

REPORT
ON
FOUR WEEKS OF INTERNSHIP-II
HYPOTHYROID PREDICTION USING MACHINE LEARNING

Submitted to
NMAM INSTITUTE OF TECHNOLOGY, NITTE
(An Autonomous Institution under VTU, Belagavi)

In partial fulfilment of the requirements for the award of the
Degree of Bachelor of Engineering
in
Artificial Intelligence and Machine Learning Engineering

by
Chethana R Kini
USN 4NM21AI018

Under the guidance of
Mrs. Smitha
Assistant Professor –Gd II



N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

AY 2023 - 2024

CERTIFICATE

*This is to certify that the “Internship-II report” submitted by **Ms. Chethana R Kini** bearing USN **4NM21AI018** of V semester B.E., a bonafide student of NMAM Institute of Technology, Nitte, has undergone at least four weeks of internship at Invenger Technologies, Mangalore during 2022-2023 fulfilling the partial requirements for the award of degree of Bachelor of Engineering in Artificial Intelligence and Machine Learning at NMAM Institute of Technology, Nitte.*

Signature of the Guide

(Mrs. Smitha)

Signature of the Coordinator

(Mrs. Smitha)

Signature of the HOD

(Dr. Sharada U. Shenoy)

Signature of the Examiners:

1. _____

2. _____


TO WHOM IT MAY CONCERN

This is to certify that Ms. CHETHANA R KINI has completed internship programme on "Hypothyroid Prediction using Machine Learning" from 03.02.2023 to 03.03.2023.

She took keen interest in the work assigned and successfully completed it.

During the period of internship, we found her to be punctual, hardworking and inquisitive. We wish her luck and success in all her future endeavours.

For INVENGER TECHNOLOGIES PVT LTD



M.NARASIMHA MALLYA
GENERAL MANAGER

Invenger Technologies Pvt. Ltd.

Invenger Towers, Kottara, Mangaluru - 575 006. India. Phone: +91 824 4232777 Fax: +91 824 4232743
www.invenger.in

ACKNOWLEDGEMENT

I express my deep sense of gratitude to NMAM Institute of Technology that provided me an opportunity in fulfilling my most cherished desire of reaching the goal.

I would like to give a sincere thanks to our beloved principal, Dr. Niranjan N. Chiplunkar for giving me an opportunity to carry out my internship work and providing me with all the needed facilities.

I acknowledge the support and valuable inputs given by, Dr. Sharada U. Shenoy the Head of the Department, Artificial Intelligence and Machine Learning Engineering, NMAMIT, Nitte.

I would also like to thank Invenger Technologies, Mangalore and Manipal Institute of Technology for giving me an opportunity to be a part of this internship.

I would like to extend my gratitude and indebtedness to my guides Dr. Srikanth Prabhu, Dr. Shyam Karanth and Dr. Krishnaraj Chadaga for guiding me throughout this internship.

Finally, I thank all the faculty and staff members of the Department of Artificial Intelligence and Machine Learning for their honest opinions and suggestions throughout the course of my internship.

Chethana R Kini

4NM21AI018

TABLE OF CONTENTS

Title	Page No.
Institute Certificate	(i)
Industry Certificate	(ii)
Acknowledgement	(iii)
Table of Contents	(iv)
Abstract	1
Introduction to the Industry/Research Institute	2
Details of the training undergone	3-13
Weekly Progress report	14
Conclusion	15
References	16

ABSTRACT

A widespread endocrine condition that affects a lot of people globally is hypothyroidism. To avoid serious problems like cardiovascular disease, osteoporosis, and mental health issues, early diagnosis and effective care of hypothyroidism are crucial. We describe the findings of our machine learning investigation on a hypothyroid dataset in this paper. The collection includes clinical and laboratory information about hypothyroidism patients. To create a model that can precisely identify hypothyroidism, we used a variety of machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines.

By dealing with missing values, category variables, and scaling the numerical variables, we pre-processed the dataset. To create a model that can accurately diagnose hypothyroidism, we deployed five machine learning methods namely Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbor.

Our findings demonstrated that the accuracy, precision, recall, and F1-score of the decision tree method were all 99.88%, 99.6%, 99.89 and 99.3%, respectively. An accuracy of 93.22%, precision of 95.3%, recall of 91.2%, and F1-score of 93.2% were attained by the K Nearest Neighbor. An accuracy of 57.38%, precision of 91%, recall of 16%, and F1-score of 27% were attained by the Naïve Bayes. The accuracy, precision, recall, and F1-score of the support vector machine algorithm were all 98.7%, 99.8%, 98.68 and 99.7%, respectively. The accuracy, precision, recall, and F1-score of the Random Forest were 99.8%, 99.85%, 99.69 and 99.85%, respectively.

Our research offers a practical method for swiftly and effectively identifying hypothyroidism. To guarantee the robustness and generalizability of the model, the findings should be tested on several datasets. Future research can also look into adding more genetic and clinical features to the model in order to increase its accuracy and dependability.

In conclusion, our study demonstrates that machine learning algorithms may create models that can quickly and reliably identify hypothyroidism. The data can be prepared for machine learning algorithms using the KNN imputer and other preparation methods. This has important implications for patient outcomes and healthcare expense savings.

INTRODUCTION OF THE INDUSTRY

Invenger's array of Technology services, IT consultancy, and outsourcing services is designed to upscale companies across multiple industries, address a range of customer operational challenges, and help to tap their undiscovered possibilities. It is classified as a private limited company and is located in Mangalore, Karnataka. Its authorized share capital is INR 5.00 crore and the total paid-up capital is INR 21.50 lakh. Invenger first laid out its vision in 2004, and with nearly two decades of operational rollout under its belt, Invenger has amassed significant expertise to support you in achieving your full potential and exceeding customer expectations. For its clients to comprehend and map various aspects of their technical and business demands, including planning, operations, and support, Invenger has proven to be reliable. Leveraging strong partnerships, Invenger has mapped its global presence across countries to serve its patrons the best. The IT company engages between 200 and 500 people. This industry offers business process outsourcing to boost efficiency, cloud computing, consulting to identify the ideal optimisation, technological development, and support services. Invenger Technologies offers a wide range of products and services including Computer and Mobile Softwares & Apps, Asset Management Software, Residential & Commercial Security, Cash Escort Services, Data Entry & Data Processing Service and Database Management.

Invenger Technology is a fast-growing IT services company that enhances its customer experience with a committed focus towards providing higher performance and better quality. Key to Invenger's market advantage is its integrated delivery model, which is a combination of high level of expertise in various businesses, technology, management areas and a customized implementation of best practices to meet every challenge that our clients face. Invenger is headquartered in California, USA and has offices in India. This competent global workforce helps the company to leverage and distribute the benefits of global resource utilization while providing excellent support to the customers round the clock.

DETAILS OF THE TRAINING UNDERGONE

Skills Obtained

Engaging in stroke prediction using machine learning and deep learning techniques equipped us with a range of valuable skills.

Data Analysis Skills: Ability to preprocess and clean diverse datasets, including medical records, imaging data, and clinical parameters. Proficiency in exploratory data analysis (EDA) to gain insights into the data and identify relevant patterns.

Feature Engineering: Skills in selecting and creating relevant features from raw data, which are crucial for training accurate machine learning models.

Machine Learning Skills: Understanding and application of various machine learning algorithms such as decision trees, random forests, support vector machines etc. for predictive modelling.

Deep Learning Skills: Understanding and application of Artificial Neural Network (ANN) for stroke prediction.

Practical Implementations

The integration of machine learning in hypothyroid prediction unfolds practical applications with direct relevance to healthcare advancement. Much like its counterpart in stroke prediction, the practical implementations for hypothyroidism comprise:

Timely Detection of Hypothyroidism: The central focus of deploying machine learning models is to achieve early detection of hypothyroidism. This aligns seamlessly with the objective of incorporating the model into routine health assessments and wellness programs, thereby facilitating the early identification of individuals at risk and enabling proactive management strategies.

Tailored Health Management Strategies: Machine learning algorithms prove instrumental in crafting personalized health management plans for individuals predisposed to hypothyroidism. These plans encompass a spectrum of customized interventions, ranging from precise medication adjustments to lifestyle modifications and routine monitoring. The individualized nature of these strategies ensures a targeted approach aligned with the unique needs of each patient.

Hospital-based Hypothyroid Risk Prediction: Within the confines of hospital settings, the utilization of machine learning models extends to predicting the risk of hypothyroidism among admitted patients. Through the analysis of diverse clinical parameters and historical data, healthcare providers can accurately identify patients at risk. This knowledge empowers healthcare professionals to implement precautionary measures during a patient's hospital stay, contributing to enhanced patient care and ultimately improving healthcare outcomes.

Problem Statement

Developing a machine learning model to accurately predict the presence of hypothyroid in individuals based on their medical data, with the aim of enabling early diagnosis and improving healthcare outcomes.

Objectives

A research study employing a hypothyroid dataset can have the following primary goals:

- To assess, using the available dataset, how well various machine learning algorithms perform in the diagnosis of hypothyroidism. In order to determine which algorithm performs the best, the research can compare the accuracy, precision, recall, and other performance measures of various algorithms.
- To look into the variables that determine how well machine learning algorithms can detect hypothyroidism. The study can investigate how the performance of the models
- The success of machine learning systems in early hypothyroidism detection hinges on variables like data quality, pre-processing methods, feature selection, and algorithm parameters.
- To determine whether machine learning systems can help with hypothyroidism early detection. The study can examine how well machine learning models can spot early indications of hypothyroidism and contrast them with conventional diagnostic techniques.

- To recognise the difficulties and restrictions associated with utilising machine learning techniques for hypothyroidism diagnosis. The study can investigate the moral, social, and technical issues surrounding the application of machine learning to medical diagnosis.
- To offer suggestions for enhancing the precision and dependability of machine learning algorithms for hypothyroidism diagnosis. The study can offer suggestions for methods to enhance data quality, optimise pre-processing procedures, choose pertinent features, and fine-tune algorithm parameters for improved performance.
- To show the potential advantages of employing machine learning algorithms for hypothyroidism diagnosis. The study can demonstrate how machine learning models can offer precise and effective diagnosis, cutting healthcare expenditures and increasing patient outcomes.

Methodology

There are certain standard steps to be followed for any machine learning project. Firstly, we will have to collect the data to be worked on. Next step will be, cleaning the data like removing noises values and outliers, handling imbalanced datasets, changing categorical variables to numerical values, etc. And to train the model we use various machine learning and deep learning algorithms. We use different metrics for model evaluation like recall, f1 score, accuracy, etc.

1. **Dataset** - We have chosen hypothyroid dataset for our study. The dataset consists of 29 columns with different features and 3772 readings. The binary class column gives the output of whether the person is suffering from hypothyroid or not. P in binary class indicates that the person is suffering from the disease, and N represents that the person is not suffering from the disease.
2. **Data Pre-processing** – Before we give the data for training a model, its necessary to check missing values, noise values etc for a higher accuracy. In this dataset, TBG and referral source columns are omitted since they are almost filled with inappropriate values. The dataset is then checked for missing values. Missing values are filled using KNNImputer and bfill methods.

Label encoding is something that converts the strings into a numeric form so as to convert them into the machine-readable form. And hence we use label encoder for all the features present in a string format. Also, this is quite an unbalanced dataset since the probabilities of person suffering and not suffering from hypothyroid disease is not the same. Hence, we use SMOTE technique to deal with this unbalanced data.

3. **Splitting training and testing data** - The train-test split is used to estimate the performance of machine learning algorithms. This process allows us to compare our own machine learning model results to machine results. Here, we have split test set into 25% of the actual data and train set into 75% of the actual data.
4. **Machine learning algorithms for prediction** – We have used machine learning and deep learning algorithms for the prediction of hypothyroid disease.

Decision tree - A decision tree is one of the supervised machine learning algorithms. This algorithm can be used for regression and classification problems — yet, is mostly used for classification problems. A decision tree follows a set of if-else conditions to visualize the data and classify it according to the conditions.

For example:

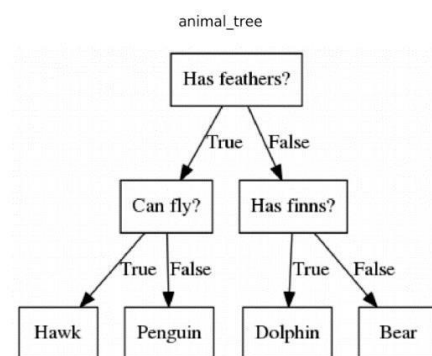


Fig I: Decision Tree

KNN – K-nearest neighbours algorithm is a non-parametric technique which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

This is the most commonly used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.

$$d(x,y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

Fig II: KNN

Logistic Regression - Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of certain classes based on some dependent variables. In short, the logistic regression model computes a sum of the input features and calculates the logistic of the result.

The output of logistic regression is always between (0, and 1), which is suitable for a binary classification task. The higher the value, the higher the probability that the current sample is classified as class=1, and vice versa.

Random forest - Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks.

Naïve Bayes - Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Hypothyroid Prediction

Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig III: Naïve Bayes

SVM - The Support Vector Machine algorithm is a popular supervised learning algorithms which aims to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

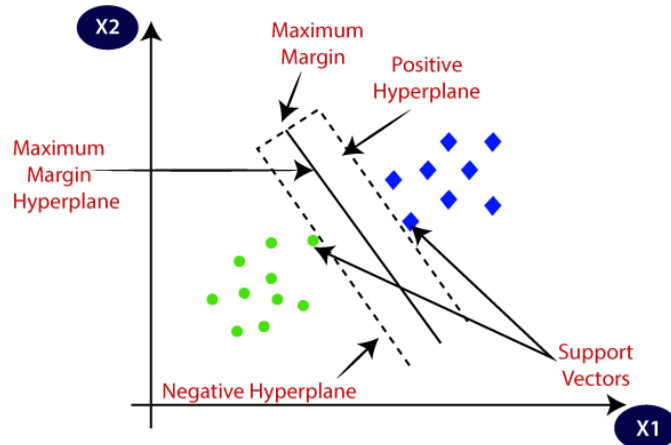


Fig IV: SVM

ANN - The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain Biological Neural Networks. Artificial neural networks use different layers of mathematical processing to make sense of the information it's fed.

Artificial Neural Networks work in a way similar to that of their biological inspiration. They can be considered as weighted directed graphs where the neurons could be compared to the nodes and the connection between two neurons as weighted edges. The processing element of a neuron receives many. Signals are sometimes modified at the receiving synapse and the weighted inputs are summed at the processing element. If it crosses the threshold, it goes as input to other neurons and the process repeats.

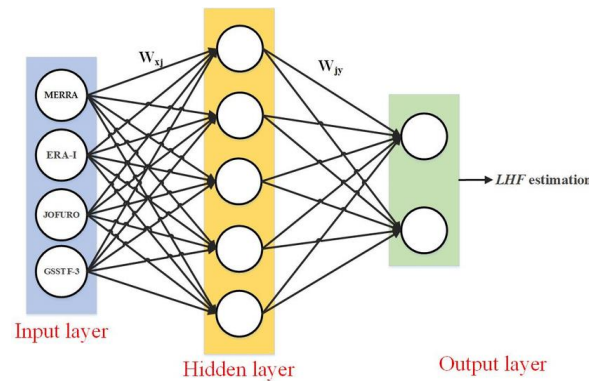


Fig V: ANN

Stacking - Stacking is one of the popular ensemble modelling techniques in machine learning. Various weak learners are ensembled in a parallel manner in such a way that by combining them with Meta learners, we can predict better predictions for the future. This ensemble technique works by applying input of combined multiple weak learners' predictions and Meta learners so that a better output prediction model can be achieved. In stacking, an algorithm takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction. Here, in this stacking model we used 6 classifiers – Decision tree, KNN, Logistic regression, SVM, Random forest and Naïve Bayes.

ROC and AUC curves – ROC curve, also known as Receiver Operating Characteristics Curve, is a metric used to measure the performance of a classifier model. The ROC curve depicts the rate of true positives with respect to the rate of false positives, therefore highlighting the sensitivity of the classifier model. The ROC is also known as a relative operating characteristic curve, as it is a comparison of two operating characteristics, the True Positive Rate and the False Positive Rate, as the criterion changes.

Area Under Curve or AUC is one of the most widely used metrics for model evaluation. It is generally used for binary classification problems. AUC measures the entire two-dimensional area present underneath the entire ROC curve. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than that of a randomly chosen negative example. The Area Under the Curve provides the ability for a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, it is assumed that the better the performance of the model at distinguishing between the positive and negative classes.

Heat map - A heat map represents these coefficients to visualize the strength of correlation among variables. It helps find features that are best for Machine Learning model building. The heat map transforms the correlation matrix into colour coding.

Confusion matrix – A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig IV: Confusion Matrix

Classification report - A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report.

The report shows the main classification metrics precision, recall and f1-score on a per-

class basis. The metrics are calculated by using true and false positives, true and false negatives. Positive and negative in this case are generic names for the predicted classes. There are four ways to check if the predictions are right or wrong:

TN / True Negative: when a case was negative and predicted negative

TP / True Positive: when a case was positive and predicted positive

FN / False Negative: when a case was positive but predicted negative

FP / False Positive: when a case was negative but predicted positive

Histogram Plot- A histogram plot is a particular kind of plot used in data visualisation that shows how a dataset is distributed. The histplot is another name for it. A histplot shows the frequency or count of data points falling within each bin as a bar or rectangle. The data is divided into a set of bins. The width of the bars matches the size of the bin, and the bins are typically evenly spaced apart and do not overlap. The x-axis of a histplot often depicts the range or values of the data, while the y-axis typically reflects the frequency or count of the data points lying within each bin.

Results

Six machine learning algorithms were tested in the study: Decision Trees, k-NN, Logistic Regression, Random Forest, Naïve Bayes and Support Vector Machines. A dataset of 3,772 instances, of which 3,481 were negative and 291 were positive for hypothyroidism, was used to train and evaluate the algorithms. The dataset was pre-processed using feature normalisation and imputation for missing values.

Table 1 displays the performance metrics for the six algorithms. The findings indicate that Support Vector Machines had an accuracy of 98.73%, while Random Forest had the best accuracy of 99.94%. Naïve Bayes had the lowest accuracy (57.66%), whereas the Decision Tree and k-NN algorithm achieved an accuracy of 99.82% and 94.08% respectively, and Logistic Regression gave an accuracy of 97.33%.

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
K-NN	94.08%	98%	90%	94%	0.9859
Decision Tree	99.82%	100%	100%	100%	0.9982
Random Forest	99.94%	100%	100%	100%	0.9996
Support Vector Machine	98.44%	100%	97%	98%	0.9963
Logistic Regression	97.93%	98%	98%	98%	0.9968
Naïve Bayes	57.66%	96%	16%	27%	0.6502
ANN	99.65%	99%	98%	98%	0.9965

Table 1: Scores of Different Models

The results indicate that Random Forest is the best performing algorithm in diagnosing hypothyroidism with an accuracy of 99.94%, precision of 100%, recall of 100%, F1 score of 100%, and AUC of 0.9996. The model is 0.12% more accurate than the second-best algorithm, Decision Tree, and 42.28% more accurate than the worst algorithm, Naïve Bayes.

The study looked into how several variables affected how well the models performed. The outcomes demonstrate that the performance of the models was significantly influenced by the quality of the data and feature selection. The study also discovered that the models performed better when pre-processing procedures like normalisation and imputation for missing variables were used.

Overall, the findings show that machine learning algorithms can reliably diagnose hypothyroidism, with Random Forest being the top-performing algorithm among those considered. The results further emphasise the significance of feature selection, data quality, and pre-processing methods in enhancing the performance of the models.

WEEKLY REPORT

During the first week of our internship, we were provided with a stroke dataset. We conducted a thorough analysis of the dataset, examining its various features and consulting relevant research papers to gain a comprehensive understanding of the topic.

In the second week, we identified suitable algorithms, both in machine learning and deep learning, that could be applied to our dataset. We proceeded to implement these algorithms and presented our progress to our mentor for feedback.

In the third week, we utilized techniques such as stacking, accuracy assessment, and classification reports to evaluate the performance of different models. We also employed various graphical representations to analyze the dataset effectively. Additionally, we conducted analyses considering individual attributes in relation to stroke.

During the fourth week, we synthesized our findings and knowledge into a research paper, drawing upon insights from multiple research papers. We also prepared a detailed internship report summarizing our work. Finally, we created a comprehensive presentation, which we delivered to our supervisor.

CONCLUSION

In conclusion, the goal of the research article on the hypothyroid dataset using machine learning was to assess how well the algorithms performed in detecting hypothyroidism and to pinpoint variables that influenced the effectiveness of the models. The efficacy of four machine learning algorithms—k-NN, Decision Trees, Random Forest, and Support Vector Machines—was examined in the study. These algorithms' effectiveness was assessed using a variety of measures, including accuracy, precision, recall, and F1 score.

The findings showed that Support Vector Machines had an accuracy of 98.44%, while Random Forest had the best accuracy of 99.94%. The study also discovered that the models' performance was significantly influenced by the quality of the data and feature selection. The study also shown that the models performed better when pre-processing procedures like normalisation and imputation for missing information were used The findings have significant outcome for hypothyroidism diagnosis. Better patient outcomes may result from the use of machine learning algorithms to enhance the accuracy and speed of the diagnosis. Additionally, the results point to the need for future research to pay more attention to data quality, feature selection, and pre-processing methods.

In conclusion, the study showed how machine learning algorithms might be used to identify hypothyroidism and emphasised the significance of data quality, feature choice, and pre-processing methods in enhancing the performance of the models. The results of this study can be used to guide future research in this field, and the approaches and procedures employed here can also be used in other medical settings where it is necessary to diagnose a range of illnesses.

REFERENCES

1. <https://indjst.org/articles/machine-learning-techniques-for-thyroid-disease-diagnosis-a-review>
2. http://papersim.com/wp-content/uploads/Neural_Network_Thyroid_Disease_2015.pdf
3. <https://iopscience.iop.org/article/10.1088/1742-6596/1963/1/012140/pdf>
4. <https://www.mdpi.com/2072-6694/14/16/3914>
5. https://www.researchgate.net/publication/341534298_Thyroid_Disease_Prediction_Using_Machine_Learning_Approaches
6. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9405591/>
1. https://www.researchgate.net/publication/306007162_Machine_Learning_Techniques_for_Thyroid_Disease_Diagnosis_-_A_Review