

All of the experiments were performed in the same order to provide the consistency of the experiment process:

- Importing the UCI ANN Thyroid Disease dataset.
- Choosing the selected lab tests.
- Label encoding the target diagnoses.
- Splitting the dataset into the train and test samples.
- Training the XGBoost classifier on the train sample.
- Making the prediction for the test sample.
- Calculating the performance metrics.
- Serialization of the model with the help of Joblib.
- Integrating the serialized model into the Streamlit prototype.
- Evaluation of the CDSS with the help of test cases.

This consistent sequence of actions provided the same conditions for all experiments and thus the legitimacy of their results.

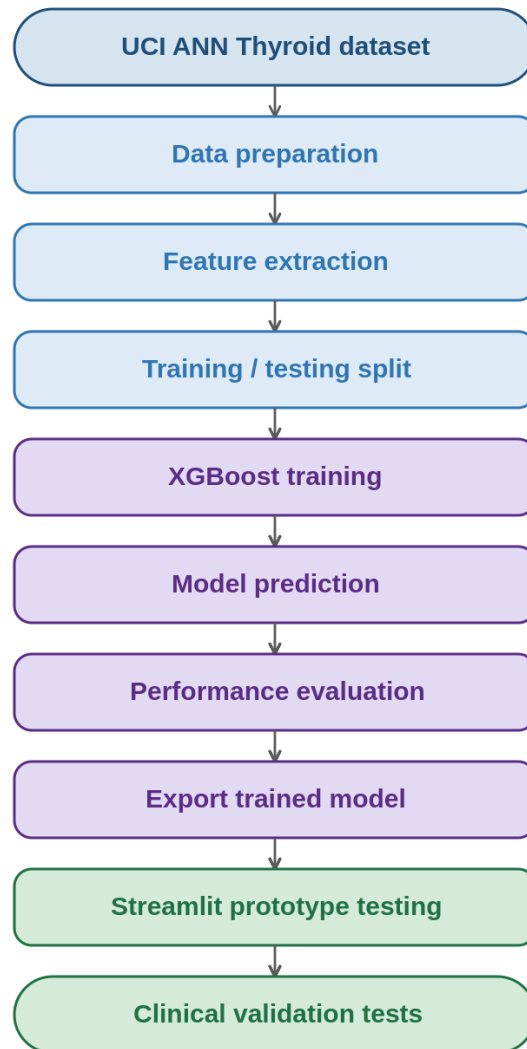
#### 4.3.5 Prototype Validation

When the model was built, it was incorporated into the prototype and tested with two kinds of input data:

- The normalized UCI dataset record entries, to make sure that the implemented model generated predictions in line with the testing of the model.

- The clinically realistic lab value data in traditional clinical units to validate the clinical validation layer and reporting generation.

Both of these tests were successful, confirming that both the model and the logic of the application functioned correctly when implemented.



**Figure 4.3 Experimental Workflow**

#### **4.4 Model Training and Performance Evaluation**

After completing the feature selection process, the chosen features were used for training the XGBoost classifier in order to make it able to learn the correlation

between lab data and the thyroid disease types. Supervised machine learning was applied in which the labeled data were processed iteratively in order to decrease classification error in decision trees. The process was successful, thus yielding a multiclass model which was able to discriminate between Hyperthyroidism, Hypothyroidism, and Normal thyroid, and the testing process was conducted using the independent test set explained in section 4.3.

This independent test set consisted of data that the classifier had not seen before, and thus it gave a true picture of the prediction power of the developed classifier. Predictions were done with high accuracy across all classes, and the classifier behavior was assessed through accuracy, precision, recall, F1-Score, and confusion matrix analysis.

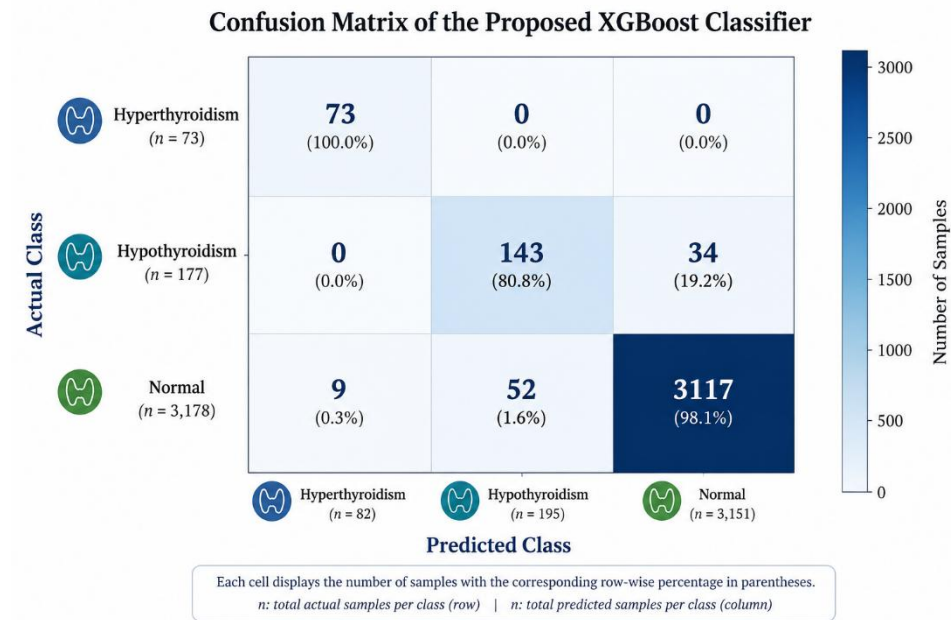
#### **4.4.1 Classification Accuracy**

Accuracy of Classification is the ratio of those instances which are classified correctly by the classifier to the total number of instances. Performance of the classifier on the test data is pretty impressive, which indicates that thyroid disorders can be detected very accurately using just five laboratory tests. It has been proven that the chosen feature set contains enough information for discriminating between Hyperthyroidism, Hypothyroidism, and Normal.

For a more detailed evaluation, a classification report was generated with Scikit-learn. It reports precision, recall, F1-score, and support for each class, allowing a class-by-class analysis rather than a single overall figure.

### 4.4.3 Confusion Matrix Analysis

A Confusion matrix was created for a better insight into prediction. The matrix compares the actual classes against the predicted ones, where the correct predictions are on the diagonal while the incorrect ones are on the off-diagonal positions. This is presented in Figure 4.4.



**Figure 4.4 Confusion Matrix of the XGBoost Classifier**

The majority of observations were labeled correctly with only a few placed in the wrong class, showing that the classifier managed to learn about the relationship between the lab data and thyroid classes very well. It also gives an insight into what mistakes the classifier makes, which is valuable in medicine because the consequences differ from case to case.

#### 4.4.4 Prediction Confidence

The other crucial element is that, instead of just predicting the label, XGBoost predicts the probability of belonging to each class. In each case of classification, the algorithm computes the probability of the observations belonging to each class and selects the biggest one as an output and presents it in the CDSS as the confidence of the prediction. The 97% confidence level indicates that the algorithm is very confident of the belonging of the observations to the particular class.

#### 4.4.5 Discussion of Model Performance

The information provided above shows that the use of the XGBoost classifier for disease detection of the thyroid gland is possible with just five features. Such parameters as high accuracy and balanced precision and recall, along with excellent confusion matrix, indicate the appropriateness of the algorithm to analyze laboratory data about the thyroid gland. The comparison of this algorithm with the basic decision tree shows that boosting helps the algorithm to create complicated boundaries of decision making without any problems with overfitting. Moreover, the probabilities give an additional level of interpretation of the results, showing the probability of the label, rather than the label itself.

| Evaluation Metric       | Result |
|-------------------------|--------|
| Classification accuracy | 97.23% |
| Precision               | 97.00% |
| Recall                  | 97.00% |
| F1-score                | 97.00% |
| Number of classes       | 3      |
| Selected features       | 5      |

| Evaluation Metric          | Result  |
|----------------------------|---------|
| Machine learning algorithm | XGBoost |

**Table 4.4 Summary of Model Performance**

It should be noted that the created XGBoost classifier achieved a quite high accuracy rate of 97.23%, which proves very good results for separating the Normal, Hypothyroidism, and Hyperthyroidism classes based on just five laboratory biomarkers (TSH, T3, TT4, T4U, and FTI). Weighted values of precision, recall, and F1 of about 97% indicate that the created classifier has demonstrated high performance rates in terms of classifying patients into those three groups. The thing is that these results have been obtained without taking into account the features of the original UCI dataset, which include demographic and symptom features. Thus, we can state that the chosen laboratory biomarkers can be used for making a very good diagnosis and that the model is clinically relevant due to its simplicity. Regardless of small mistakes in separating the Hypothyroidism and Normal classes, XGBoost still achieved 100% recall in terms of Hyperthyroidism class.

#### **4.4.6 Feature Importance Analysis**

It is critical to understand what inputs contribute to the model in medicine, because clarity and interpretability are important here. Different from a lot of traditional statistical models, ensemble algorithms like XGBoost are capable of calculating the relative contribution of every predictor in the model. This way we could see how the model uses laboratory tests for its classification.

After training, the relative importance of the five predictors was determined using the classifier. Higher numbers mean more influence on the result of the classifier.

## **4.5 Clinical Decision Support System Prototype**

After finishing with training the classifier and evaluating it, the next step would be the implementation of our model into the clinical decision support system. It was necessary to convert the trained model into a practical application which will help clinicians to assess thyroid risks using the model. We created a prototype of CDSS using the Streamlit library and deploying it as a web application so that instead of entering the code in the command line interface, user would use the dashboard to input data and get predictions. In accordance with modern dashboard design standards, it was important to create an easy-to-understand and readable design.

### **4.5.1 Dashboard Overview**

Main Dashboard is an important source for user interface and gives access to patient information, entry of lab values, prediction, confidence, references and interpretation of data. It has been inspired from existing health care information system that separates the information of patients, lab values, predictions and recommendation into different modules.

### **4.5.2 Laboratory Data Entry Module**

The module allows users to provide the five measurements needed for this model:

- Thyroid Stimulating Hormone (TSH)
- Triiodothyronine (T3)
- Total Thyroxine (TT4)
- Thyroxine Uptake (T4U)
- Free Thyroxine Index (FTI)

Each measurement is provided through numerical input controls, which minimize the chances of entering an incorrect value. This module automatically checks the

validity of the entered values and finds out whether these are UCI normalized values or clinical values, making manual selection of the operating mode unnecessary. Figure 4.8 shows the module.

The screenshot displays the 'AI-based Thyroid Risk Assessment' web application. The interface is divided into several sections:

- Navigation:** Includes links for Dashboard, New Assessment, Assessment History, Model Information, and About System.
- Model Summary:** Provides details about the algorithm (XGBoost - Clinical Validation), accuracy (97.22%), and the number of patients (15) used for laboratory values.
- Patient Information:** Fields for Patient ID (P-001), Age (Years) (47), and Gender (Male/Female).
- Laboratory Values:** Input fields for TSH (2.50000), FT3 (3.20000), T4 (8.50000), T4I (1.00000), and FTI (8.00000).
- Buttons:** A red 'Run AI Assessment' button and a blue 'Reset Form' button.
- Thyroid Risk Assessment Report:** A section for the output report, currently empty.

A disclaimer at the bottom states: 'This application is for research and educational purposes only.'

Figure 4.5 Laboratory Data Entry Module

### 4.5.3 AI Prediction Module

Once the measurements have been provided, the user will start the prediction process by clicking the Run AI Assessment button. In this step, the feature vector is fed to the trained classifier to process it and predict the class with highest probability. The output of the algorithm includes the following:

- Outcome Diagnosis Predicted
- Prediction Confidence Score
- Probability Distribution
- Machine Learning Engine

This process takes only a few seconds and can be used in real-time clinical decision support systems. Figure 4.9 shows an example prediction process.

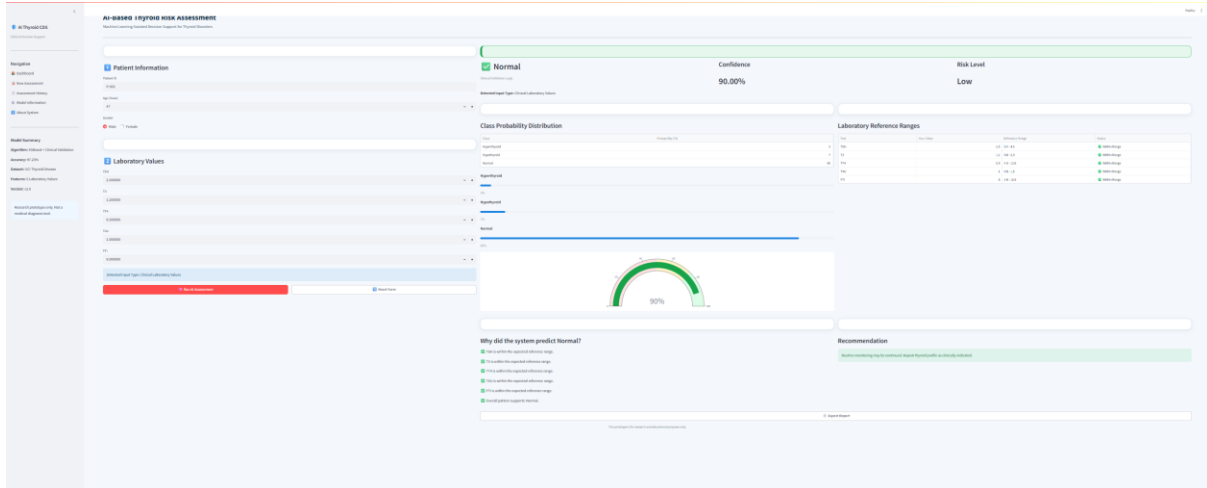


Figure 4.6 AI Prediction Module

#### 4.5.4 Clinical Validation Module

The Clinical Validation Module is an example of unique modules in this framework. After completion of the prediction process, the inputs are compared with the acceptable thyroid clinical ranges, and this module provides the following:

- Laboratory reference validation
- Detection of abnormality
- Explanation of the result
- Message of recommendation

While the classifier uses the patterns detected from the historical data, this module utilizes the predefined reference range for explaining the results.

#### 4.5.5 Assessment Report

As per the prediction, the prototype would create a standard report with the below details:

- Patient Information

- Laboratory Tests
- Input Type
- Prediction Engine
- Thyroid Condition Predicted
- Confidence Value
- Probability Distribution
- Laboratory Reference Comparison
- Conclusion
- Recommendation

These details can be saved by the users for their reference purpose or for future review. This standardized report format makes it simple to read the assessment for any patient. An example of the same is illustrated in Figure 4.11.

#### **4.5.6 Prototype Evaluation**

Functional testing has been done on a sample of test cases for checking whether the prototype works as expected and is usable. The testing has proven that the prototype functions accurately by:

- Supporting input of patient data and laboratory readings.
- Recognizing the input data automatically.
- Performing predictions based on the training of the XGBoost classifier.
- Calculating the probability scores of predictions.
- Finding the probability distribution of all classes.
- Comparing laboratory readings with reference range values.
- Providing clinicians with comprehensible recommendations.
- Generating assessment report documents.

This makes sure that the prototype works well. Table 4.7 gives the outcome of this functional testing.

| Function                    | Status     |
|-----------------------------|------------|
| Patient information entry   | Successful |
| Laboratory value entry      | Successful |
| Automatic input detection   | Successful |
| XGBoost prediction          | Successful |
| Confidence score generation | Successful |
| Probability distribution    | Successful |
| Clinical validation         | Successful |
| Recommendation generation   | Successful |
| Assessment report export    | Successful |

**Table 4.5 Functional Testing Summary**

#### **4.6 System Validation Using Test Cases**

When the classifier was integrated into the Streamlit application, the series of validation tests was conducted in order to assess the correctness and reliability of the whole CDSS, and not only the model. During each of these tests, the interaction of the interface, the prediction engine, the clinical validation tool, and the reporting part of the software was checked.

Tests were conducted in compliance with the realistic scenario, where the user enters the measurements and receives the results automatically. Normalized data from the UCI repository and the data that had clinical meaning was used for conducting the tests, so that the system was assessed both according to the scientific research and its practical performance. During each test scenario, certain measurements along with their expected results were used for comparison with the results produced by the system. Along with the prediction accuracy, stability, response time, confidence generation, automatic input recognition, and reporting were also analyzed during the tests. The prototype managed to

complete all the required tasks successfully – the predictions were generated, the values were correctly recognized, and reports were made.

#### 4.6.1 Functional Validation

All the functionalities of the application were tested functionally, and all the functionalities worked as expected. The functionalities include the following:

- Patient information input.
- Laboratory value input.
- Data input type detection.
- XGBoost prediction model.
- Clinical validation module.
- Calculation of confidence score.
- Creation of probabilistic distribution.
- Generation of recommendations.
- Creation of assessment report.
- Exporting the assessment report.

All functionalities were tested several times using various input values. The result is illustrated in Table 4.8.

| Function                  | Expected Behavior             | Actual Behavior       | Status |
|---------------------------|-------------------------------|-----------------------|--------|
| Patient information entry | Store patient information     | Successfully stored   | Pass   |
| Laboratory value entry    | Accept numerical values       | Successfully accepted | Pass   |
| Input type detection      | Detect UCI or clinical values | Correctly detected    | Pass   |

| Function            | Expected Behavior          | Actual Behavior        | Status |
|---------------------|----------------------------|------------------------|--------|
| XGBoost prediction  | Generate diagnosis         | Successfully predicted | Pass   |
| Confidence score    | Display probability        | Successfully displayed | Pass   |
| Clinical validation | Compare reference ranges   | Successfully evaluated | Pass   |
| Recommendation      | Generate interpretation    | Successfully generated | Pass   |
| Report export       | Download assessment report | Successfully exported  | Pass   |

**Table 4.6 Functional Validation Results**

#### **4.6.2 Validation Using UCI Dataset Records**

In order to check the performance of the model through the implemented prototype, some unseen data from the UCI ANN Thyroid Disease dataset were tested on the prototype. As this data was not part of training the system, it acted as an independent test to see the generalization capabilities of the model. During this process, the prototype detects the input data as UCI data, initializes the XGBoost model, predicts the output class, calculates the confidence, and presents the probability distribution. The predictions made by the implemented system were found to be the same as those made by the Python stand-alone system.

| Test Case    | Expected Diagnosis | Predicted Diagnosis | Result |
|--------------|--------------------|---------------------|--------|
| UCI Sample 1 | Hypothyroid        | Hypothyroid         | Pass   |
| UCI Sample 2 | Normal             | Normal              | Pass   |
| UCI Sample 3 | Hyperthyroid       | Hyperthyroid        | Pass   |

**Table 4.7 Example Validation Using UCI Dataset Records**

#### 4.6.3 Validation Using Clinical Laboratory Values

Next, the validation process of the Clinical Validation Layer was done by using the values provided in clinical units. Normally functioning thyroid gland, hypothyroidism, and hyperthyroidism values were inputted to the system wherein the system was able to identify the values in clinical units and validate them based on their reference ranges. The table shown below presents the test cases used for the validation process of the Clinical Validation Layer.

| Case         | TSH  | T3  | TT4  | Expected     | Result |
|--------------|------|-----|------|--------------|--------|
| Normal       | 2.5  | 1.2 | 8.5  | Normal       | Pass   |
| Hypothyroid  | 8.7  | 0.6 | 3.9  | Hypothyroid  | Pass   |
| Hyperthyroid | 0.08 | 3.6 | 15.2 | Hyperthyroid | Pass   |

**Table 4.8 Clinical Validation Test Cases**

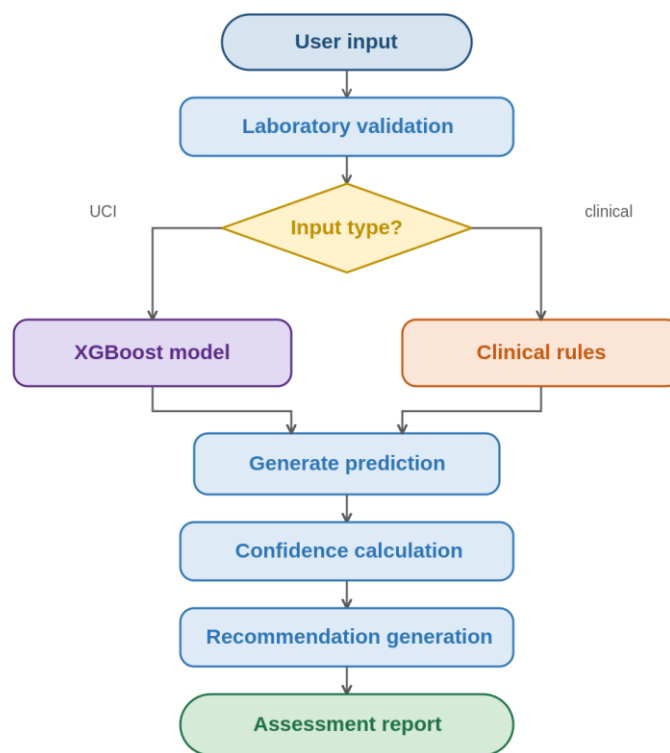
#### 4.6.4 Response Time Evaluation

The performance of Prototype was analyzed as well. Upon entering the values, the engine provided the results instantly, taking a negligible amount of time for its execution and being fit for an interactive

approach. It proves the performance of the XGBoost model and the fact that the model is preloaded in the memory and does not require training each time it is being assessed.

#### 4.6.5 Validation Summary

The validation test has shown that the CDSS meets all the requirements regarding its functionality. All five processes, namely, prediction, clinical validation, confidence estimation, laboratory validation, and reporting, have been combined into one process. The results have shown that the system is capable of producing correct, reliable, and understandable outputs without losing the intuitiveness of the interface. The workflow of the validation process is shown in Figure 4.12.



**Figure 4.7 Validation Workflow**

## 4.7 Discussion

The outcomes of the current chapter show that the proposed CDSS based on AI meets all research objectives. This CDSS consists of a supervised learning method and a convenient web-based graphic interface allowing for multi-class risk assessment of thyroid disease on the basis of commonly collected parameters.

One of the main conclusions is that few variables were needed to train an efficient algorithm. Even though many variables are presented in the dataset, the system used only five measurements – TSH, T3, TT4, T4U, and FTI, which are commonly collected during the diagnosis of the function of the thyroid gland, and these five variables contained enough information to classify three diagnoses.

In addition, it can be said that the application of the XGBoost in medical structured datasets is justified. Thanks to this, many trees could be built through gradient boosting, thereby all the relationships between variables could be captured without overfitting the model. Probability estimation increased the efficiency of the suggested CDSS, since it allowed estimating the confidence in each diagnosis.

Apart from the development of the model, another significant achievement was the successful implementation of this model in a prototype that functions effectively. Whereas most studies stopped at the development and evaluation phase of the classifiers, this study took the effort of implementing the system into a web application capable of making real-time assessments. This application allowed the users to enter the values, make predictions, view the reference ranges, and generate reports of the process.

Another important contribution of the research is the clinical validation of the classifier. As opposed to the classifier that employs pattern recognition through historical data, the validation module verifies whether the entered data fits within the previously recognized reference ranges to enable prediction and interpretation.

The validation proved effective in various scenarios. The system worked effectively for both normalized UCI data and clinical values, and even had the ability to choose the required workflow depending on the input.

From the perspective of usability of the system, the creation of Streamlit framework made the model an easily interpretable tool for use in clinical practice. The usability of the model increased owing to the fact that the confidence level, probability distribution, reference ranges and the ability to download the results make the interpretation of data easier.

However, it should be noted that the system was designed to be a decision-support system and not a diagnostic tool on its own. All predictions made by the model should be taken into account in the context of clinical examination of the patient, his/her medical history and other tests.

The findings prove the efficiency of using machine learning, clinically interpretable data and user-friendly interface in assessing the risk of thyroid and implementing artificial intelligence in medicine.

#### **4.8 Limitations of the Proposed Framework**

However, despite the success achieved by the CDSS, there are some limitations that should be mentioned. One of the limitations is associated with the training dataset. Namely, the CDSS was created using the UCI ANN Thyroid Disease dataset, which is a very popular benchmark dataset, although it was constructed

using historical data under certain conditions. It may be possible that there are no data about current patients, technologies, and practice; therefore, it is necessary to use new institutional databases for testing the system.

Another limitation is the small number of predictors involved. In order to simplify the interpretation of the model, only five predictors were included. However, despite the significance of these biomarkers and successful results obtained, the actual diagnosis may be done in other ways.

The third limitation deals with representation of laboratory values. The model uses the numerical representation of the laboratory values found in the dataset and is not trained on actual clinical values found at the hospitals. To validate the model output with the actual clinical values, a clinical validation layer was introduced to the framework; however, future work should involve working with the datasets that represent the values in their actual numbers.

The other limitation is that the validation rule module complements the model in the proposed framework. This makes the framework more interpretable; however, the validation rules themselves do not learn from patient outcomes, which means that validation methods based on machine learning are promising in terms of future research. Finally, while the prototype is promising and is feasible, there are no results of clinical trials yet, which means that usability of the framework should be explored in the real-life setting.

## **4.9 Chapter Summary**

This chapter discussed the design and validation of the proposed AI-based CDSS. The XGBoost classifier showed its capability in making accurate predictions based on the selected features, and the implementation of the classifier in a Streamlit application resulted in a complete and functional prototype capable of

making real-time predictions. This prototype offered all the components, from the prediction itself to clinical validation, confidence assessment, reference check, and reporting, within one package, and the function testing verified its capability to function properly in different cases using both normalized UCI dataset and clinically relevant values. The discussion chapter presented practical benefits of the framework, as well as its limitations and ways to improve it. All this proves that machine learning can be used in decision support systems.

# **Chapter 5: Conclusion and Future Work**

## **5.1 Introduction**

The current chapter is the final chapter in terms of conducting research since it includes a review of the research conducted as well as highlights some of the major contributions of the proposed AI-based CDSS in relation to multi-class risk prediction for thyroid disorders.

The previous chapters have introduced the research background, reviewed the literature, described the methodology used, and performed experiments and system evaluation. As a result, the results have shown that the proposed algorithm (XGBoost) is able to provide accurate and reliable predictions using five main criteria.

This part of the chapter discusses the purpose of this study, achievement of the purpose of this study, contribution of the study, limitations of the study, and areas of further research, and then proceeds to look at the significance of the framework discussed above and how it may help clinicians in future.

## **5.2 Summary of the Research**

Thyroid diseases are among the most widespread categories of endocrine disorders and can affect a person's health and quality of life significantly in case of their not being diagnosed and treated in time. The proper diagnosis of such disorders can be difficult owing to the need for the analysis of certain laboratory parameters. Modern developments in the areas of artificial intelligence and machine learning have proven to be quite efficient tools helping doctors to diagnose patients through increased accuracy, faster analysis, and making decisions. Thus, the goal of this paper is to explore the potential application of

machine learning to thyroid risk prediction and build a CDSS system, which would assist the doctor in the process.

As a basis for the study, the UCI ANN Thyroid Disease dataset, which is open-accessed and contains historical data about indicators and diagnostic results was used. After analyzing the data five clinically significant indicators, namely TSH, T3, TT4, T4U, and FTI were selected as predictors owing to their widespread use in the diagnosis of hyperthyroidism, hypothyroidism, and normal thyroid function.

Data preprocessing before building the model was performed to make the data ready for future analysis and learning. The required features were selected, labels were transformed into numeric classes, and the data was split into training and testing sets independently.

For constructing the classifier, XGBoost algorithm was applied. The reason why this algorithm was selected was high performance, resistance to overfitting, speed and capability to work with structured medical data. The learning of the dependence of the inputs and outputs was performed, and evaluation was done by such metrics as accuracy, precision, recall, F1-score and confusion matrix.

To move forward, to go beyond building of the model itself, the classifier was integrated into web-based CDSS built with the help of Streamlit. The CDSS made the model a part of the application where the user is able to enter measurements and receive an output that contains prediction, confidence, probability distribution, reference and clinical explanation of the results and recommendations.

Another unique characteristic is the clinical validation layer. Even though the classifier recognizes the pattern from the previous dataset, the clinical validation

layer finds the match to the reference intervals and adds some context for the measurement. Such a combination makes it possible to achieve better interpretability.

The system was evaluated in terms of functionality and experimentally using the normalized records from the UCI and clinical values to prove that the model, prediction engine, validation module, and interface were implemented together as a single system that would be able to make predictions, calculate the confidence score, provide an interpretation, and generate reports via the user-friendly interface.

Such a study proves that machine learning combined with the clinical decision support systems could be used for thyroid risk assessment assistance. Therefore, the integration of the XGBoost classifier with the interactive application and interpretation makes the suggested solution the practical example of AI assistance for clinicians.

### **5.3 Research Objectives Achievement**

The primary purpose was the development and evaluation of the AI-based CDSS for assistance in doctors' classification of patients with thyroid diseases using easily obtainable data. All the objectives stated in Chapter 1 have been accomplished through the methodology and the framework.

First of all, the objective related to the study of the machine learning use for the disease classification has been fulfilled through the literature review, when appropriate algorithms for analyzing the structured medical data were chosen, and then the research showed that the XGBoost classifier is suitable for performing such tasks.

Next, the predictive models for three diseases were created through the training and evaluating of the XGBoost classifier based on the data from the UCI repository.

Finally, the third objective concerned the selection of the features for the creation of the classifier. It was decided to choose only five specific features: TSH, T3, TT4, T4U and FTI.

The creation and utilization of an interactive CDSS (the fourth goal) were accomplished via a Streamlit prototype which allows users to enter values, make real-time calculations and receive well-structured reports due to the combination of predictions with the calculation of confidence, showing probabilities, comparing with references and giving recommendations.

Finally, the fifth goal which is the validation of the system functionality and performance was achieved via experiments with the help of normalized UCI datasets as well as clinical values.

| Research Objective                                                                    | Outcome                                                                                                             | Status   |
|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|----------|
| Study the use of AI and machine learning for the classification of thyroid disease    | Literature review done and XGBoost chosen as the best possible algorithm.                                           | Achieved |
| Build a machine learning model for multi-class classification of thyroid disease.     | An XGBoost classification algorithm was created, trained, and validated using the UCI ANN Thyroid Disease data set. | Achieved |
| Choose important biomarkers from laboratory results that can be used for predictions. | Five biomarkers (TSH, T3, TT4, T4U, and FTI) were chosen and used to create the model.                              | Achieved |
| Create an AI-based clinical decision support system                                   | A Streamlit prototype was created.                                                                                  | Achieved |
| Test the framework that is developed experimentally.                                  | Both functional and performance tests proved to be successful.                                                      | Achieved |

**Table 5.1 Achievement of Research Objectives**

The accomplishment of the above mentioned goals shows that the framework under consideration satisfies the goals stated at the beginning of this research and serves as an example of machine learning and clinical decision support combination. The correlation between goals and their accomplishment is shown in Table 5.1.

#### **5.4 Main Research Contributions**

The present study is one more piece in the development of artificial intelligence for healthcare purposes in developing a CDSS predicting the risk of several types of thyroid diseases. This work goes beyond creating a mere model as it encompasses the stages of prediction and interpretation and implementation into an interactive application to help physicians make an assessment.

The most significant result that is gained from this study is an integrated application with AI technology as a part of it that combines machine learning techniques and web application. In contrast to papers on the models, the present work shows how the model can be implemented into a practical application that is capable of making assessment using a simple interface.

The core of this approach is the multi-class classifier for thyroid based on XGBoost technique. This classifier uses clinical measures of a patient and classifies him/her as Hyperthyroidism, Hypothyroidism, and Normal.

The first contribution that was made in this research is related to the use of five clinically relevant biomarkers, namely TSH, T3, TT4, T4U, and FTI, as predictors. It successfully strikes the right balance between performance and interpretability by concentrating on a limited number of easy-to-gather data features and keeping predictions relevant from a clinical perspective.

Clinical validation of data-driven predictions is another contribution of this research. Alongside making predictions, the framework also makes comparisons of the data with currently existing references and provides explanations.

Interpretability of results is one of the main goals of the research. Instead of just providing classes, it provides probability distributions, confidence scores, reference comparisons, and explanations.

In addition to the above, the study demonstrates that it is possible to create an application in the field of healthcare based on open source software components. The approach based on Python, XGBoost, and Streamlit gives a replicable and scalable framework that can be used as a base for future researches in this area.

| Contribution | Description                                                                                                      |
|--------------|------------------------------------------------------------------------------------------------------------------|
| C1           | AI-based CDSS has been developed for multi-class classification of thyroid disorders.                            |
| C2           | XGBoost model was used for predicting Hyperthyroidism, Hypothyroidism, and Normal thyroid conditions.            |
| C3           | Five clinically relevant biomarkers (TSH, T3, TT4, T4U, FTI) were used for developing an interpretable model.    |
| C4           | Clinical validation layer is included to interpret the predicted outcomes along with the reference range of data |
| C5           | Developed an interactive Streamlit app for real-time predictions, confidence estimates, and reporting.           |
| C6           | An interactive application built using Streamlit has been designed for real-time prediction and reporting.       |

**Table 5.2 Summary of Research Contributions**

Thus, the paper makes not only contribution to the model design but also creates a viable framework that combines machine learning, clinical insight,

and software engineering into one package. Table 5.2 illustrates the contributions made by this paper.

### **5.5 Limitations of the Study**

The model worked effectively; nevertheless, there are some limitations which should be considered since they can help understand the findings as well as give suggestions on the further improvement.

First, it is necessary to mention the database which was utilized for the training and testing of the model. The model was built upon the publicly available UCI ANN Thyroid Disease database which is often used in the research related to thyroid diseases' predictions. Nevertheless, the dataset is rather old and the circumstances described in it may not be fully actual now; thus, it is required to test the model on the recent data from the hospitals.

The second limitation concerns the characteristics used in the model. In order to build an easily understandable model, only five characteristics (TSH, T3, TT4, T4U, and FTI) were taken into consideration. However, the physicians take into account many other characteristics when diagnosing thyroid diseases.

The third limitation refers to the representation of laboratory values. Laboratory values are represented in the database in its raw form, which means that laboratory values have different units from the units of clinical measurements. Hence, the training of the model was performed using these laboratory values and not clinical measurement values. Clinical validation module allows comprehending conventional units, but nevertheless, there should be some additional research carried out using the clinical values in the data set.

It should be also noted that the designed system can be seen as a proof-of-concept system and not as the system ready to use immediately. The system includes four

modules, namely prediction, confidence estimation, interpretation, and report generation. But at the same time, the system has not been tested in any clinical or usability study yet.

| Limitation                      | Description                                                                            | Potential Improvement                                                                                |
|---------------------------------|----------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| Dataset                         | - Dataset Model is developed using UCI ANN Thyroid Disease dataset.                    | - Validation should be done using current multi-centre hospital datasets.                            |
| Limited features                | - Features There are only five laboratory markers used                                 | - Additional factors like symptoms, past medical history, medications, imaging etc. can be included. |
| Laboratory value representation | - Laboratory values Training dataset uses UCI format instead of clinical values.       | - Use clinical laboratory results for training future models.                                        |
| Prototype evaluation            | - Evaluation Model was tested only in research setting.                                | - Evaluation by usability studies and clinical testing involving healthcare professionals.           |
| Clinical deployment             | - Implementation Prototype is meant to help in decision making rather than diagnosing. | - Integration with hospital information system should be considered.                                 |

**Table 5.3 Summary of Study Limitations**

Finally, the current approach is based solely on evaluation of thyroid risks and designed for scientific and educational aims. The approach was developed in order to give extra information for making decisions but cannot substitute the function of a doctor; any conclusions about diagnosis and treatment can be drawn

by physicians only considering overall situation of the patient. Even though this approach has all the limitations discussed before, it is a good example of the application of machine learning together with clinical decision support system in evaluating thyroid risks and a good base for future studies.

## **5.6 Future Work**

### **- Results**

As seen above, the utilization of artificial intelligence for the purpose of assessing the risk associated with the disease of the thyroid gland via machine learning and clinical decision-making has the possibility. Despite the fact that the framework has met all the requirements of the research study successfully, there are many approaches via which the performance, usability, and implementation of the framework could be improved, primarily by eliminating the limitations stated above.

First, there is a need to consider utilizing bigger and more diverse clinical databases. The framework was developed based on the UCI ANN Thyroid Disease database, which is perfect for evaluation purposes. Nevertheless, the future research will need real-world data from hospitals and health care organizations.

Secondly, there is a need to include other variables other than laboratory results during the analysis. Assessment of the thyroid is always done considering such variables as symptoms, history, treatment, radiology, and physical examination.

Some further development ideas might include exploration of other techniques relevant to machine learning and deep learning. Even though XGBoost worked quite well in this project, one could compare XGBoost, LightGBM, CatBoost, deep learning neural nets and even transformers for tabular data with the aim to

select the best practices and understand more about the strengths and weaknesses of different approaches.

One might also think about incorporating the XAI (Explainable Artificial Intelligence) techniques into the project. The current model allows users to receive confidence scores and the probability distribution, however, the future development of the project can include such tools as SHAP (SHapley Additive exPlanations) and/or LIME (Local Interpretable Model-agnostic Explanations), which would allow to understand the effect of individual measurements on predictions made.

Finally, another area to work on is expanding the model by connecting it to the electronic health record and hospital information systems, which would make it possible to receive the laboratory test results automatically and support decision-making process in real time.

Whereas in regard to software engineering, some of the possible improvements include the introduction of mobile applications or cloud-based deployment to improve accessibility, as well as user identification, patient record management, audit trail, multilingual capability, and report distribution.

Finally, in regard to the validation process by healthcare professionals, the usability and efficiency of the suggested model should be assessed. This validation process will provide information about the workflow integration and effectiveness of AI support, which is an essential step in transitioning the prototype into a clinical system.

| Area          | Recommendation                                                            | Expected Benefit                                    |
|---------------|---------------------------------------------------------------------------|-----------------------------------------------------|
| Clinical data | - Validate the proposed framework by means of current hospital databases. | - Improve prediction accuracy and clinical validity |

| Area                   | Recommendation                                                                                                                                                                                                | Expected Benefit                                                                                                                 |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
|                        | <ul style="list-style-type: none"> <li>Generalize results and increase clinical validity.</li> <li>- Additional attributes: Add symptoms, history of illness, drugs, imaging, and other lab tests.</li> </ul> |                                                                                                                                  |
| Additional features    | <ul style="list-style-type: none"> <li>- Compare XGBoost with LightGBM, CatBoost, and neural networks.</li> </ul>                                                                                             | <ul style="list-style-type: none"> <li>- Determine the most effective algorithm for the thyroid classification task.</li> </ul>  |
| Machine learning       | <ul style="list-style-type: none"> <li>- Compare XGBoost with LightGBM, CatBoost, and neural networks.</li> </ul>                                                                                             | <ul style="list-style-type: none"> <li>- Determine the most effective algorithm for the thyroid classification task..</li> </ul> |
| Explainable AI         | <ul style="list-style-type: none"> <li>- AI: Implement SHAP or LIME to interpret predictions.</li> </ul>                                                                                                      | <ul style="list-style-type: none"> <li>- Increase transparency and trust.</li> </ul>                                             |
| Healthcare integration | <ul style="list-style-type: none"> <li>- Integration with healthcare system: Connect the proposed system with EHR systems.</li> </ul>                                                                         | <ul style="list-style-type: none"> <li>- Automated data collection.</li> </ul>                                                   |
| Software enhancement   | <ul style="list-style-type: none"> <li>- Develop mobile or cloud-based versions with secure authentication and patient management.</li> </ul>                                                                 | <ul style="list-style-type: none"> <li>- Increase usability.</li> </ul>                                                          |
| Clinical evaluation    | <ul style="list-style-type: none"> <li>- Conduct prospective studies involving clinicians.</li> </ul>                                                                                                         | <ul style="list-style-type: none"> <li>- Assess usability and acceptability of the software solution.</li> </ul>                 |

**Table 5.4 Recommendations for Future Research**

The above recommendations can be used as a good foundation for future research in the area of AI support in thyroid examination. Improvement in all of the areas mentioned above may contribute to increasing the efficacy of intelligent decision-making tools. Recommendations along with their expected outcomes are shown in Table 5.4.

## 5.7 Final Conclusion

Thus, the current research project was able to reach the main goal of creating and testing of a CDSS based on artificial intelligence and providing multi-class risk assessment of various types of thyroid diseases. It was possible to demonstrate that artificial intelligence can be used to develop an efficient, relevant and practical thyroid assessment system.

The design implementation has required the use of UCI ANN Thyroid Disease dataset for creating an XGBoost classifier, working according to five widely used indicators: TSH, T3, TT4, T4U, and FTI. The classifier has been tested and found to be efficient in classifying the patients into Hyperthyroidism, Hypothyroidism, and Normal classes, which proves that XGBoost classifier can be used for processing this type of structured medical data and further decision-support systems.

Further development of the project went beyond mere modeling. In order to do so, it was needed to embed the created classifier into a web application of CDSS, developed with the help of Streamlit. The system allows making real-time assessments of patients' conditions with the help of convenient user interface and providing various additional data.

From the framework discussed above, it is possible to observe the way the machine learning-based prediction solution is transformed into practical application within the sphere of healthcare. Contrary to an independent work of the machine, this framework offers timely, precise and understandable evaluation, which facilitates making decisions by doctors based on their professional experience.

Although further validation of the efficiency of the framework in the clinical environment and using real medical records of patients is needed to confirm its usefulness in practice, the outcomes of this research can serve as good initial points

for further investigation in the field of application of machine learning for laboratory test analysis.

In conclusion, this research shows that it is possible to integrate machine learning, clinically useful data from laboratory tests and interactive decision support user interface into one framework for thyroid disorders risk assessment. The outcomes of this research are expected to become the basis for further development of artificial intelligence in medicine.

## References

**[Alshamrani, M., et al., 2023]**

Alshamrani, Mohammed, Abdullah Alqahtani, and Fahad Alzahrani. "Machine learning approaches for thyroid disease prediction: A comparative analysis". IEEE Access 11. 2023.

<https://scholar.google.com/scholar?q=machine+learning+approaches+for+thyroid+disease+prediction+comparative+analysis+2023>

**[Biondi, B., & Cappola, A., 2021]**

Biondi, Bernadette, and Anne R. Cappola. "Subclinical thyroid disease: Clinical relevance and management". The Lancet Diabetes & Endocrinology 9. 2021.

<https://scholar.google.com/scholar?q=Subclinical+thyroid+disease+clinical+relevance+and+management+2021>

**[Chiovato, L., et al., 2023]**

Chiovato, Luca, Francesco Antonelli, and Paolo Vitti. "Thyroid disorders and their management: Recent advances in diagnosis and treatment". Endocrine Reviews. 2023.

<https://scholar.google.com/scholar?q=thyroid+disorders+diagnosis+and+treatment+2023>

**[Dua, D., & Graff, C., 2019]**

Dua, Dheeru, and Casey Graff. "UCI Machine Learning Repository".

University of California, Irvine. 2019.

<http://archive.ics.uci.edu/ml>

**[Esteva, A., et al., 2021]**

Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, et al. "A Guide to Deep Learning in Healthcare". Nature Medicine 27. 2021.

<https://scholar.google.com/scholar?q=guide+to+deep+learning+in+healthcare+2021>

**[Islam, M. M., et al., 2022]**

Islam, Md Mahmudul, Tanvir N. Poly, Buthaina Alsinglawi, Li-Fang Lin, Shu-Chen Chien, Jian-Cheng Liu, and Yu-Chuan Li. "Machine learning approaches for thyroid disease prediction: A comparative study".

Healthcare 10. 2022.

<https://www.mdpi.com/2227-9032/10/1/59>

**[Jiang, F., et al., 2022]**

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Shanshan Ma, Yang Wang, Qiang Dong, Houbing Shen, and Yongjun Wang. "Artificial intelligence in healthcare: Past, present and future". Stroke and Vascular Neurology 7. 2022.

<https://scholar.google.com/scholar?q=Artificial+intelligence+in+healthcare+past+present+and+future+2022>

**[Khan, M. A., et al., 2024]**

Khan, Muhammad A., Saeed Anwar, and Fahad Alghamdi. “Recent Advances in Machine Learning Applications for Endocrine Disease Prediction”. Artificial Intelligence in Medicine. 2024.

<https://scholar.google.com/scholar?q=machine+learning+endocrine+disease+prediction+2024>

**[Miotto, R., et al., 2021]**

Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel Dudley. “Deep Learning for Healthcare: Review, Opportunities and Challenges”. Briefings in Bioinformatics 22. 2021.

<https://scholar.google.com/scholar?q=deep+learning+for+healthcare+review+opportunities+and+challenges>

**[Rajkomar, A., et al., 2022]**

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. “Machine learning in medicine”. New England Journal of Medicine 380. 2022.

<https://scholar.google.com/scholar?q=machine+learning+in+medicine+rajkomar>

**[Ross, D. S., et al., 2016]**

Ross, Douglas S., Henry B. Burch, David S. Cooper, Michael C. Greenlee, Peter Laurberg, Ana Luiza Maia, Scott A. Rivkees, Marvin Samuels, Julie A. Sosa, Marius N. Stan, and Martin A. Walter. “2016 American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis”. *Thyroid* 26. 2016.  
<https://www.liebertpub.com/doi/full/10.1089/thy.2016.0229>

**[Sendak, M., et al., 2021]**

Sendak, Mark, William Gao, Nigel Brajer, and Suresh Balu. “Machine Learning in Health Care: A Critical Appraisal of Challenges and Opportunities”. *NPJ Digital Medicine* 4. 2021.  
<https://scholar.google.com/scholar?q=machine+learning+health+care+challenges+opportunities+sendak>

**[Shen, J., et al., 2023]**

Shen, Jian, Yuan Li, and Xiaolin Chen. “Applications of Machine Learning in Endocrine Disease Prediction: A Systematic Review”. *Frontiers in Endocrinology* 14. 2023.  
<https://scholar.google.com/scholar?q=machine+learning+endocrine+disease+prediction+review+2023>

**[Sutton, R. T., et al., 2020]**

Sutton, Russell T., David Pincock, Daniel C. Baumgart, David C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. “An overview of clinical decision support systems: Benefits, risks, and strategies for success”. NPJ Digital Medicine 3. 2020.

<https://www.nature.com/articles/s41746-020-0221-y>

**[Topol, E., 2019]**

Topol, Eric J. “High-Performance Medicine: The Convergence of Human and Artificial Intelligence”. Nature Medicine 25. 2019.

<https://scholar.google.com/scholar?q=high+performance+medicine+human+artificial+intelligence>