Hyperthyroid and Hypothyroid Prediction using Machine Learning Techniques

by

Qiyuan Zeng

A dissertation submitted for the degree of

Master of Science
of
University College London

Supervisor: **Lina Dahye Song**
Co-supervisor: **Qinquan Cui**

Department of Physics and Astronomy
MSc. Scientific and Data Intensive Computing

September 2022

I, Qiyuan Zeng, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this dissertation.

The code for the implementation of this project are in the GitHub repository: https://github.com/zengxueyuan/PHAS0077_QiyuanZeng_2022.git

Abstract

The thyroid gland produces thyroid hormone, which regulates the body's metabolism including controlling body temperature, managing blood pressure, and regulating heart rate. Hypothyroid and hyperthyroid are the two main thyroid diseases caused by an excess or shortage of thyroid hormones, respectively. In this study, we adopt the thyroid gland dataset from UCI databases. First, a new comprehensive dataset is obtained through data integration and cleaning. Data analysis is then performed to determine the distribution of some features in the new data. There are ten machine learning algorithms in this research used to train the prediction models, which are XGBoost, AdaBoost, Gradient Boost, CatBoost, Random Forest, Decision Tree, Extra Trees, Bagging classifier, MLP classifier, and KNN. The results show the excellence of accuracy, for example, the F1-score of the random forest algorithm can reach 98.87% which is the best score in all the algorithms. In addition, other algorithms also can get above 98% accuracy. We have obtained high-quality results from this study, so it might be used as a reference by hospital decision-makers and researchers in the future as a new technique.

Keywords: Hyperthyroid, Hypothyroid, Machine Learning, Classification, Thyroid Disease

Acknowledgement

First and foremost, I am deeply grateful for the guidance I received from my project supervisor, Prof. Lina Dahye Song, and co-supervisor, Dr. Qinquan Cui. During weekly meetings and via emails, they provided me with constructive guidance and explanations. My knowledge of data analysis has been enhanced by their resources and detailed interpretations. I was able to gain confidence to achieve such a difficult task due to their constructive instruction which aided me in having a clear thoughts on this project.

Especially, my sincere gratitude is also extended to University College London's faculties for their perseverance and support throughout this challenging year. Greetings to all! I hope all is well with you!

Contents

Chapter 1

Introduction

1.1 Background

The thyroid gland is located in the lower part of the neck and its shape resembles a butterfly. Additionally, it contributes to the secretion of thyroid hormones, which regulate metabolism and protein synthesis and keep the body moist. There are many ways thyroid hormones can control metabolism in the body, including controlling body temperature, managing blood pressure, and regulating heart rate [1]. Therefore, if the thyroid gland does not function properly, the body's weight may fluctuate, it may even make the heart palpitate, and it may affect Fertility, motor coordination, and temperature sensitivity. Thyroid disease is one of the most prevalent endocrine disorders, second only to diabetes, and the incidence of thyroid disease is increasing [2]. In the thyroid gland, iodine is the main component to create thyroid hormone. In some regions, dietary iodine deficiency may lead to goiter or active thyroid nodules, with prevalence rates as high as 15% [3]. Furthermore, severe hyperthyroidism and end-stage hypothyroidism may result in death [4]. With such a high prevalence and serious influence of thyroid diseases, it is crucial to improve diagnosis accuracy as well as understand the risk factors for thyroid diseases.

A thyroid gland produces thyroid hormones which mainly are T4 (levothyroxine) and T3 (triodothyronine). In addition to regulating body temperature, these hormones facilitate protein synthesis and energy production [1]. Two of the most common diseases affecting the thyroid gland are hypothyroidism and hyperthyroidism and they are most frequently induced by an autoimmune process [2]. A deficiency of thyroid hormone causes hypothyroidism, and an increase in thyroid hormone levels causes hyperthyroidism [5]. In medical terms, hypo means deficient or less. The two primary causes of hypothyroidism are inflammation and thyroid gland injury. Symptoms of this condition include weight gain, neck swelling, intestinal problems, drowsiness, forgetfulness, intolerance to low temperatures, and low pulse rate. A person with hyperthyroidism, or overactive thyroid, may suffer from elevated blood pressure and pulse rate as a result of an excessive amount of thyroid hormone released by the thyroid gland, dry skin, heavy sweating, nervousness, insomnia, neck enlargement, shortened menstrual cycles, and unstable emotions. [5–9]. It is possible for these symptoms to deteriorate over time if they are not treated [10]. Therefore, these diseases need to be taken seriously and there is a special need to distinguish

between these two diseases.

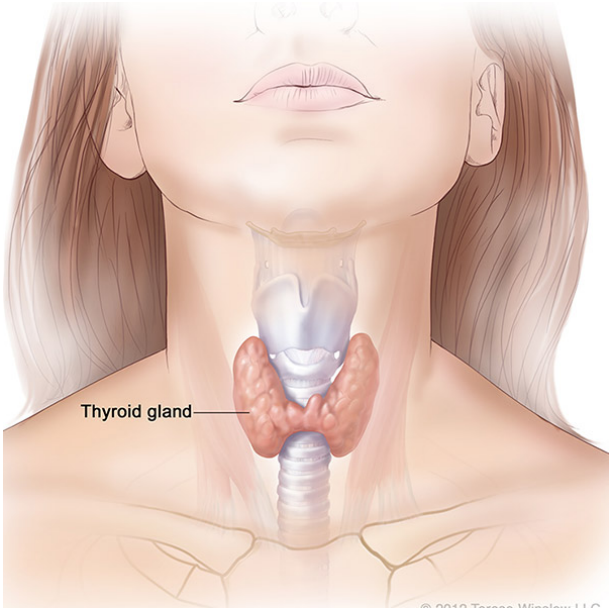

Figure 1.1: Thyroid Gland [1]

1.2 Research Aims and Objectives

In light of the above background, this research mainly focuses on the classification of thyroid disease into three specific types. In this case, we can use a number of technologies employed to assist doctors in diagnosing thyroid disorders. The first step was to integrate and clean the UCI dataset to obtain a large number of datasets. The second step was to determine which characteristics are more likely to indicate thyroid disease and use this integrated 25 thousand dataset to analyze the data. For example, Females are more likely to have thyroid nodules than males [11]. A key part of this study was to use machine learning algorithms to predict and classify thyroid diseases and to compare the strengths and weaknesses of different algorithms to determine which was best at predicting thyroid diseases. This report compares the performance of 10 classification algorithms like K-Nearest Neighbor (KNN), AdaBoost (AdB), XGBoost (XGB), Gradient Boosting, Random Forest, Decision Tree, Bagging classifier, CatBoost, MLP classifier, and Extra Trees analyzing five thyroid hormone labels (TSH, T3, TT4, FT4, and T4U) and other high related labels in UCI thyroid disease dataset to classify the type of thyroid disease into three classes, namely negative (normal), hypothyroid, and hyperthyroid.

1.3 Research Necessity and Contributions

It is challenging to diagnose, evaluate, and manage thyroid nodules. In order to identify a thyroid disorder from a laboratory test report, extensive knowledge and experience are required. The diagnosis of thyroid disease is traditionally achieved

[1] https://www.fastnewsfeed.com/health/thyroid-disease-causes-symptoms-and-treatment/4/

through blood tests that measure thyroid hormone levels, imaging tests that detect thyroid nodules, and physical exams, and then the experts make a diagnosis based on these results [12, 13]. However, one of the most misunderstood and undiagnosed diseases is thyroid disease, which is a subset of endocrinology [8, 14]. For example, if you go for a blood hormone test in the morning, your body is not affected so it is "normal", but when you exercise, are hungry, etc., you will get feedback from hormone regulation and it will affect the hormone levels in the current condition. Therefore, it might influence the diagnosis result and make incorrect feedback from doctors. Early disease detection, diagnosis, and treatment can prevent disease progression and even death [15]. Therefore, we can use machine learning techniques to help doctors diagnose thyroid disease. Nowadays, machine learning algorithms appear to be a useful diagnostic tool for identifying thyroid diseases and their progressed stages with high accuracy [16, 17]. Therefore, with technological advancements, thyroid gland diagnosis problems can be solved well. Many researchers worked with various algorithms to predict thyroid disease already. For example, Shahid et al. proposed three classification models based on Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) algorithms and achieving 95.81% accuracy to diagnose the thyroid disease [18]. If these algorithms are executed properly, they can substantially reduce human error and provide a reliable replacement for medical professionals [19]. Therefore, machine learning is predicted to become more prevalent in healthcare in the near future [20].

In spite of this, most of the data samples they used had limitations and the algorithms they selected were not comprehensive enough. As a result, this study addresses these issues by focusing primarily on:

1) By integrating all thyroid nodule disease-related documents which is the most comprehensive thyroid database available from UCI, the sample size will be increased.

2) Finding relationships between common features in the data, such as the correlation between thyroid nodule disease and age, gender, and hormonal indicators of patients, as small findings.

3) Using ten common and emerging commercially available machine learning methods, the best models were trained and compared to address the problem of old and insufficient algorithms. Compared to other papers, our results show that the accuracy of the best performance model (Random Forest algorithm) can reach 98.85% which is an very excellent result.

1.4 Project Structure

A total of six chapters are included in this project. Upon completing the introduction, the second chapter will analyze thyroid-related diseases and compare the literature related to machine learning for thyroid disease prediction. In Chapter 3, the data processing process and analysis of the processed data are described in detail. Specifically, chapters 4 and 5 are the focus of this project. Chapter 4 introduces and

classifies the machine learning required, while Chapter 5 compares the prediction results under different data processing methods. A conclusion to the project is provided in Chapter 6, where the limitations of the study are discussed and an idea for further research is provided.

Chapter 2

Literature Review

In this chapter, there are two sections which are the thyroid disease mechanism and machine learning techniques. The former explains how thyroid disease occurs and how hormones are involved and the latter makes a comparison of existing machine learning techniques used in thyroid disease classification is presented.

2.1 Thyroid Disease Mechanism

Health begins and ends with the balance of the endocrine system and the thyroid gland is a very important part of the endocrine system. The main functions of the thyroid gland are to maintain body temperature and normal metabolism, to participate in the three major energy supplies, and to regulate gonadal metabolism acting directly on the ovaries. As for how it works, it has been demonstrated as a process diagram in figure 2.1. When the body senses a caloric deficit, the adrenal cortex sends instructions to the thyroid and pituitary glands. The hypothalamus releases TRH (thyrotropin-releasing hormone) acts with the pituitary gland, which then releases TSH (thyroid stimulating hormone) acts on the thyroid gland. The thyroid gland is activated and produces T4 (a very low vitality hormone), which reaches the liver and is transformed into vital T3, which enters the bloodstream and is transported to all parts of the body by thyroglobulin. The cells become active in response to T3 and accelerate heat production [1, 21, 22]. However, it has a multi-layered feedback mechanism. Lack of T3/T4 means no inhibition of TRH/TSH release so elevated levels of TSH. The increasing level of TSH will stimulate thyroid gland growth that's why hypothyroid patients who have insufficient T3 and T4 would have a big neck which thyroid goitres form "benign goitre".

According to the latest research by Taylor et al. [23], the main causes of thyroid disease may be: gender, insufficient iodine intake, excessive iodine intake, the transition from iodine deficiency to adequate iodine intake, other autoimmune conditions, genetic risk factors, smoking, alcohol consumption, drug abuse, selenium deficiency, infection and syndrome. A patient with abnormal thyroid function normally falls into three categories, as we discussed earlier: hypothyroidism, hyperthyroidism, or a normal thyroid gland. Hypothyroidism is characterized by an insufficient amount of hormones being produced by the thyroid gland, which is reflected in increased TSH levels and decreased FT4 levels. Hyperthyroidism occurs when the thyroid gland produces more hormones than the body needs, resulting in decreased TSH

and increased FT4 levels. The final one means no thyroid disease. In order to assess the various functions of the thyroid gland, thyroid function tests such as TSH, T3, T4U, Total T4, and Free T4 Index (FTI) are performed. Normally, the thyroid gland produces a normal TSH and normal T4, hyperthyroidism is indicated by a low TSH and high T4, and primary hypothyroidism by a high TSH and low T4. Testing for T3 determines the severity of hyperthyroidism or diagnoses it [24]. Especially for the female, hypothyroidism and premature ovarian failure, often accompanied by decreased estrogen, luteal insufficiency, and non-ovulation, may lead to a significant increase in maternal and infant risk for pregnant women, including preterm birth and stillbirth, and in fetal and infant life, thyroid hormone deficiency may cause brain damage and mental deficiency [25]. A pregnant woman and her baby may suffer health problems as a result of it. A genetic background is found to be responsible for approximately 70% of the risk of developing thyroid disease [26]. In addition, systemic lupus erythematosus, which is an autoimmune disease, is associated with thyroid disease risk. Thyroid disease and low vitamin D levels are associated with some evidence [27].

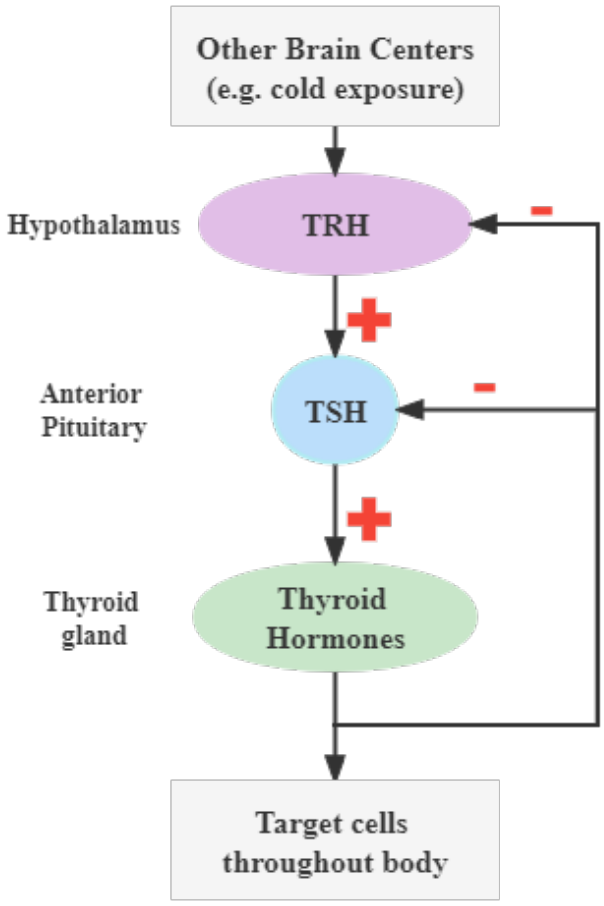

Figure 2.1: Thyroid Hormones Production

In addition, thyroid nodules and abnormal thyroid function are not the same thing but can be intertwined. Many years ago, big neck disease (simple goiter) is due to iodine intake is insufficient, the thyroid gland diffuse enlargement line disease. But now that nutrition is abundant and most of the salt eaten contains iodine, goiter is basically invisible. In its place, there are three other common types of nodules [28]. The first type of nodule, usually not operated, is a benign nodule, but if it is too large or displaced, compressing the surrounding blood vessels or airways, it needs to be removed. The second type of thyroid adenoma, hyperthyroidism, hypothyroidism or Hashimoto's thyroiditis (an autoimmune disease in which a series of inflammatory conditions develop because of autoantibodies). This is characterized by a slow decline from hyperthyroidism to a stable hypothyroid state. The third category, thyroid cancer, is steadily increasing by 10-20% per year, making it one of the highest malignancies. It is not clear how it comes about, but it is related to family history, radiation, smoking, obesity, increased stress, etc. Compared to other cancers, the age of onset is very low and there are more women than men. It is highly prevalent and there is no way to prevent it. However, thyroid cancer is one of the four major cancers, and the other three are basal cell carcinoma, Luminal breast cancer and prostate cancer [29]. But thyroid cancer has very low mortality rates. Therefore, in the Japanese guidelines for thyroid cancer, even if surgery is not performed, observation and regular ultrasound review are sufficient for a single small central nail cancer, and if the growth is slow, no surgery will be performed all the time, and if it progresses rapidly, surgery will be performed immediately. This option takes into account both life and quality of life, because if surgery is done, there may be other complications that affect life and one cannot live without thyroid medication to replenish the thyroid gland [30].

2.2 Machine Learning Techniques

Labeled data is used for learning in classification algorithms through supervised learning. A large number of features or attributes are required in the labeled training sets for these algorithms. Based on the available query to the literature we found that the majority of the literature using the UCI thyroid disease database only selected one of the files with data size of 215 [4, 31, 32]. For such a data size of 215, although it can also be used to train the model, the results obtained are not convincing even if they are relatively good. For example, in this paper, a hybrid algorithm IEDA combining immune algorithm and estimated distribution algorithm is proposed, however, even if such accuracy of 98.39% is obtained [31] may be hard to be fully convinced. But if the amount of data is increased then it will also improve the accuracy of the model significantly. Not to mention that there is also 69.77% such accuracy generated by the back propagation algorithm [32]. So we need to expand the number of datasets.

Researchers have used various machine learning algorithms to diagnose thyroid disease. Although their datasets are sourced from UCI, they have taken only a few machine learning algorithms and have not put together a large number of machine learning algorithms for comparison. The main point is that the same algorithm

does not produce exactly the same results when analyzed by different people. For example, the accuracy of the model trained with SVM algorithm is 96.75%, 99.02% and so on [18, 33] and the SVM algorithm has the best result in their research. In addition to this, in other literature it is shown that Random Forest or MLP is the best algorithm to classify thyroid disease with an accuracy of 98.98% and 96.4% respectively [34]. However, since their accuracies are based on the results obtained with different data processing and different environments, they do not constitute a comparative relationship. The main problem is that these papers do not show the boosting algorithm well. For example, XGBoost, AdaBoost and other algorithms are popular algorithms in recent years especially in the field of classification are widely used, but they are not common in these research.

Therefore, this study will not only address the problem of insufficient data sets, but also introduce ten machine learning algorithms, including the recently popular boosting algorithm, and the common bagging algorithm, as well as other machine learning algorithms that do not belong to these two categories, to compare which one is more suitable for the classification of thyroid diseases.

Chapter 3

Thyroid Data Analysis

3.1 Introduction of Dataset

The thyroid datasets taken from the UCI Machine Learning Repository were used [1]. There are total eight files to distinguish different types of thyroid disease. Other files are not satisfied with our requirement which means they have no specific hypothyroid or hyperthyroid classification. In this research, these data sets which satisfy the analysis requirement will be merged into one file first so that we can acquire as much data as possible, and also it is easier to analyse and process the thyroid data. There are 19 attributes after data integration which consist of (age, sex, on_thyroxine, query_on_thyroxine, on_antithyroid_medication, sick, pregnant, thyroid_surgery, query_hypothyroid, query_hyperthyroid, lithium, goitre, tumor, TSH, T3, TT4, T4U, FTI, and Target). For these float attributes, TSH means thyroid stimulating hormone, T3 means triiodothyronine which is a hormone that the thyroid gland produces, TT4 means thyroxine which is a type of the thyroid hormone that regulates metabolism, T4U mean T4 uptake, and FTI mean Free T4 Index.

3.2 Data Preprocessing

3.2.1 Data Integration

Thyroid disease classification is the main purpose of this research. According to a previous survey, it will be divided into three classes which are: hypothyroid, hyperthyroid and normal. However, the target in each data file is different which means we need to determine the meaning of the categories within each of the different files and group them into one of the three categories based on the instruction documents. The Target information of these eight data set files are demonstrated below:

- *allhyper-train.csv* and *allhyper-test.csv*: [hyperthyroid, T3 toxic, goitre, secondary toxic and negative] They are collectively referred to 'Hyperthyroid' class
- *allhypo-train.csv* and *allhypo-test.csv*: [hypothyroid, primary hypothyroid, compensated hypothyroid, secondary hypothyroid and negative] They are

[1] https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

collectively referred to 'Hypothyroid' class

- *ann-train.csv* and *ann-test.csv*: [normal (not hypothyroid), hyperfunction and subnormal functioning] normal is 'negative', hyperfuction means 'hyperthyroid', and subnormal functioning is 'hypothyroid' based on the description of the data we determine

- *hypothyroid.csv*: [hypothyroid and negative] The name is same as the class

- *thyroid0378.csv*: [A,B,C,D - hyperthyroid conditions, E,F,G,H - hypothyroid conditions, I,J - binding protein, K - general health, L,M,N - replacement therapy and R - discordant results] The target in this class uses letters to indicate, which each letter represents a thyroid disorder. They have surely be grouped into six broad categories. However, the class hyperthyroid conditions and hypothyroid conditions are the tow satisfy our requirements. Therefore, we label A,B,C,D into 'hyperthyroid' and E,F,G,H into 'hypothyroid'. Other data that do not meet the conditions are simply discarded.

After data integration, it is the final clean dataset that combines all the raw data together which has **25453** entries and 19 attributes (shown below) which are demonstrated above. The three targets are: hyperthyroid, hypothyroid, and negative.

Column	
age	continuous, float
sex	F-Female, M-Male
on_thyroxine	t/f
query_on_thyroxine	t/f
on_antithyroid_medication	t/f
sick	t/f
pregnant	t/f
thyroid_surgery	t/f
query_hypothyroid	t/f
query_hyperthyroid	t/f
lithium	t/f
goitre	t/f
tumor	t/f
TSH	thyroid stimulating hormone
T3	triiodothyronine
TT4	thyroxine
T4U	T4 uptake
FTI	Free T4 Index
Target	hyperthyroid, hypothyroid, negative

3.2.2 Data Cleaning

A set of missing data in this dataset was identified that were age, sex, TSH, T3, TT4, T4U and FTI whose missing rate are 1.76%, 2.448%, 7.577%, 17.452%, 4.185%, 6.673%, and 6.626% respectively which are not very large amount. Therefore, we can consider drop them first. However, we have noticed that the number of hyperthyroid and hypothyroid class are much smaller than the normal class. Therefore, we decided to drop these lost data whose class is negative and replace the rest missing value of their corresponding mean value. As for the sex column, there are only 58 missing values which are filled with female because female patients are more than male based on the previous research. After working in this way we have dropped 5892 instances and the final total number is **19555** instances.

3.3 Data Analysis

Figure 3.1 shows the whole data analysis of the thyroid data set. We analyze the distribution of samples first which is the first pie chart. The proportion of negative, hypothyroid, and hyperthyroid are 90.5%, 7.2%, and 2.3% respectively. It is extremely uneven and certain measures are needed to deal with the sample imbalance. We will discuss the solution in the next chapter. The second bar chart shows the age distribution who have suffered from hyperthyroid or hypothyroid. We can notice that the elderly are more likely to suffer from thyroid dysfunction compared to the younger population which is more than 50% of the elderly (50-100) associated with thyroid disease. Because age affects thyroid hormone secretion, metabolism, and action in several ways. It is common for serum levels of thyroid stimulating hormone and T3 to decrease with aging, while serum-free T4 levels tend to remain unchanged [35]. In addition, middle-aged people (30-50) also are easily influenced by thyroid disease which is second only to the elderly. With such a high incidence of thyroid diseases, it is necessary to gain a clearer understanding of the risk factors for thyroid diseases, as well as improve the accuracy of diagnosis. Especially, there is also a clear difference in the gender distribution of patients that the females more than males whose the proportion is 78.36% and 21.64% in the fourth chart of figure 3.1. It is because the thyroid hormone has sex-specific effects on neural circuit organization. Due to this, thyroid hormone-related problems are on the rise among females [36]. Mood disorders might also be complicated by thyroid dysfunction, which is common among females. Therefore, it was innovative and important for the laboratory to test by age and gender. Furthermore, even though some patients have done thyroid surgery, 6.49% of them still suffer from thyroid disease. This means that after thyroid surgery, such as removal of the thyroid gland, there is still a chance that you will still have thyroid disease. There is a study result showing that about one in seven patients who undergo thyroid surgery develop hypothyroidism requiring thyroid hormone therapy which supports above finding. Especially, there is an increased risk for patients with preoperative TSH levels exceeding 1.5 IU/mL, low free T4 levels, and Hashimoto's thyroiditis [37]. It is therefore important to determine whether surgery is necessary and what the post-operative recovery process will be like. It is important to consider carefully whether surgery is necessary and to

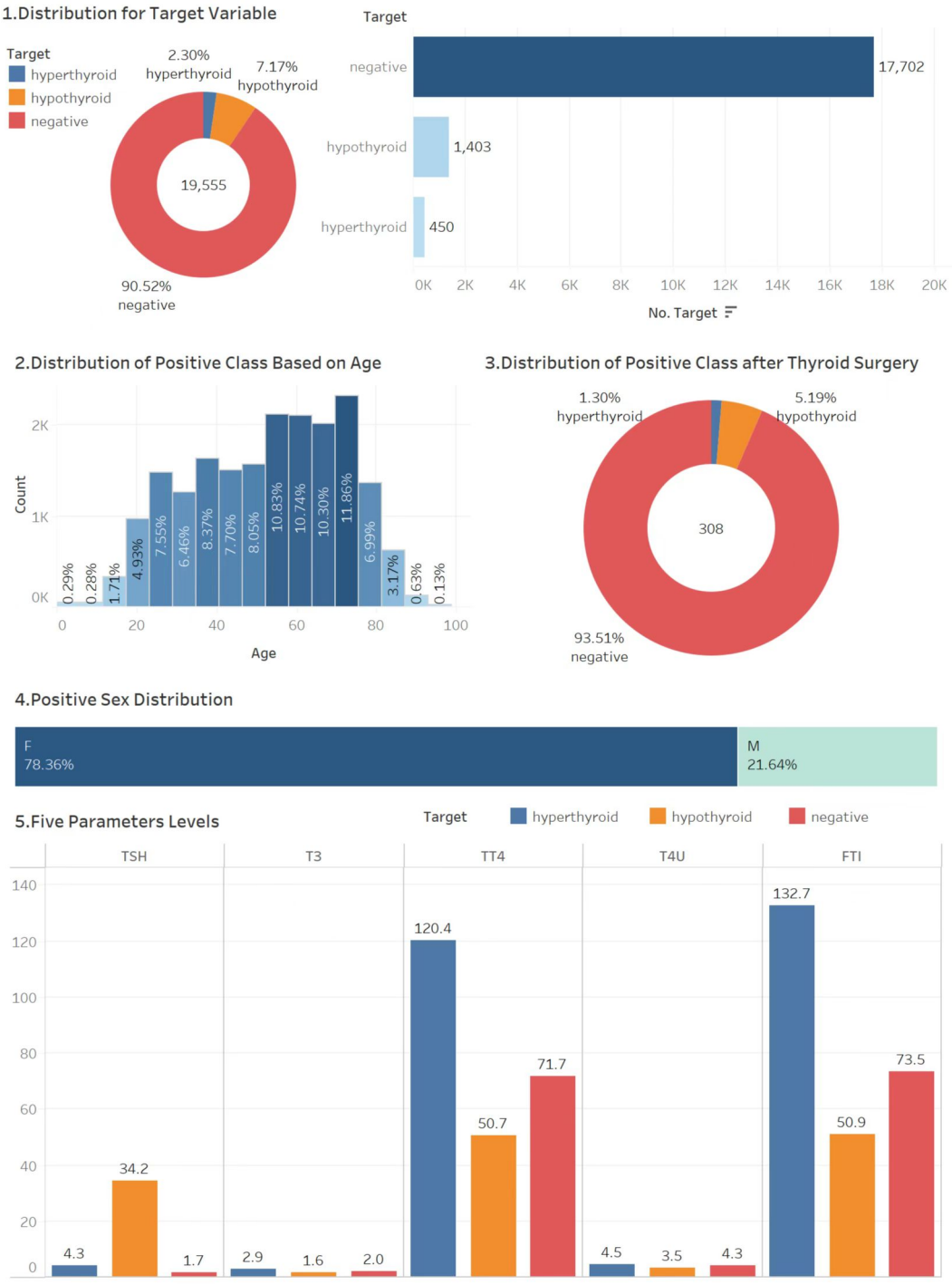

Figure 3.1: Thyroid Disease Data Analysis

follow the advice of the surgeon after surgery.

From the thyroid hormone level chart, we can know that hypothyroid patients have extremely higher TSH, but hyperthyroid patients comparatively lower TSH. Hyperthyroid patients have higher TT4 and FTI, but hypothyroid patients have lower TT4 and FTI. As we mentioned before, a hypothyroid condition is characterized by an insufficient thyroid hormone production. Therefore, T3 and TT4 will be lower in comparison and this phenomenon shows in the figure too. TSH promotes thyroid production but it is affected by negative feedback from thyroid hormones, so hypothyroid patients will have higher TSH levels. In contrast, for the hyperthyroid patients, they have higher levels of T3 and TT4, but lower TSH levels. However, these are just some basic principle of thyroid disease judgement. We need more other related hormone levels and the reaction of thyroid patients to determine the disease. Furthermore, the levels of TT4, T4U, FTI are highly related. Because T4U and FTI can determine how much T3 can transfer to T4. The five hormones in the figure are five routine indicators of thyroid disease screening, so all five should be put into machine learning as features in the prediction model.

Chapter 4

Methodology

4.1 Process Overview

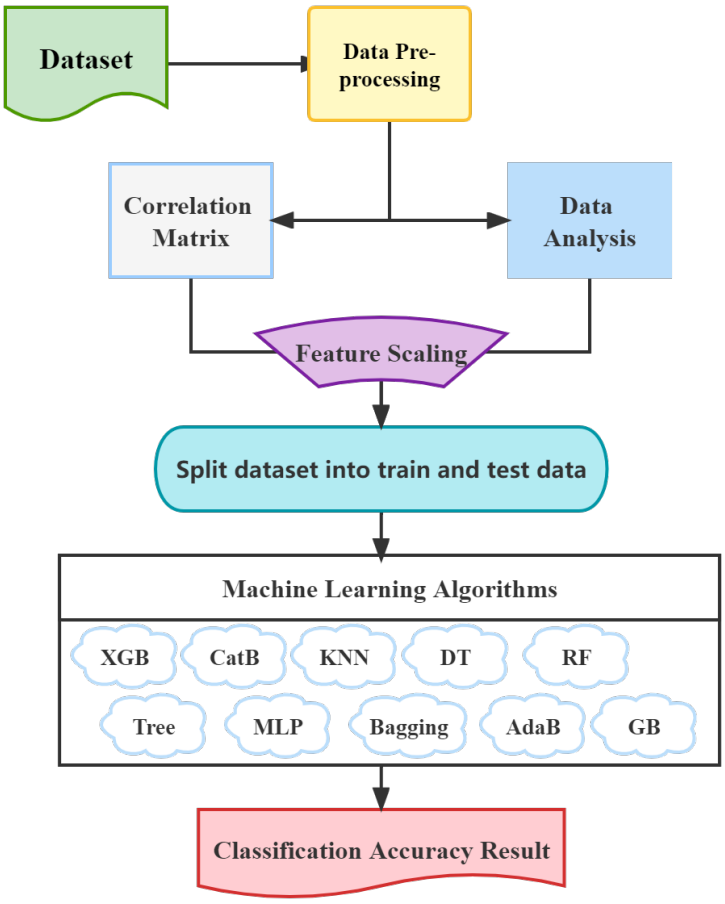

Figure 4.1: Process overview

The figure showing above is the whole process in this project. First, a set of datasets will be integrated, pre-process, and analysis of the feature distribution which we have demonstrated in the chapter 3. Then, we will consider together based on the

data analysis results and combined with the value of correlation matrix to choose the useful features which implements the feature scaling. The thyroid datasets of the patients need to be analyzed in order to discover meaningful information regarding thyroid disease diagnosis. After that, the new dataset will be split into two data which are train data and test data. The most important part is modeling that there are ten different machine learning algorithms used to train the models. In spite of this, imbalance data problem would be solved and the created new dataset will be used to train the models. Compared the results in different processing methods through calculating the precision, accuracy, recall and F1-score. In our study, the key aim of using machine learning techniques is to differentiate between three forms of thyroid disease which are hyperthyroid, hypothyroid and negative who do not have any thyroid issues. Therefore, we mainly focus on the algorithms that can be implemented well in classification.

4.2 Feature Scaling

In order for a given classifier to perform more efficiently, features must be scaled. It is the age of big data, some datasets store a plenty of non-relative features. As a result, these extra features increase the traning time and increase overfitting risk exponentially. It is possible to reduce the average time for predicting and training by using feature selection technique. In order to save time and cost, these selected features were then used for training and testing. Classification results are greatly impacted by such techniques [12]. In our study, we use correlation matrix to recognize the level of relationship between the training features and Target. We would select the highest correlations, plus a few features that we think are of interest in medical data analysis. After that, models would be trained based on these features.

4.3 Ensemble Learning Methods

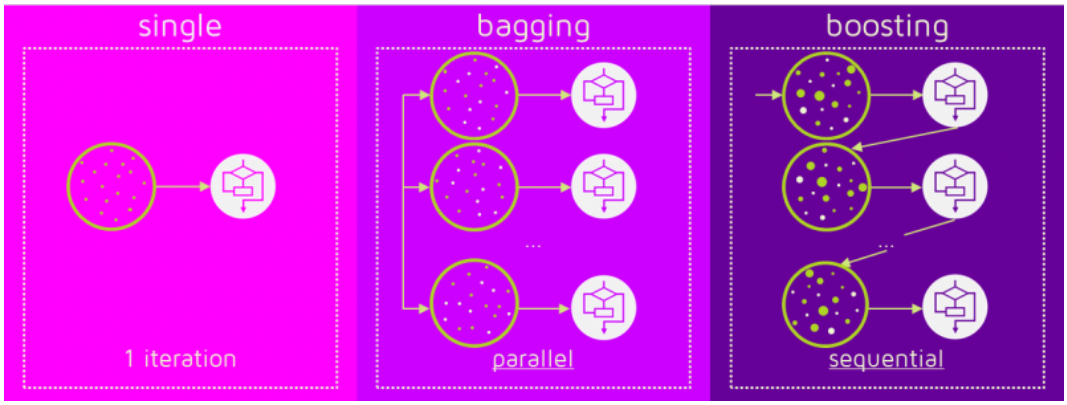

Figure 4.2: Bagging vs Boosting

Ensemble Learning models combine multiple models to produce powerful models. By combining multiple weak learners, Ensemble Learning produces better results, stability, and predictive power than by using any one model alone. In the

bioinformatics domain, ensemble learning method is more suitable for handling classification problems [38]. Therefore, some ensemble learning algorithms are used here to predict the thyroid disease. It is mainly divided into bagging, boosting and stacking. In this research, we not only can train a model to predict the thyroid disease more accurately, but also the results of different ensemble learning methods are compared. It is helpful to find what type of algorithm is more suitable for bioinformatics domain.

4.3.1 Bagging

Bagging (Bootstrap aggregation) which is a parallel ensemble learning method is the most basic type of meta-algorithm for decision trees and it works best with strong and complex models. By using the voting scheme, the bagging result is the most appropriate label for classification. It is well known that machine learning models are constantly subject to bias-variance trade-offs. The main advantage of bagging is that it reduces variance while keeping bias virtually unchanged because it is not very susceptible to overfitting with noisy data. Bias usually increases with a reduction in variance. The downside, however, is that this increases bias [39]. In our experiment, decision trees and bagging classifier are used as a representative of bagging.

4.3.1.1 Random Forest

Random forest is very similar to bagging. The only difference is that random forests do not split all the features of a component tree when splitting a node. Suppose there are m features overall, the size of the subset can be any number from 1 to m-1, with the most common choices are $\sqrt{m}$ and $log(m)$. As deep decision trees are averaged from different samples of training data, RF is less affected by noise in attributes [40]. It is a highly accurate, robust, and stable method because a large number of decision trees participate in it. There is no overfitting problem with it. This is because it averages all the predictions, canceling out any biases [41, 42]. Since RF has a higher degree of robustness, its performance is better than AdaBoost, another ensemble learning technique we will discuss later. However, as a result of having multiple decision trees, random forests have a slow time in generating predictions. Trees make predictions for the same input every time the forest makes a prediction, then vote on it. This process takes a long time [43].

4.3.1.2 Extremely Randomized Trees (Extra Tree)

Extra Tree is a variation of Random Forest. Each decision tree's training set is selected by random sampling bootstrapping, whereas extra trees do not use random sampling. Instead, they use the original training set for each decision tree. RF decision trees divide the points by an optimal value based on Gini coefficients, mean squared deviations, and so on, just like traditional decision trees. A feature value is randomly selected to divide the extra trees in a more radical manner [44].

4.3.2 Boosting

In contrast to bagging, boosting is a sequential ensemble learning method. By iteratively shifting the focus on problematic observations that were difficult to predict,

they try to boost the performance of a weak learner. By adding the weak learners, a strong learner is formed. As the figure 4.3 shown, a model was created after the first training. Afterwards, the wrong feature point weight is enlarged for the second training since there is a wrong classification in the middle. The final model is then obtained by following the same steps

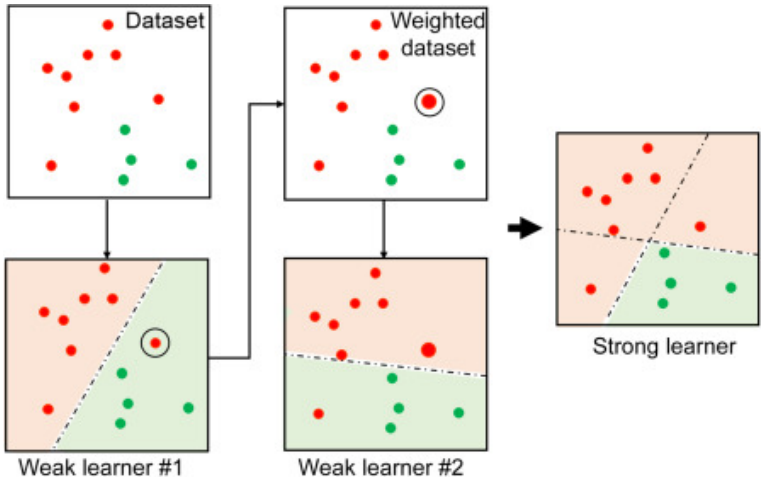

Figure 4.3: Boosting Method

4.3.2.1 Gradient Boost Decision Tree (GBDT)

Although GBDT also consists of many decision trees, it differs from random forests in a number of ways, one of which is that the trees in GBDT are all regression trees. Another difference is that the trees in GBDT all build on the previous tree, rather than through a voting mechanism. GBDT has a wide range of applications in areas such as search, advertising, and recommender systems, and is highly explanatory. However, because of the interdependence between trees, strong training time is required.

4.3.2.2 AdaBoost

The AdaBoost algorithm is also a boosting algorithm that combines several weak-learner classifiers into a robust one. Based on the decision tree stump, different weight coefficients are assigned to the classifiers [45]. Observations that had previously been misclassified are up-weighted. The final classifier for each weak iteration $h(x_t)$ after i iteration is,

$$E_i = \sum E[F_{(i-1)}(x_t) + \alpha_i h(x_t)]$$

Here, α_i is the assigned coefficient.

However, high-weighted data points identify the ”shortcomings.” The exponential loss of AdaBoost gives a higher weight to samples that fit poorly. It is considered to be a special case of Gradient Boost in terms of loss function, in which exponential losses occur.

4.3.2.3 XGBoost

Tianqi Chen developed XGBoost in order to push the limits of computations for boosted algorithms [46] which takes the best parts of AdaBoost and random forests and adding additional features. They are sequential tree growing, minimizing loss function using gradiant descent, parallel processing to increase speed, and regularization parameter. Therefore, it is very popular in recent years.

4.3.2.4 CatBoost

In 2017, Yandex, the Russian search engine giant, opened sourced CatBoost, a machine learning library. In this implementation, symmetric decision trees (oblivious trees) are used as the best learner since they have fewer parameters, can support categorical variables, and are highly accurate. In CatBoost, Categorical and Boosting are combined to emphasize the efficient and reasonable processing of categorical features. By addressing Gradient Bias and Prediction Shift, CatBoost reduces overfitting and improves the accuracy and generalisation of the algorithm [47].

4.4 Other Machine Learning Techniques

4.4.1 K-Nearest Neighbor (KNN)

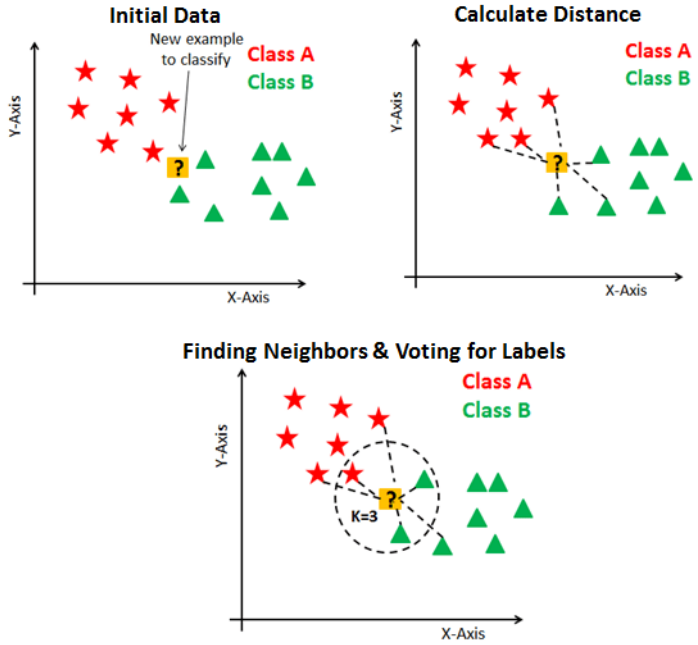

Figure 4.4: KNN Principle

The K-nearest neighbors algorithm (KNN) is one of the most popular supervised machine learning algorithms. Instead of creating a model first, k-nearest neighbor classifies a data point based on the classification of its neighbors [48]. By using the Euclidean distance function, it calculates the distance between two new instances based on a voting strategy. Figure 4.4 demonstrates the working principle of KNN. There are two classes in this figure which are class A and class B. The yellow box

point "?" is a new example which would be predicted by the algorithm. Each point has their own feature vector and their distance would be calculated based on the feature vector. Finally, according to the number of K (k=3 in the figure), three points would be selected which are closest to the new point. The class of the new point would be decided by voting.

As a result, KNN is easier to use than other algorithms since there is no need for special model construction [40]. There are several drawbacks to KNN, including the need for a metric to measure distances between data objects to calculate the case neighborhood. In class membership estimation, there is only one variable that needs attention: k, which indicates how many nearest neighbors will be used. This study aims to optimize KNN model performance by varying its value and comparing it with other similar models to obtain the best classification performance when the value of K is changed.

4.4.2 Multi-Layer Perceptrons (MLP)

With MLP, you can quickly adjust the network weight as well as determine the new weight and bias using gradient descent. Like a human neuron, a multilayer perceptron generates outputs from inputs just like a feedforward artificial neural network. MLPs use backpropagation for learning and consist of several layers of input nodes. It is widely used in supervised learning problems, computational biology, and parallel distributed processing analysis because it supports parallel implementation, generalization, fault tolerance, and parallel distributed processing analysis. There are several applications, including speech recognition, image recognition, and translation. [49].

4.5 Imbalance Data

There are two methods to solve imbalance data problem. The first is oversampling, which is increase the number of instances for these less number classes. Another is undersampling, which is decrease the number of instances to make all the classes have the same size.

4.5.1 Oversampling - SMOTE

SMOTE is a feature space-based oversampling method that synthesizes new features from a few sample classes and their neighbors, then composes new samples [50]. By artificially synthesizing samples, SMOTE alleviates overfitting caused by randomly copied samples, although it has some limitations like we cannot define the quality of synthesizing samples.

4.5.2 Undersampling - Resample

In this situation, to change the class distribution, resampling methods are used to remove examples from the training dataset. It delete or merge examples in the majority class.

4.6 Performance Evaluation Metrics

Before we get the most accurate results, we analyse the performance of the classifiers with the default parameters, and then select the best ones for further tuning to reduce the time spent on the selection of bad classifiers. These measures are derived from confusion matrix for the purpose of evaluating diagnostic tests. In a confusion matrix, there are four distinct terms: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). True positive means that the system predicts the outcome correctly and the result also confirms the prediction. False positives occur when the system predicts the outcome as correct, but the result is incorrect. It is true negative if the system predicts a false outcome and also predicts a false result. When a system predicts a false value, but the outcome turns out to be true, it is called a false negative [51].

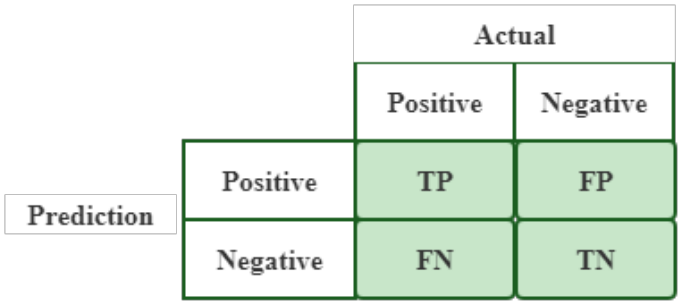

Figure 4.5: Confusion Matrix

4.6.1 Fitness Function

Following performance metrics are used to analyze the performance of different Machine Learning classifiers by representing the fitness value of each individual. The equation of accuracy, precision, recall, and f1-score are showed by the following.

- Accuracy measures the fraction of predictions our model got right

  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

- Precision is used to measures the classifier exactness

  $Precision = \frac{TP}{TP+FP}$

- Recall measures the classifier completeness

  $Recall = \frac{TP}{TP+FN}$

- F1-Score measures the balance between precision and recall

  $F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall}$

4.6.2 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic is the full name of this analysis tool, which is a two-dimensional curve called the ROC curve. False positive rates (FPR) are horizontally distributed in the plane, while true positive rates (TPR) are vertically distributed. The performance of a classifier on the test sample can be used to obtain a pair of TPR and FPR points. As a result, ROC points can be assigned to the classifier. This ROC curve can be obtained by adjusting the threshold used to classify this classifier. Generally, this curve should be above the line between (0, 0) and (1, 1). A random classifier can be represented by the ROC curve formed by the (0, 0) and (1, 1) lines. An intuitive remedy if the classifier results in a negative classification is to reverse all the predictions. If the classifier produces a positive classification, then the final classification is also negative. A ROC curve can be useful in representing the performance of a classifier because it is intuitive and intuitive. To measure how good or bad a classifier is, one would always prefer a numerical value. Hence the Area Under roc Curve (AUC) came into existence. AUC represents the area under the ROC curve as indicated by its name [52]. In our study, classification models are measured by AUC (Area Under ROC Curve).

Chapter 5

Results and Discussion

5.1 Feature Selection

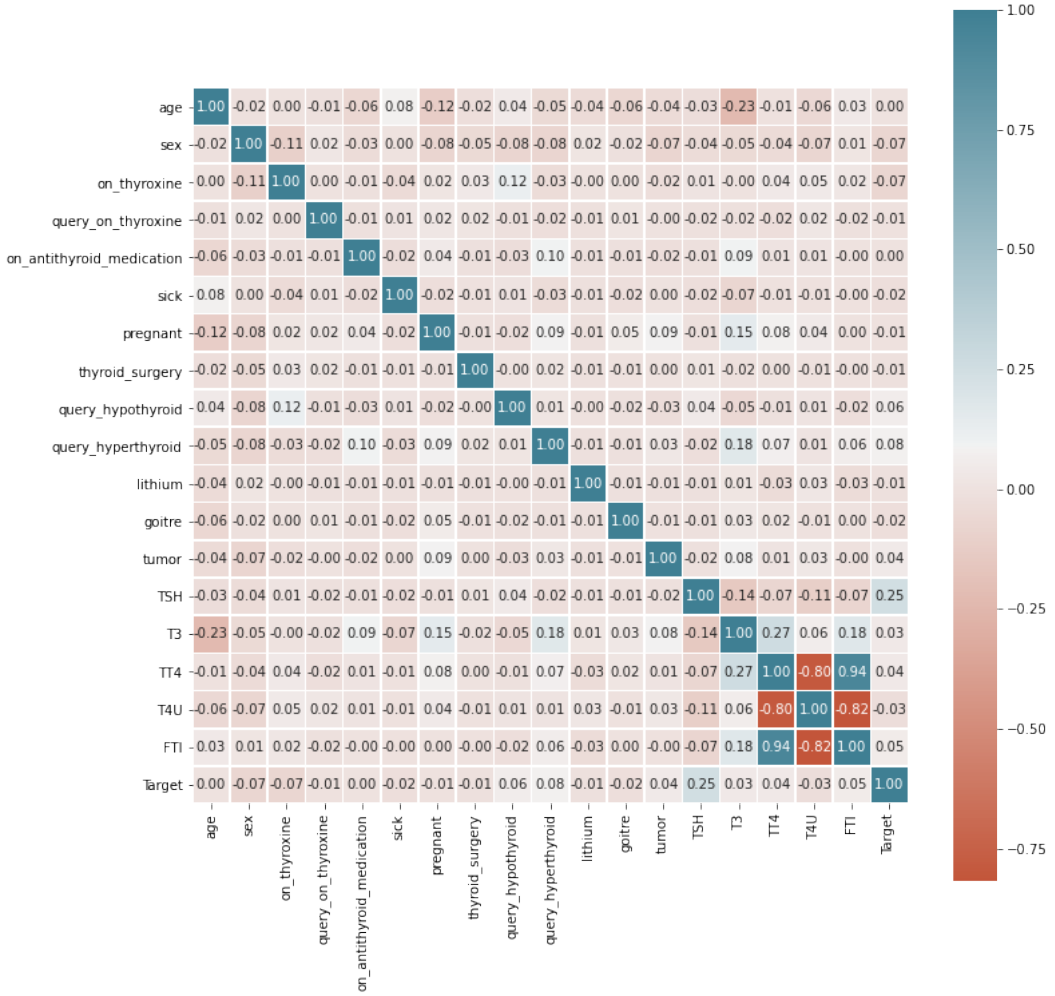

Figure 5.1: Correlation Matrix

Figure 5.1 shows the correlation matrix of this dataset. It is obvious that three thyroid hormone levels have extremely high relationship than others. However, we

need to pay attention to the correlation value of Target. Indicators with relatively high correlation will be retained and used as one of the features of the training model. In this situation, the column whose correlation value is more than 0.04 will be selected as the training features. They are sex (7.15%), on_thyroxine (6.91%), query_hypothyroid (6.07%), query_hyperthyroid (7.91%), tumor (4.45%) and the five hormones which are TSH, T3, TT4, T4U, and FTI. No matter how many these hormone correlation values are, they are the necessary features to train the thyroid disease prediction model based on the data analysis results. Also, these relevant values also confirm the results of the previous data analysis. For example, TSH is one of the most important attributes to diagnose thyroid disease, with a value of 25%. Based on the results of data analysis, we have known that a high TSH means he/she would suffer from hypothyroid, and a low TSH means he/she would get hyperthyroid. Similarly, gender is also a very important attribute to measure whether a person is prone to thyroid disease, i.e., women are much more likely to have the disease than men, so its relation value of 7% is highly relative among these attributes. These five types of hormones influence each other, so they should be one of the indicators of training to keep modeling, regardless of the relevant values.

Features	Information
sex	Female(F) or Male(M)
on_thyroxine	True(t) or False(f)
query_hypothyroid	True(t) or False(f)
query_hyperthyroid	True(t) or False(f)
tumor	True(t) or False(f)
TSH	float64
T3	float64
TT4	float64
T4U	float64
FTI	float64

Following are the explanation of these arttributs:

- on_thyroxine: whether they have taken thyroxin medicine.
- query_hypothyroid: patients query if they have hypothyroid.
- query_hyperthyroid: patients query if they suffer from hyperthyroid.
- tumor: whether the patients have tumor.
- TSH: Thyroid Stimulating Hormone.
- T3: Triiodothyronine
- TT4: Thyroxine
- T4U: T4 Uptake
- FTI: Free T4 Index

5.2 Modelling

Ensemble learning techniques are mainly used in this research to predict thyroid disease. The bagging methods are Decision Tree, Bagging Classifier, Random Forest, and Extra Trees. For the boosting methods, they are XGBoost, CatBoost, AdaBoost, and Gradient Boosting. There are other two machine learning algorithms that are used to compare with bagging and boosting to check if they have a better performance which are K-Nearest Neighbors and MLP Classifier.

As we mentioned before, the thyroid dataset in this research is imbalanced. The percentage of the 'negative' class is much higher than the other two which are 'hyperthyroid' and 'hypothyroid'. Therefore, it might influence the result of the modeling. However, we can try first to see the original performance of the ten techniques. From the results of table 5.1, it is obviously shown that all the methods have high accuracy results which are all above 90%. Some models, however, have very low scores which are below 0.70, based on the F1-score. Accuracy is the most commonly used metric in classification problems and it calculates the ratio of the number of correctly classified predictions to the total number of predictions. Unbalanced data sets, however, make Accuracy an unsuitable metric. As a result of the following reasons.

Assume we have 100 images that contain 91 'dogs', five 'cats', and four 'pigs', and we want to train a triple classifier that can correctly identify these animals. Dogs make up the majority of them. When the number of samples in the majority class (dogs) far exceeds that of other classes (cats, pigs), if Accuracy is used to evaluate the classifier, then even if the model performance is poor (e.g., predicting "dogs" regardless of the input images), a high Accuracy Score (e.g., 91%) can be obtained. Despite the high Accuracy Score, it is not very meaningful at this point. Similarly, in our study, the distribution of the data set was previously analyzed to be quite unbalanced in the ratio of 90:7:3. Therefore, the shortcomings of the Accuracy evaluation method are particularly significant when the data are unusually unbalanced. Therefore, we need to introduce Precision (accuracy), Recall (recall), and F1-score evaluation metrics into the multi-classification model.

Table 5.1: Performance with original data

Classifier	Accuracy	Precision	Recall	F1-Score
GradientBoosting Classifier	0.9701	0.8472	0.9073	0.8749
XGB Classifier	0.9674	0.8490	0.8999	0.8731
CatBoost Classifier	0.9655	0.8472	0.8699	0.8582
Random Forest	0.9641	0.8380	0.8801	0.8581
Decision Tree	0.9614	0.8265	0.8763	0.8498
ExtraTrees	0.9554	0.8099	0.8370	0.8224
MLP Classifier	0.9570	0.8504	0.7690	0.7954
AdaBoost Classifier	0.9526	0.8467	0.7362	0.7732
Bagging Classifier	0.9375	0.7887	0.6873	0.7301
K Nearest Neighbors	0.9343	0.8174	0.6286	0.6972

Therefore, according to the F1-Score ranking, the top three models Gradient Boosting classifier, XGBoost classifier and CatBoost classifier are boosting techniques whose scores are above 0.85 which means that performance of boosting is overall slightly better than bagging methods except for the AdaBoost classifier whose score is 0.5573 which is the lowest one in this situation. In addition, the other two algorithms KNN and MLP have worse results even though they also have above 90% accuracy, because F1-score is better evidence to judge a good model which has worse results which are 0.7954 and 0.5573. Next, we can solve the imbalance data problem to get some better models. Two solutions are available here, they are oversampling and undersampling.

5.2.1 Oversampleing - SMOTE results

Oversampling is to increase the size of data, which means using algorithms to create new instances for the 'hyperthyroid' and 'hypothyroid' two classes. SMOTE technique is a way to increase the number of instances. Through this method, the number of data has gone from 19,555 to 31,863. The performance results have been demonstrated below in table 5.2. The overall performance is better than the first time when we use the original data, whose accuracy, precision, and recall are all maintained above 95% except for the AdaBoost classifier whose is 0.8634 score. Furthermore, the results of accuracy and F1-score are almost the same which means these models are not influenced by the imbalanced data and they all have better performance. In this situation, the results of bagging and boosting all performed well. Especially, the f1-score of random forest is reaching 98.87% which is an extremely high score. In other words, the SMOTE technique which solves imbalanced data issues has significantly improved the performance of the thyroid disease prediction model.

Table 5.2: Performance with oversampling data

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9887	0.9888	0.9888	0.9887
XGB Classifier	0.9879	0.9880	0.9880	0.9879
CatBoost Classifier	0.9855	0.9856	0.9856	0.9854
ExtraTrees	0.9851	0.9851	0.9852	0.9851
Decision Tree	0.9844	0.9844	0.9845	0.9845
GradientBoosting Classifier	0.9844	0.9846	0.9846	0.9844
K Nearest Neighbors	0.9697	0.9698	0.9700	0.9697
Bagging Classifier	0.9635	0.9643	0.9639	0.9633
MLP Classifier	0.9614	0.9621	0.9613	0.9615
AdaBoost Classifier	0.8990	0.9095	0.8981	0.8979

5.2.2 Undersampling - Resample results

After that, we changed the size of the data again which is reduce the number of data in this situation, which is called undersampling. The minimal class number

is hyperthyroid which is 450. Therefore, we reduced the number of the other two classes hypothyroid and negative to 450 together too. The total number is 1,350 which is much smaller than the previous two situations. In spite of this, the models perform well in the oversampling environment, with above half of them having an accuracy of 90 percent or greater. However, it is still slightly inferior to the oversampling situation. Random Forest is still the best model whose f1-score reaches 97.25%, but three boosting models GradientBoosting, XGBoost, and CatBoost have performed well too whose results are all above 96.5% except for AdaBoost.

Table 5.3: Performance with undersampling data

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9728	0.9731	0.9731	0.9725
GradientBoosting Classifier	0.9704	0.9705	0.9707	0.9701
CatBoost Classifier	0.9679	0.9681	0.9683	0.9676
XGB Classifier	0.9654	0.9663	0.9660	0.9650
Decision Tree	0.9383	0.9382	0.9374	0.9377
ExtraTrees	0.9160	0.9156	0.9158	0.9152
MLP Classifier	0.9012	0.9009	0.9020	0.9009
K Nearest Neighbors	0.8321	0.8412	0.8300	0.8323
Bagging Classifier	0.8148	0.8138	0.8126	0.8126
AdaBoost Classifier	0.7877	0.8191	0.7895	0.7849

5.2.3 Summary

Table 5.4: Performance with oversampling data

Classifier	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9887	0.9888	0.9888	0.9887
XGB Classifier	0.9879	0.9880	0.9880	0.9879
CatBoost Classifier	0.9855	0.9856	0.9856	0.9854
ExtraTrees	0.9851	0.9851	0.9852	0.9851
Decision Tree	0.9844	0.9844	0.9845	0.9845
GradientBoosting Classifier	0.9844	0.9846	0.9846	0.9844
K Nearest Neighbors	0.9697	0.9698	0.9700	0.9697
Bagging Classifier	0.9635	0.9643	0.9639	0.9633
MLP Classifier	0.9614	0.9621	0.9613	0.9615
AdaBoost Classifier	0.8990	0.9095	0.8981	0.8979

Based on these results, demonstrates that solving the data imbalance problem can significantly improve the model's accuracy, precision, and recall. In our research, SMOTE technique is the best method because it produces the best outcome. The F1-score of Random Forest, XGBoost, CatBoost, Extra trees, Decision tree, and Gradient

Boost are 98.87%, 98.79%, 98.54%, 98.51%, 98.45%, and 98.44% which are quite high results. Therefore, we use the data which is created by SMOTE technique as our training and testing dataset. The ROC figure based on the oversampling data is shown in figure 5.2 below which demonstrates an excellent outcome in which all the results are above 0.99. The best result in ROC is XGBoost classifier whose AUC is 0.999 approching to 1.

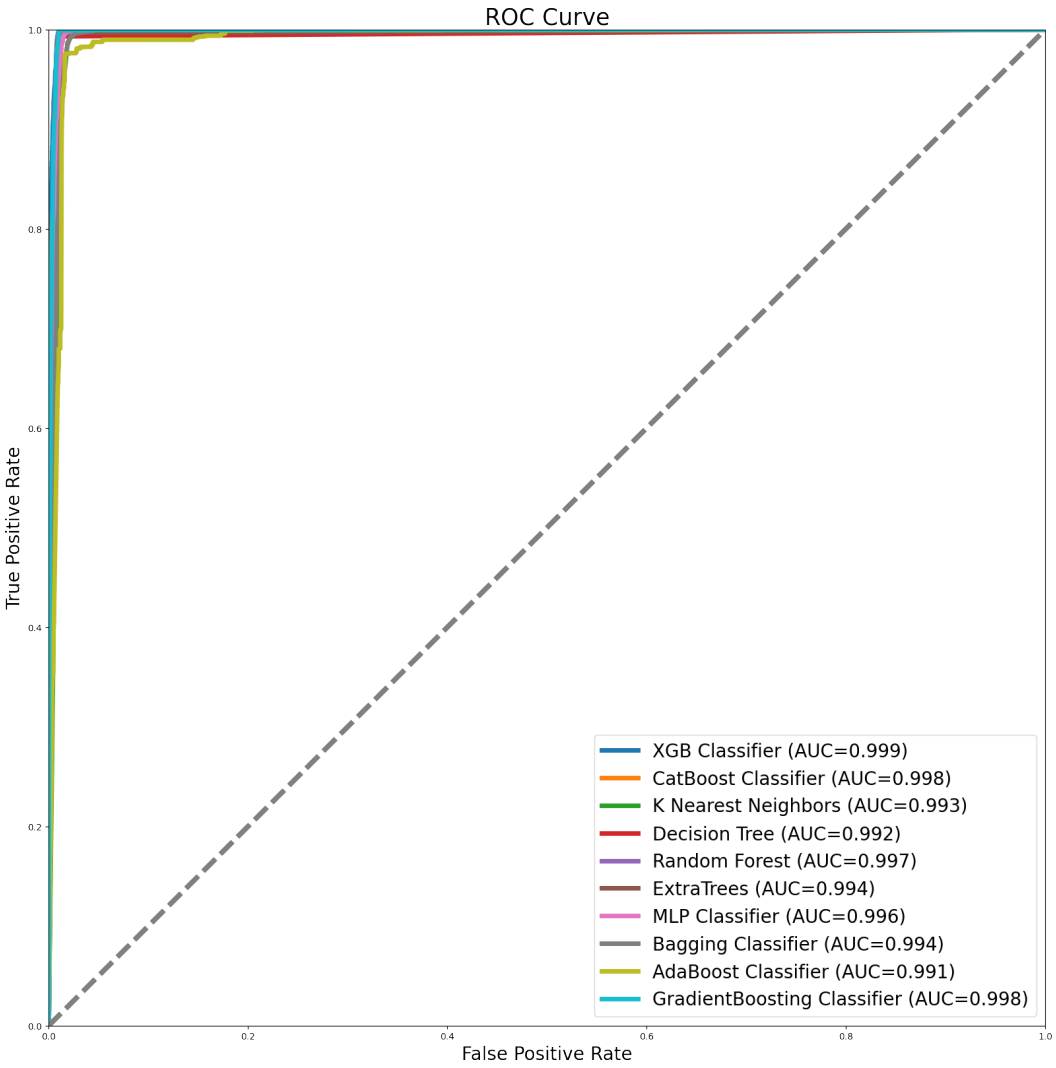

Figure 5.2: ROC

However, we notice that the results of AdaBoost always performed worse in oversampling and undersampling these two situations. The reason might be Adaboost (Adaptive Boost) emphasizes adaptive, by constantly modifying the sample weights increasing the weight of misclassified samples and decreasing the weight of misclassified samples, and constantly adding weak classifiers for boosting. Therefore, the disadvantage of AdaBoost is that it is sensitive to anomalous samples, which may receive higher weights in the iterations and affect the prediction accuracy of the

final strong learner.

Although the random forest model ranks first in terms of accuracy results, other boosting models are not inferior to it. The reason why XGBoost has better results in boosting models is that GBDT only uses first-order Taylor expansions, while XGBoost performs second-order Taylor expansions on the loss function. . Secondly, XGBoost borrows from random forests and supports column sampling, which not only reduces overfitting, but also reduces computation. This is also a feature that makes XGBoost different from traditional GBDT.

Chapter 6

Conclusion

6.1 Summary

This article firstly discusses that the thyroid gland releases thyroid hormones that regulate the body's metabolism and therefore if there is a problem with the thyroid gland it can have a huge impact on the body, such as affecting fertility, motor coordination and temperature sensitivity. Secondly, thyroid disease is a very common endocrine disorder with a very high prevalence, especially in women. In our data set, the ratio of women to men is 78.36% to 21.64%, which is due to the inextricable relationship between thyroid hormones and sex hormones. In addition to this, the prevalence of thyroid disorders is high at every stage, but is particularly concentrated in middle-aged and elderly people over 50%. However, the main focus of this paper is on the use of ten different machine learning algorithms, which are divided into three main categories, boosting techniques are Gradient Boosting, XGBoost, CatBoost and AdaBoost. Another category is bagging classifiers, including Random Forests, Decision Trees, Extra Trees and Bagging classifiers. In addition, there are KNN and MLP classifiers, which are used to compare with other ensembling learning methods. Since we analyzed the data distribution and found that there was a serious imbalance in the data we used. We adopted two solutions, namely, oversampling and undersampling. The former mainly using the smote technique to reach the requirement of boosting the number of samples, and the latter reducing the number of samples until all three classes were flush. Finally, we find that the overall accuracy of the models trained with the smote technique is higher than the other two. The most accurate model is the Random Forest model with 98.87% F1-socre, and the score of other bagging techniques such as Extra trees and Decision Tree are 98.51% and 98.45%. While the boosting models such as XGBoost, CatBoost, and Gradient Boost all reach 98.79%, 98.54%, and 98.44% respectively. These results are also quite outstanding. Therefore, the accuracy of bagging and boosting in our study are all quite well. This proves that these traditional machine learning algorithms can achieve more effective results in thyroid disease classification, and thus AI may soon become a very excellent tool in the medical field to assist doctors in thyroid disease diagnosis.

6.2 Limitations and Future Research

Our study is unique in that it has a more comprehensive dataset than the existing literature, where other datasets have only 215 quantities, while my dataset has a total amount of data reaching 20k after removing some items containing null values. As a second point, the existing literature analyzes classical algorithms and their role in thyroid disease diagnosis classification, but most of them are not compared together to determine which algorithms are unfavorable for thyroid disease diagnosis and which ones are highly accurate. In our study, this issue has been solved.

However, our dataset also has a serious imbalance problem, with a 90:7:3 ratio for the three categories. Algorithm bias may result from this. Another major cause of bias in the algorithm is that different races, habits, and regions can have different thyroid hormone levels. The first step in tackling these problems is to collect more detailed data sets, including patient lifestyle, changes in other sex hormone indicators when a woman becomes pregnant, thyroid surgery, thyroid hormone changes before and after surgery, and thyroid-related data from patients around the world. Our models for classifying thyroid disorders can not only be improved with more comprehensive data to make them more robust but also be trained to predict thyroid function more accurately for patients in different regions or based on specific physiological conditions (e.g., pregnancy).

References

1. Sapin, R. & Schlienger, J. Thyroxine (T4) and tri-iodothyronine (T3) determinations: Techniques and value in the assessment of thyroid function in *Annales de Biologie Clinique* **61** (2003), 411–420.
2. Duntas, L. H., Orgiazzi, J. & Brabant, G. The interface between thyroid and diabetes mellitus. *Clinical Endocrinology* **75,** 1–9 (2011).
3. Singh, G., Gupta, V., Sharma, A. K. & Gupta, N. Evaluation of thyroid dysfunction among type 2 diabetic Punjabi population. *Advanced Biomedical Research* **2,** 3–9 (2011).
4. Temurtas, F. A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications* **36,** 944–949 (2009).
5. Melish, J. S. Thyroid Disease. *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition* (1990).
6. Pal, R., Anand, T. & Dubey, S. K. Evaluation and performance analysis of classification techniques for thyroid detection. *International Journal of Business Information Systems* **28,** 163–177 (2018).
7. Li, D., Yang, D., Zhang, J. & Zhang, X. Ar-ann: Incorporating association rule mining in artificial neural network for thyroid disease knowledge discovery and diagnosis. *IAENG International Journal of Computer Science* **47,** 25–36 (2020).
8. Keleş, A. & Keleş, A. ESTDD: Expert system for thyroid diseases diagnosis. *Expert Systems with Applications* **34,** 242–246 (2008).
9. Boelaert, K. *et al.* Endocrinology in the time of COVID-19: Management of hyperthyroidism and hypothyroidism. *European Journal of Endocrinology* **183,** G33–G39 (2020).
10. Prerana, P. S. & Taneja, K. Predictive data mining for diagnosis of thyroid disease using neural network. *International Journal of Research in Management, Science & Technology* **3,** 75–80 (2015).
11. Neki, N. & Kazal, H. Solitary Thyroid Nodule-An Insight. *J Ind Acad Clin Med* **7,** 328–3 (2006).
12. Acharya, U. R. *et al.* Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images. *Knowledge-Based Systems* **107,** 235–245 (2016).

13. Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T. & Mukherjee, S. A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. *CSI Transactions on ICT* **4,** 313–319 (2016).

14. Azar, A. T., Hassanien, A. E., Kim, T.-H., *et al.* Expert System Based on Neural-Fuzzy Rules for Thyroid Diseases Diagnosis, 94–105 (2012).

15. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* **13,** 8–17 (2015).

16. Razia, S., Siva Kumar, P. & Rao, A. S. Machine learning techniques for thyroid disease diagnosis: a systematic review. *Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough,* 203–212 (2020).

17. Aswad, S. A. & Sonuç, E. Classification of VPN network traffic flow using time related features on Apache Spark in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (2020), 1–8.

18. Shahid, A. H. *et al.* A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques in *2019 International Conference on Communication and Electronics Systems (ICCES)* (2019), 930–933.

19. Nishat, M. M. *et al.* Performance investigation of different boosting algorithms in predicting chronic kidney disease in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)* (2020), 1–5.

20. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data to Health Care. *JAMA* **309,** 1351–1352 (2013).

21. Mullur, R., Liu, Y.-Y. & Brent, G. A. Thyroid Hormone Regulation of Metabolism. *Physiological Reviews* (2014).

22. McAninch, E. A. & Bianco, A. C. Thyroid hormone signaling in energy homeostasis and energy metabolism. *Annals of the New York Academy of Sciences* **1311,** 77–87 (2014).

23. Taylor, P. N. *et al.* Global epidemiology of hyperthyroidism and hypothyroidism. *Nature Reviews Endocrinology* **14,** 301–316 (2018).

24. Fitzgerald, S. P. & Bean, N. G. The Relationship between Population T4/TSH Set Point Data and T4/TSH Physiology. *Journal of Thyroid Research* **2016** (2016).

25. Karaca, N. & Akpak, Y. K. Thyroid disorders and fertility. *Int J Res Med Sci* **3,** 1299–304 (2015).

26. Roberta, M. *et al.* Thyroid and celiac disease in pediatric age: a literature review. *Acta Bio Medica: Atenei Parmensis* **89,** 11 (2018).

27. Mele, C. *et al.* Immunomodulatory effects of vitamin D in thyroid diseases. *Nutrients* **12,** 1444 (2020).

28. Dean, D. S. & Gharib, H. Epidemiology of thyroid nodules. *Best practice & research Clinical endocrinology & metabolism* **22,** 901–911 (2008).

29. Cabanillas, M. E., McFadden, D. G. & Durante, C. Thyroid cancer. *The Lancet* **388,** 2783–2795 (2016).

30. Davies, L., Roman, B. R., Fukushima, M., Ito, Y. & Miyauchi, A. Patient experience of thyroid cancer active surveillance in Japan. *JAMA Otolaryngology–Head & Neck Surgery* **145,** 363–370 (2019).

31. Chang, W.-W., Yeh, W.-C. & Huang, P.-C. A hybrid immune-estimation distribution of algorithm for mining thyroid gland data. *Expert Systems with Applications* **37,** 2066–2071 (2010).

32. Maysanjaya, I. M. D., Nugroho, H. A. & Setiawan, N. A. A comparison of classification methods on diagnosis of thyroid diseases in *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)* (2015), 89–92.

33. Saiti, F., Naini, A. A., Shoorehdeli, M. A. & Teshnehlab, M. Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM in *2009 3rd International Conference on Bioinformatics and Biomedical Engineering* (2009), 1–4.

34. Sonuç, E. *et al.* Thyroid Disease Classification Using Machine Learning Algorithms in *Journal of Physics: Conference Series* **1963** (2021), 012140.

35. Mitrou, P., Raptis, S. A. & Dimitriadis, G. Thyroid disease in older people. *Maturitas* **70,** 5–9 (2011).

36. Baksi, S. & Pradhan, A. Thyroid hormone: sex-dependent role in nervous system regulation and disease. *Biology of Sex Differences* **12,** 1–13 (2021).

37. Stoll, S. J. *et al.* Thyroid hormone replacement after thyroid lobectomy. *Surgery* **146,** 554–560 (2009).

38. Verma, A. & Mehta, S. A comparative study of ensemble learning methods for classification in bioinformatics in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence* (2017), 155–158.

39. Prasad, A. M., Iverson, L. R. & Liaw, A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9,** 181–199 (2006).

40. Shahid, A. H. & Singh, M. Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments. *Biocybernetics and Biomedical Engineering* **39,** 638–672 (2019).

41. Arfiani, A. & Rustam, Z. Ovarian cancer data classification using bagging and random forest in *AIP Conference Proceedings* **2168** (2019), 020046.

42. Ali, J., Khan, R., Ahmad, N. & Maqsood, I. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)* **9,** 272 (2012).

43. Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications* **134,** 93–101 (2019).

44. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine Learning* **63,** 3–42 (2006).

45. Fan, Z., Xu, F., Li, C. & Yao, L. Application of KPCA and AdaBoost algorithm in classification of functional magnetic resonance imaging of Alzheimer's disease. *Neural Computing and Applications* **32,** 5329–5338 (2020).

46. Abdurrahman, G. & Sintawati, M. Implementation of xgboost for classification of parkinson's disease in *Journal of Physics: Conference Series* **1538** (2020), 012024.

47. Dorogush, A., Gulin, A., Gusev, G., Ostroumova, L. & Vorobev, A. Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516* (2017).

48. Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. KNN model-based approach in classification in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (2003), 986–996.

49. Pacheco, W. D. N. & López, F. R. J. Tomato classification according to organoleptic maturity (coloration) using machine learning algorithms K-NN, MLP, and K-Means Clustering, 1–5 (2019).

50. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16,** 321–357 (2002).

51. Rahmad, F., Suryanto, Y. & Ramli, K. Performance comparison of anti-spam technology using confusion matrix classification in *IOP Conference Series: Materials Science and Engineering* **879** (2020), 012076.

52. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine* **4,** 627 (2013).