Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755230

THYROID DISEASE CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Chatlapally Shravan Kumar
PG Scholar, Department of Computer Science & Engineering, Holy Mary Institute of Technology & Science (Autonomous), Hyderabad, India.

Mr. Ravi Kumar Banoth
Associate Professor, Department of Computer Science & Engineering, Holy Mary Institute of Technology & Science (Autonomous), Hyderabad, India.

ABSTRACT

Machine learning algorithms and data mining techniques play a significant role in handling data due to the overwhelming volume of information that can be challenging to handle, particularly in the health system. We employed machine learning algorithms with thyroid illness in our investigation. A thorough project aimed at applying machine learning algorithms to categorize thyroid illness based on patient data is the Thyroid illness Classification Project using Machine Learning. The research incorporates a dataset with different variables associated with patient characteristics and thyroid function testing. The method attempts to precisely categorize individuals into several thyroid disease categories, such as hypothyroidism, hyperthyroidism, or euthyroidism, by utilizing supervised machine learning algorithms. We used data from Iraqi people, some of whom have an overactive thyroid gland and others who have hypothyroidism, to work toward the objective of classifying thyroid illness into three categories: hyperthyroidism, hypothyroidism, and normal. All of the algorithms were employed in this study. Support vector machines, k-nearest neighbors, logistic regression, random forest, decision trees, naïve bayes, multilayer perceptron (MLP), and linear discriminant analysis. to categorize thyroid conditions. The thyroid gland is an essential hormone gland that affects the body's growth, development, and metabolism. By continuously delivering a regular amount of thyroid hormones into the circulation, it aids in the regulation of several bodily activities. Numerous follicular cells found in the thyroid gland store the thyroid hormones inside the thyroglobulin molecule until the body requires them. Almost all of the body's cells are impacted by the thyroid hormones, also known as the primary metabolic hormones. The hypothalamic-pituitary-thyroid system's development and the availability of tyrosine and iodine are prerequisites for the synthesis and release of thyroid hormones. Interruption of its development, as occurs with preterm birth, results in insufficient synthesis of thyroid-stimulating hormone and thyroxin, leading to a range of physiologic disorders. When there is an issue with the thyroid gland or inadequate synthesis of thyroid hormone, pathologic problems arise. Laboratory testing are helpful in detecting disorders of the thyroid gland. Certain thyroid problems can also be diagnosed with the use of a comprehensive history, clinical signs, and radiologic results. In addition to diagnosing and treating thyroid issues, nurses are crucial in offering newborns and their families supportive care. By offering a tool for the early diagnosis and categorization of thyroid problems, this research advances medical diagnostics and helps healthcare providers make well-informed decisions.

Keyword: knn (k-nearest neighbors), Logistic Regression, Support Vector Machine, Decision Tree, Random Forest

1. INTRODUCTION

Thyroid gland problems are among the most widespread endocrine disorders in the world, second only to diabetes, according to the World Health Organization. Hyper function hyperthyroidism and hypothyroidism afflict roughly 2% and 1% of persons, respectively. Men have roughly a tenth of the frequency of women. Hyper-and hypothyroidism may be induced by thyroid gland dysfunction, secondary to pituitary gland failure, or tertiary to hypothalamus malfunction. Due to dietary iodine shortage, goitre or active thyroid nodules may become prominent in some places, with a prevalence of up to 15%. The thyroid gland may also be the location of different kinds of tumors and can be a hazardous region where endogenous antibodies wreak havoc (autoantibodies). The thyroid gland is a butterfly-shaped gland found at the base of the neck. It consists two active thyroid hormones, levothyroxine and triiodothyronine, which are involved in brain activities such as body temperature control, blood pressure management, and heart rate regulation. Early illness identification, diagnosis, and care, according to experts, are crucial in preventing disease progression and even death. For various distinct sorts of abnormalities, early detection and differential diagnosis enhances the probability of effective therapy. Thyroid Disease is one of the most widespread disorders globally, and it is mainly caused by a shortage of iodine, although it may also be caused by other causes. The thyroid gland is an endocrine gland that secretes hormones and transfers them via the circulation. It is positioned in the center of the front of the body. Thyroid gland hormones are crucial for assisting in digestion as well as maintaining the body wet, balanced, and so on. Thyroid gland therapies such as T3 (triiodothyronine), T4 (thyroid hormone), and TSH (thyroid stimulating hormone) are used to measure thyroid activity (thyroid stimulating hormone). Thyroid condition is split into two types: hypothyroidism and hyperthyroidism. Hyperthyroidism is a disease in which the thyroid gland releases so many thyroid hormones. Hyperthyroidism is caused by a rise in thyroid hormone levels .Dry skin, higher temperature sensitivity, hair thinning, weight loss, increased heart rate, high blood pressure, excessive perspiration, neck enlargement, anxiousness, menstrual periods shortening, irregular stomach motions, and hands trembling are some of the indications .Hypothyroidism is a disorder in which the thyroid gland is underactive Hypothyroidism is characterized by a reduction in thyroid hormone production. Hypo denotes insufficient or less in medical words. Inflammation and thyroid gland damage are the two basic causes of hypothyroidism. Obesity, low heart rate, increased temperature sensitivity, neck swelling, dry skin, hand numbness, hair troubles, heavy menstrual periods, and digestive problems are some of the symptoms. If not addressed, these symptoms might increase over time. Machine learning algorithms are one of the finest answers to many problems that are tough to address Classification is a data extraction technique (machine learning) used to predict and identify many diseases, such as thyroid disease, which we researched and classified here because machine learning algorithms play a significant role in classifying thyroid disease and because these algorithms are high performing and efficient and aid in classification. Although the application of computer learning and artificial intelligence in medicine stretches back to the early days of the profession, there has been a fresh trend to address the need for machine learning-driven healthcare solutions. As a result, analysts expect that machine learning will become mainstream in healthcare in the near future. With recent technical breakthroughs in data processing and computation, machine learning and deep learning techniques have been applied in various research projects for thyroid illness prediction. Prediction of this disease in its early stages and its categorization as cancer, Hypothyroidism, or Hyperthyroidism is useful for prompt treatment and recovery. The literature survey is undertaken utilizing peer-reviewed article databases such as Google scholar and Scopus. The searches were done within the span of the previous five years to identify the recent works in our study. The keywords "Thyroid disease", "Thyroid cancer", "machine learning", and "deep learning" combinations were used to find the relevant articles. As the amount of returned results is significantly larger for discovering the relevant articles, we have further adjusted the search queries and utilized a tight keyword search. Overall, more than 100 relevant publications were recognized during our first scan. We further evaluated those articles and picked 25 articles that are directly relevant to our work. Machine learning and deep learning approaches are employed both for thyroid illness identification and thyroid cancer detection. As the procedure of implementing these strategies is different for each objectives, they are explained separately.

2. LITERATURE SURVEY

Rasitha Banu, G. [1] One of the most prevalent ailments affecting people is thyroid disease. The University of California, Irvine (UCI) data repository provided the hypothyroid data utilized in this investigation. Throughout the whole study project, the Waikato Environment of Information Analysis platform (WEKA) will be utilized. The decision stump tree approach was shown to be less successful than the J48 strategy. In the medical field, diagnosing diseases is a challenging task. A multitude of data mining techniques are employed in the decision-making process. In this investigation, we defined hypothyroidism using J48 and decision stump data mining classification approaches, and we utilized dimensionality reduction to choose a subset of variables from the original findings. The precision and error rate of the classifier output are evaluated using the uncertainty matrix. In addition to having a lower mistake rate than Decision stump, the J48 Algorithm has an accuracy of 99.58 percent, which is greater than Decision stump tree accuracy.

Rafi Ahmad Khan, Dr. Syed Mutahar Aaqib, and Umar Sidiq [2] Classification is a widely used supervised learning data mining approach that is employed to categorize preset data sets. The categorization is frequently used in the healthcare industry to support medical diagnosis, administration, and decision-making. Data for this research was obtained from a reputable laboratory in Kashmir. The ANACONDA3-5.2.0 platform will be used for the whole research study. Classification techniques like k closest neighbors, support vector machines, decision trees, and Nave Bayes may be applied in an experimental investigation. With an accuracy of 98.89 percent, the Judgment Tree is the most accurate of the other classes.

Mrs. K. Sindhya [3] Globally, thyroid dysfunction is a chronic ailment that affects people. Healthcare data mining is yielding outstanding outcomes in the prognosis of many illnesses. Data mining techniques have a high prediction accuracy and a cheap prediction cost. The fact that prediction requires relatively little time is another important advantage. In this work, I analyzed thyroid data using classification algorithms and produced a conclusion. There are two main criteria that influence a model's effectiveness. forecast time comes in second, and forecast precision comes in first. Our results show that Nave Bayes forecasted in under 0.04 seconds. It is not as accurate as Random Forest and J48, though. In terms of prediction accuracy, 99.3 percent was achieved with the Random Forest model. The model took longer to assemble than the previous two rounds, though. Since J48's accuracy is among the greatest at 99 percent and it takes 0.2 seconds to run—much less time than the Random Forest model—we may infer that it is the best model for hypothyroid prediction.

Göksu, AKGÜL, et al. [4] The purpose of this research is to provide a data mining-based approach that combines test findings with patient queries to improve the accuracy of diagnosing hypothyroidism. Minimizing the dangers associated with dialysis interventional studies is another objective. The sensible outcome Data from the UCI machine learning database was used to evaluate if the new samples were hypothyroid. Of the 3163 samples, 151 were hypothyroid, and the remaining samples were hypothyroid. To eliminate the imbalanced distribution, several sampling approaches were employed throughout the data gathering process. Logistic regression, K-Nearest Neighbor, and Support Vector Machine classifiers were utilized in the development of models designed to detect hypothyroidism. In this sense, the thesis illustrated how sampling strategies affect the diagnosis of hypothyroidism. Out of all the models developed, the Logistic Regression classifier yielded the best results. For this study, which was trained on the data set using over-sampling approaches, the accuracy was 97.8%, the F-Score was 82.26 percent, the region under the curve was 93.2 percent, and the Matthews correlation coefficient was 81.8 percent.

Kumar, K. Vijiya, et al. [5] The purpose of this work is to develop a machine learning strategy that uses the Random Forest algorithm to predict diabetes in a patient in an early and accurate manner. A popular kind of ensemble learning system for classification and regression applications is the Random Forest method. The performance ratio is greater when compared to other methods. The findings showed that the prediction system is capable of accurately, effectively, and most significantly, promptly forecasting diabetes illness. The recommended model yields the best outcomes for diabetic prediction.

3. EXISTING SYSTEM

Every Segmentation is based on a search. Gathering the data and segmenting it appropriately is difficult, and the results are not very reliable since the clustering is not near enough to identify precise centroids.

4. PROPOSED SYSTEM

Create a system to obtain simple visualization methods. To get more accurate findings, expand the data collection to include more data points. Directly divide the merchandise into consumer segments. Instead than using paper forms, gather client data using alternative techniques.

5. SYSTEM ARCHITECTURE

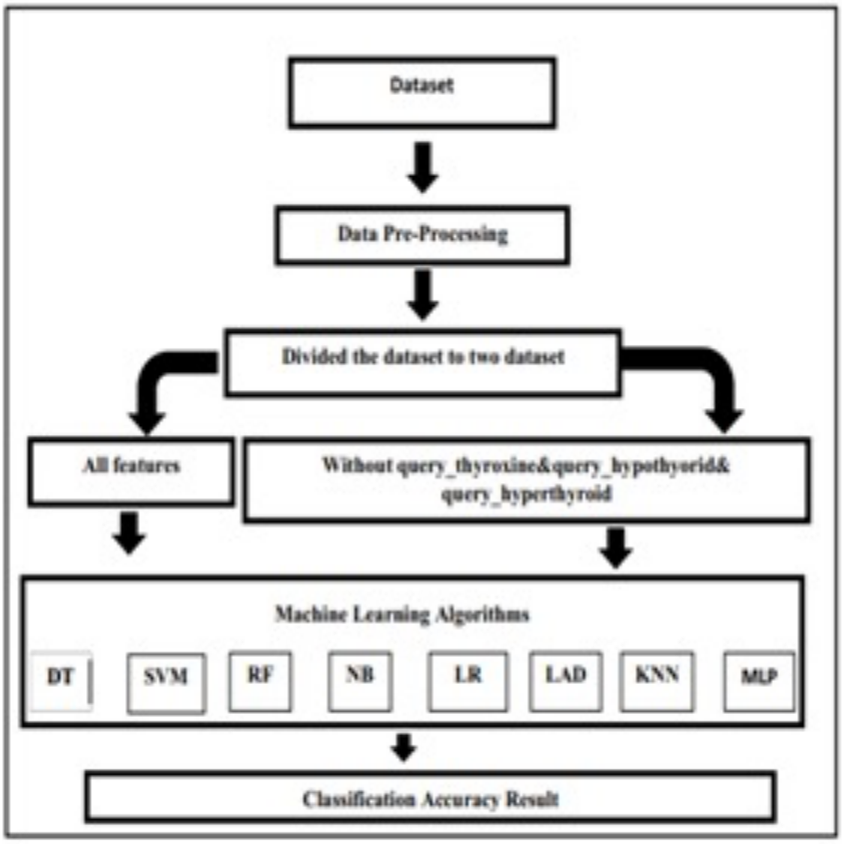

Dataset
Data Pre-Processing
Divided the dataset to two dataset
All features
Without query_thyroxine&query_hypothyorid& query_hyperthyroid
Machine Learning Algorithms
DT
SVM
RF
NB
LR
LAD
KNN
MLP
Classification Accuracy Result

Fig.5.1 architecture

6. METHODOLOGY

6.1 Data collection

Since machine learning algorithms are becoming increasingly important in the medical profession and aid in the diagnosis and classification of illnesses, they are utilized in the quick and early identification of thyroid disorders as well as other illnesses. because of this, we were able to gather a sizable quantity of information on thyroid disorders, and we are currently using this information to work on our study on the categorization of illnesses. The data that I used for our study came from outside hospitals and laboratories that specialize in analyzing and diagnosing illnesses. The sample of data that I took from these sources was the data pertaining to thyroid disease, which included 1250 male and female subjects whose ages ranged from one to one year. ninety years, since these samples include both healthy individuals without thyroid illness and individuals with thyroid disease who experience hyper- and hypothyroidism. The primary objective of the data collection, which took place over the course of one to four months, was to categorize thyroid disorders using machine learning algorithms. Gender, age, T3 (triiodothyronine), T4 (thyroid hormone), TSH (thyroid stimulating hormone), and a variety of other variables are included in this data. Since all 17 variables or attributes—i.e., id, age, gender, query thyroxin,

on_antithyroid_medication, ill, pregnant, thyroid surgery, query_hypothyorid, query hyperthyroid, TSH_M, TSH, T3_M, T3, T4, Category—were collected for our study, they make up the data gathered.

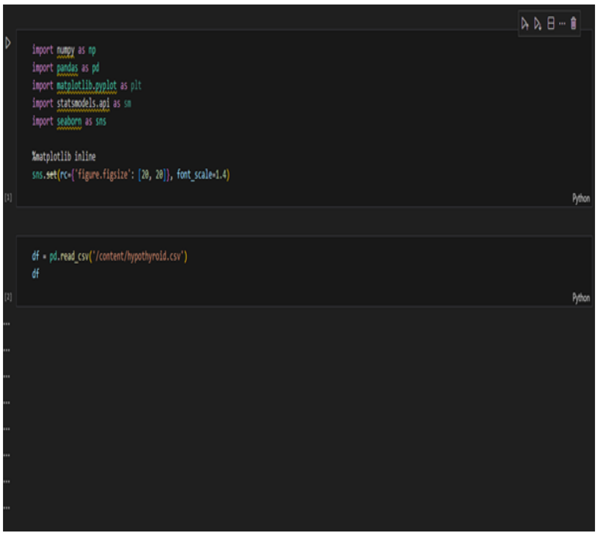
6.2 Data preprocessing

Preparing the data is a crucial phase in the data mining process and has a positive impact on the data itself. It is used to uncover information by carefully examining and analyzing the data, perhaps uncovering lost information. Data preparation, cleansing, and other tasks are part of the pre-processing procedure. We first cleaned and organized the data we were able to obtain. We then found a set of missing data in which the missing features were identified. Among these properties, T4 by number 151 and T3 by number 112. We were able to process this lost data by substituting the mediator's value, and after doing this, we were able to obtain the data in a better and more efficient manner, free from lost data. The data was also arranged, good, and free from any defects or issues, allowing us to work on it efficiently and effectively. We also employed the MLP algorithm using normalizing techniques.

6.3 Data machine learning techniques

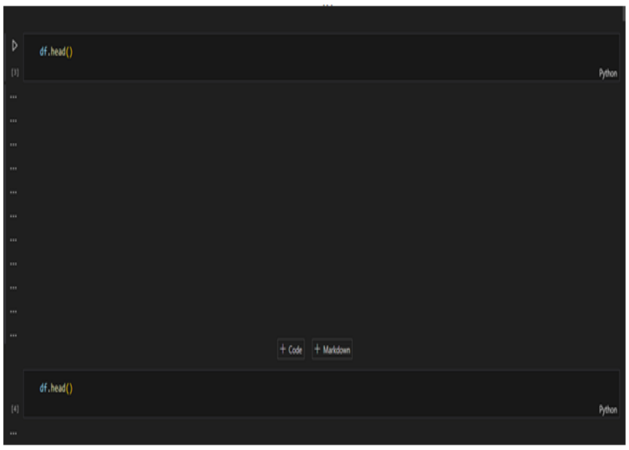
The ability to distinguish between the three types of thyroid illness is the main goal of applying machine learning algorithms. The first three conditions are hyperthyroidism, hypothyroidism, and stable individuals without thyroid problems.

7. OUTPUT SCREENS



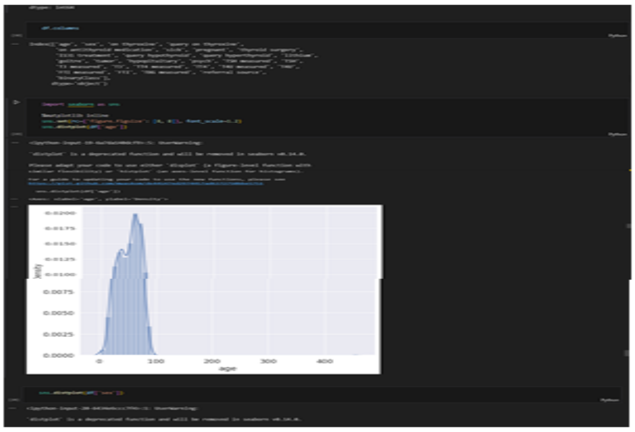
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import seaborn as sns

%matplotlib inline
sns.set(rc={'figure.figsize': [20, 20]}, font_scale=1.4)
```

```
df = pd.read_csv('/content/hypothyroid.csv')
df
```



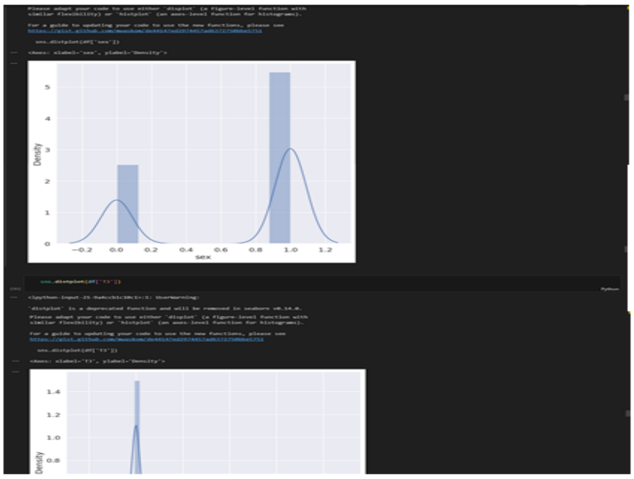
```
df.head()
```

```
df.head()
```



```
dtype: int64

df.columns

Index(['age', 'sex', 'on thyroxine', 'query on thyroxine',
       'on antithyroid medication', 'sick', 'pregnant', 'thyroid surgery',
       'I131 treatment', 'query hypothyroid', 'query hyperthyroid', 'lithium',
       'goitre', 'tumor', 'hypopituitary', 'psych', 'TSH measured', 'TSH',
       'T3 measured', 'T3', 'TT4 measured', 'TT4', 'T4U measured', 'T4U',
       'FTI measured', 'FTI', 'TBG measured', 'referral source',
       'binaryClass'],
      dtype='object')

import seaborn as sns

%matplotlib inline
sns.set(rc={'figure.figsize': (8, 6)}, font_scale=1.2)
sns.distplot(df['age'])

<ipython-input-19-[illegible]>:5: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['age'])

<Axes: xlabel='age', ylabel='Density'>
```

```
sns.distplot(df['sex'])

<ipython-input-20-[illegible]>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.
```



```
Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['sex'])

<Axes: xlabel='sex', ylabel='Density'>
```

```
sns.distplot(df['T3'])

<ipython-input-21-[illegible]>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df['T3'])

<Axes: xlabel='T3', ylabel='Density'>
```

8. 8. CONCLUSION

The goal of the machine learning-based Thyroid Disease Classification Project is to meet the urgent demand for a precise and prompt diagnosis of thyroid illnesses. The project aims to enhance the efficiency of thyroid illness detection and make a contribution to the area of medical informatics by investigating sophisticated algorithms, extensive datasets, and cutting-edge technologies. An extensive and heterogeneous dataset including patient demographics, thyroid hormone levels, and clinical aspects was examined. To guarantee the accuracy and applicability of the data, thorough preprocessing procedures were carried out, such as feature selection, normalization, and management of missing values. Numerous machine learning algorithms were taken into consideration, spanning from sophisticated ensemble techniques and neural networks to more conventional decision trees. The method that showed the best accuracy, precision, recall, and overall robustness for classifying thyroid diseases was selected by assessing each algorithm's performance. The Thyroid Disease Classification Project is a major advancement in the application of machine learning to improve healthcare. The model's effective creation and assessment highlight the potential for technologically advanced solutions to improve medical diagnosis, paving the way for future advancements at the nexus of machine learning and healthcare. The project lays the groundwork for further improvements and expansions, such as the integration of new data sources, modification to meet changing medical standards, and ongoing development based on input from medical experts.

REFERENCES

[1] Spline Interpolation, pages 141–173. Springer New York, New York, NY, 2006.

[2] Izdihar Al-muwaffaq and Zeki Bozkus. Mltdd : Use of machine learning techniques for diagnosis of thyroid gland disorder. 2016.

[3] A. Begum and A. Parkavi. Prediction of thyroid disease using data mining techniques. In 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), pages 342–345, 2019.

[4] Mario Luca Bernardi, Marta Cimitile, Fabio Martinelli, and Francesco Mercaldo. Keystroke analysis for user identification using deep neural networks. In International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019, pages 1–8. IEEE, 2019. [5] G. Biau, B. Cadre, and L. Rouviere. Accelerated gradient boosting. ` Machine Learning, 108(6):971–992, 2019.

[6] Shiva Borzouei, Hossein Mahjub, NegarAsaad Sajadi, and Maryam Farhadian. Diagnosing thyroid disorders: Comparison of logistic regression and neural network models. Journal of Family Medicine and Primary Care, 9:1470, 06 2020.

[7] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. J.Artif. Int. Res., 16(1):321357, June 2002.

[9] Essam Al Daoud. Comparison between xgboost, lightgbm and catboost using a home credit dataset. International Journal of Computer and Information Engineering, 13(1):6 – 10, 2019.

[10] Leonidas H. Duntas and Jacqueline Jonklaas. Levothyroxine dose adjustment to optimise therapy throughout a patient’s lifetime. Advances in Therapy, 36(2):30–46, 2019.

[11] Yoav Freund and Robert E Schapire A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[12] T. Karaylan and . Kl. Prediction of heart disease using neural network. In 2017 International Conference on Computer Science and Engineering (UBMK), pages 719–723, 2017.