International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 4th Apr 2021. Link :http://www.ijiemr.org/downloads.php?vol=Volume-10&issue=ISSUE-04

DOI: 10.48047/IJIEMR/V10/I04/120

Title **Early Prediction of Thyroid Disease Using Machine Learning Classifiers**

Volume 10, Issue 04, Pages: 590-600

Paper Authors
Kishor Kumar Reddy C, Anisha P R

International Journal for Innovative Engineering and Management Research
A Peer Reviewed Open Access International Journal
www.ijiemr.org

Early Prediction of Thyroid Disease Using Machine Learning Classifiers

Kishor Kumar Reddy C [1] and Anisha P R [2]
[1,2]Stanley College of Engineering & Technology for Women, Hyderabad, India

Abstract:

In India and other countries, there is a significant problem caused due to thyroid diseases. Various researchers have studies on thyroid disease and estimates that about 42 million people in India suffer from thyroid diseases. Thyroid looks like small, butterfly-shaped gland located at the base of your neck with an intricate network of glands called the endocrine system. Endocrine system is responsible for coordinating and controlling many of your body's activities. The thyroid gland creates and produces hormones that regulate your body's metabolism and play a very important role in many different systems throughout your body. When your thyroid gland releases either too much or too little of these important hormones, it leads to thyroid disease. Different types of thyroid disease, like hypothyroidism, hyperthyroidism, thyroiditis, thyroid nodules, and thyroid cancer. So, with the help of Machine Learning techniques (i.e., Bagging and Boosting algorithm) and different machine learning classifiers, predicting the thyroid disease are implemented for supporting the health care field in making decisions.

Keywords: Thyroid gland, Hormones, Hypothyroidism, Hyperthyroidism, Machine Learning, Boosting algorithms.

1. Introduction

Thyroid sickness could be a common term accustomed to describe a condition in the medical field that allows our thyroid for creating proper hormones. The thyroid gland produces hormones that allow our body to function normally. Thyroxine, additionally known as T4, and it's a primary secretion created by the secretor. once the blood passed to the tissues, a tiny low indicates little or no growth of the T4 discharged from the secretor regenerated by liothyronine (T3), which is the active secretion [15].

The advancement in computational biology has been going on used in the healthcare field. It allows us to collect and maintain the data of patients for helping the medical field in disease prediction. They are very less intelligent machines available for predicting the different diseases and for diagnosis of thyroid disease at initial stages, intelligent systems can't easily analyse the thyroid root problem. As time goes on changing, machine learning has played an important part in solving hard and nonlinear types of problems for building a predictive model [17].

Healthcare field is a big business and produces large amounts of data. This hidden information within the healthcare field is

going to be used for effectively making the decision to help the individual to improve health conditions. Therefore, this area needs some improvement because many of the industry never use this data for making decisions [16].

Here important work is gaining information in a form of data or information because the volume of raw data is very huge so, with the help of machine learning, some helpful information for analysis thyroid dataset can be carried out. The expectation from this paper is to predict the thyroid i.e., gland disease early to make possible treatments for the patients who have been suffering from disease, which may reduce the risk of life and cost to get treatment. For this types of problems, Machine Learning approaches will be used to train and test the thyroid disease, and the data pre-processing techniques will be considered.

According to statistical analysis, thyroid diseases are on the increase in India. Approximately 1 out of 10 in Indian suffer from a thyroid problem. It has been estimated that 42 M persons have suffered from any type of thyroid disease. Predicting gland disorder by a doctor may be a tedious process that may cause negative predictions, only an experienced doctor can examine the case properly. To assist doctors machine learning can help them in the diagnosis of disease and reduces their burden [18].

In the healthcare field, many organizations (i.e., hospitals) are facing problems in the supply of good qualitative services within the limited cost. A good service means that gland diagnosing in patients correctly and helping them with effective treatments. Poor medical services will/may lead to disastrous types of consequences which may be not acceptable in the real world. Hospitals should minimize the value of tests. Many hospitals today try to employ hospital information system to manage the patient data i.e., thyroid data. This system will try to generate a huge volume of data, but these data have never used to support in decision-making.

The need of this paper work of classification of different thyroid diseases is to protect our world from the effect of different thyroid diseases which are affecting many our world's populations for a decade. Delayed diagnosis of diseases may be a fundamental problem thanks to a shortage of medical professionals. Automatic tools may help in reducing the work rate for doctors and reduce the death percentage that occurs due to different diseases. It is correct to say that, if anyone is suffering from any problem you get to know how tough time others go through in the same problem that you are suffering from [19].

2. Relevant Work

To indicate the difficulty level of Thyroid and to diagnosis it early, it important to add previous data that should be recorded for further progress. During this survey, we had reviewed the related work which is related to thyroid disorder and help to do prediction using various classification methods. "The diagnosis of abnormality in human is driven by a framework known as Computer-Aided. This specially encapsulates three

components as follows: a file as an input, Extraction of features and selecting the important features. we have reviewed the previous work done by researchers based on: Data Extraction methods and procedures, Feature selection approaches, Classification algorithms which are used by them, and the accuracy of the work" [20].

G. Rasitha Banu et al. [1], "Title of the paper: Predicting thyroid disease with the help of Linear Discriminant Analysis (LDA) data processing Technique". This system tries to explain about people to understand and identify the thyroid disease and to understand the prediction details and level of disease anywhere within the world. They used the classification method to find the prediction details. Here researchers try to predict, hypothyroid disease is with the help of data mining. The dataset in this is taken from the study on hypothyroid from the UCI repository. Here dataset has 3772 instances from which 3481 belong to negative and the rest belongs to hypothyroid disease. An experiment is done using Linear Discriminant Analysis for better results.In this, they had used the LDA for data-processing and for classification also and classify the hypo-thyroid disease. Validation known as k-fold is also considered.

Sanjeev Kumar, SunilaGodara et al. [2], "Thyroid Disease Prediction with the help of Machine Learning". In this paper, an effort is formed to research Logistic regression and as well as Support Vector Machine for multiclass label to do classification of thyroid dataset. Here dataset contains 3772 records and 30 features. Performance of these techniques is on basis of Precision, Recall, F measure, and accuracy. They have analysed that logistic regression is efficient when they have compared it with SVM for multi-class classification for thyroid disease datasets. Binary logistic regression will take two values 0 or 1 (y=b0+b1x=e) is used here to predict the disease. To predicted probability's sigmoid function is considered.

Dr. D Anitha, Mrs. S Sathya Priya, et al. [3], "Performance Improvement with the help of several approaches to Predict gland disease caused by thyroid". In this work, dimensionality reduction is used for selecting the subset of attributes from original dataset sample and have applied J48 and decision stump with the help of data mining for doing classification of class which are used to classify types of hypothyroid disease. Evaluated the model with the help confusion matrix for understanding error rate. J48 has performed well than decision stump and J48 gives a minimum error rate than the Decision stump. In this, future work is that the similar techniques can be applied for other disease too such as heart disease, Lung cancer, breast cancer.

Pushpanathan G, Gowthami Singh, and Anil Kumar et al. [4], "Comparative thyroid disease Analysis based on hormone-levels using Data Mining". In this work, they have used ML algorithms like SVM, classifier known as decision tree, and Bayesian approach such as Naïve Bayes, and used distance-based approach known as

knn, at last logistic regression. All this together is used to spot thyroid disease such as a software tool like anaconda and programming language is python for implementing these algorithms. At last, comparison of accuracy for logistic regression, random forest is done and then represented them in visualization form known as graphical manner.

Suresh Kumar Kashyap, Dr. NeelamSahu, et al. [5], "A Comparative Study of Machine Learning Based Model for Thyroid Disease Prediction". This work predicts the thyroid disease by using a classifier refers as Decision Tree and a neural network model such as Artificial Neural Network. This paper tries to focus and explains how the data is classified and well separated in simplest way. As a result, a set of operation had been carried out and completed in both segregation modes, here accuracy will be compared with the confusion matrix. It has been concluded that ANN provides better accuracy than other decision tree classifiers.

Marissa Lourdes De Ataide et al. [6], Paper title: "Thyroid Disease Detection with the help of Soft Computing Techniques". This work is based on applying a multilayer perceptron (i.e., which consists of 2 layers) classifier for disease diagnosis into three classes which refers as thyroid, hyperthyroidism and hypothyroidism, and to classify the hypothyroid disease as primary-level hypothyroidism, secondary-level hypothyroidism, and last tertiary level of hypothyroidism and trying to focus on max accuracy in very less time.

Sayyad Rasheed Uddin etal. [7], This paper has tried to examine and as well as assesses the utilization of the unique strategy known as feature selection combined with the classification model and techniques. With the help of Weka tool and classification approaches has been used to measure and understand the dataset. The results have been calculated using various test cases, which includes a technique called as 10-fold validation, the percentage divided with and without the feature selection method and training datasets. Results are analysed and compared with correctly classified observation, and with a runtime, and calculating absolute mean values for the present experiments.

SuwarnaGothane et al. [8], "Data Mining technique for classification of hypo thyroids detection: Association Women Outnumber Men". This paper tries to present thyroid patients data and analyse it. Proposed work results are employed for identifying important factors that are very much affecting hypothyroid. In this researcher have attempted to identify the hypothyroid earlier and analyse few parameters that can affect the thyroid. To overcome the matter researcher have come up with a strategy for a solution to identify the key factors that are affecting the hypo disease using the tool known as weka for classification.

Mir Saleem, S JahangeerSidiq. Akhtar Rasool Malik et al. [9], "Diagnosis and Classification of Thyroid Disorder using Machine Learning". In this, the researcher has provided a set of methods and has used

a new technique and as well as latest technology to convert dataset into some useful structure which can be used for supporting and making decision for identifying future outcomes. In this dataset is extracted and cleaned through several data mining approaches such as association rule for finding relationships among different features, classification for separating the variables into some set of classes, clustering, and pattern recognition like words recognition and it is very useful for experts. This survey is suggesting that the accuracy enhancement can be done by implementing a classifier such as ensemble rather than straight forward models.

Hasan makes et al. [10], have done research on seven different types of Networks knowing the top goal to support stronger & dependable systems for identifying thyroid disease. Researchers has used optimization techniques, but pre-processing is not done. Training and as well as testing the networks for identifying the disease is considered. NazariKousarrizi et al. [11,12], identified that Hyperthyroid can get stimulated by inflammation of the endocrine gland, various medical types of drugs, and lack of controlling on the secretion of several hormones. Gland disorder should not be ignored or underestimated as severe hyperthyroid and the last stage of hyperthyroidism, may lead to the death of the patient.

Decision tree [13], It's a type of predictive modeling that is used for classification and as well as prediction tasks. The decision tree makes use of divide and conquers approaches. Edge from each decision node refers to possible answer that is related with a question. Every terminal node refers to asolution for the problem. Dhyan Chandra Yadav and Saurabh Pal et al. [14], with their experiments have shown that the ensemble classification technique helps and improved evaluation accuracy and tests for thyroid datasets

3. Proposed Architecture

The proposed system will predict the different types of thyroid diseases like hypothyroidism, hyperthyroidism, no thyroid, and last is general health issues that are not related to thyroid. To predict thyroid disease, the ML model needs training, for training the ML model needs a dataset.The proposed system consists of a thyroid0387 dataset which is taken from UCI for training the machine learning model and five different classification algorithms for prediction.

There is no such automated tool or software used yet for the prediction of thyroid disease and till now thyroid diseases are predicted manually such as by the doctors and by the pathology lab members. This system will surely be helpful across the world for the prediction of different thyroid diseases.

International Journal for Innovative Engineering and Management Research
A Peer Reviewed Open Access International Journal
www.ijiemr.org

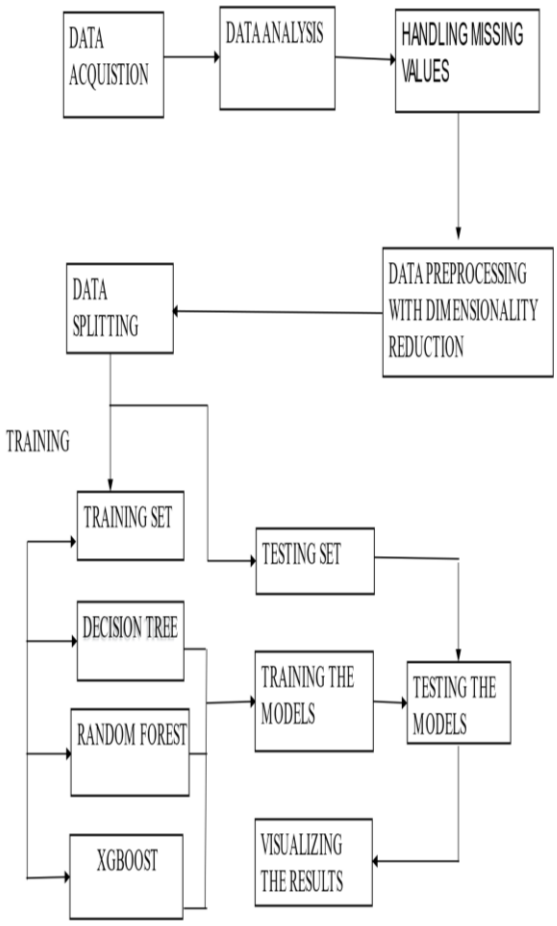

Fig 1: Proposed Architecture

Thyroid disease data is supplied by Garvan Institute, J. Ross Quinlan, at New South Wales Institute, at Sydney, in Australia.1987. The record which is provided in the UCI Machine Learning repository contains the updated version of the archive of thyroid disease diagnoses obtained from the Garvan Institute, which consists of 9172 records started from 1984 to 1987.

Here dataset consists of (29 attribute values), diagnoses [record identification].

Table 1: Thyroid Dataset

Attribute Name	Possible Values
Age of person	continuous values
Sex	M, F
On thyroxine	f, t.
Query on thyroxine	f, t.
On antithyroid medication	f, t.
Sick	f, t.
Pregnant	f, t.
Thyroid surgery	f, t.
I131 treatment	f, t.
Query hypothyroid	f, t.
Query hyperthyroid	f, t.
Lithium	f, t.
Goitre	f, t.
Tumor	f, t.
Hypopituitary	f, t.
Psych	f, t.
TSH measured	f, t.
TSH	Continuous
T3 measured	f, t.
T3	Continuous
TT4 measured	f, t.
TT4	Continuous
T4U measured	f, t.
T4U	Continuous
FTI measured	f, t.
FTI	Continuous
TBG measured	f, t.
TBG	Continuous

4. Discussion & Results

In this paper Random Forest have performed well Compared to other Supervised algorithms such as Xgboost classifier and Decision Tree because in random forest, we attempt to find ourselves with decision trees that aren't only trained on several different datasets of knowledge but also use different subset of columns to make decisions.

Why we need to consider Random Forest:

- It is more versatility and can be used for regression problems and classification tasks also, and it is also easy to visualize the relative importance for the input features.

- Random forest is a handy algorithm because the hyperparameters which is set as default produces a good result. Understanding hyper-parameters is straightforward approach, and there's very less parameters.
- One of the issues in supervised machine learning is overfitting, but in most of the situation this won't happen due to random forest classifier.

This work compares the different types of classifiers and by considering the evaluation metrics, best model will be decided according to the performance calculated. To support the medical field in disease prediction as acute as possible so that this model can be utilized by many doctors and fresher who are a part of medical practitioners for identifying thyroid disease.

Here when we have used different dataset i.e., hypothyroid dataset which contains 3163 observation and 26 features, Both the algorithms Random Forest and Decision Tree have same performance level where as Xgboost have performed very well compared to other two algorithms.

Accuracy	Classifier	Precision	Recall	FScore
0.9813	XGBClassifier	0.943	0.885	0.912
0.9782	Random Forest	0.917	0.884	0.900
0.9772	Decision Tree	0.904	0.891	0.897

Table 2: Result for Hypothyroid Dataset

Because Xgboost prunes the decision tree with a score card known as "Similarity score" before focusing on the actual purposes of modeling.Xgboost considers the "Gain" of a decision tree node as the dissimilarity between the similarity score for the decision tree node and the similarity score for the children node.

If the gain score from that node is minimal then it stops the construction of the decision tree, which can overcome the overfitting issue to an extent.Random Forest may overfit the dataset and if majority of the decision trees in the given forest are provided with samples which are similar.If decision trees are fully grown, then the model will try to collapse when the test dataset is introduced for testing.

XGBoost is considered as a good option when imbalance datasets are there but we can't expect this from random forest. The most important dissimilarity between Random Forest andXgboost is that the Xgboost tries to focus on functional space for reducing cost of a model whereas Random Forest focus more on hyperparameters for optimize the model.

According to data science and this study, Random Forest classifier has worked well because a large set of the uncorrelated models (which are not the same) in a decision tree can work as a committee and try to outperform for any individual i.e., Single model.

Dataset: (Thyroid Patient Dataset): All testing instances were taken as the testing data for which the confusion matrix is created.

International Journal for Innovative Engineering and Management Research
A Peer Reviewed Open Access International Journal
www.ijiemr.org

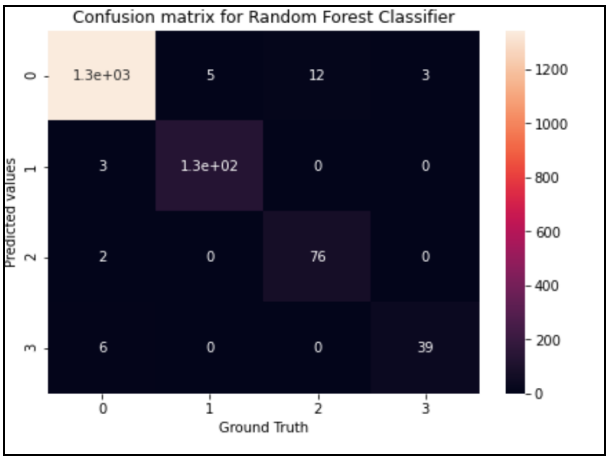

Figure 2: Confusion matrix for Random Forest of the dataset

classification report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1363
1	0.96	0.97	0.97	137
2	0.85	0.96	0.90	78
3	0.95	0.84	0.89	45
accuracy			0.98	1623
macro avg	0.94	0.94	0.94	1623
weighted avg	0.98	0.98	0.98	1623

Figure 3: Classification report for Random Forest of the dataset

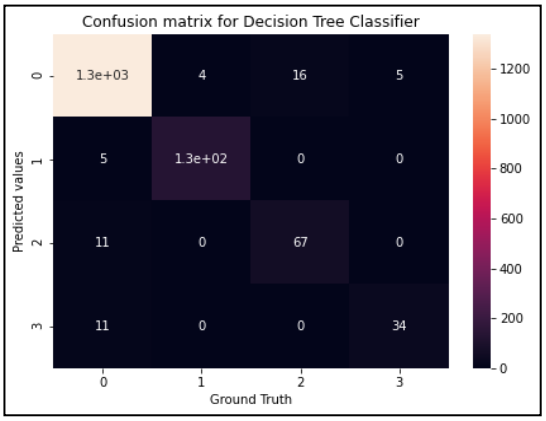

Figure 4: Confusion matrix for Decision Tree of the dataset with feature selection

classification report:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	1363
1	0.96	0.94	0.95	137
2	0.83	0.79	0.81	78
3	0.89	0.89	0.89	45
accuracy			0.97	1623
macro avg	0.91	0.90	0.91	1623
weighted avg	0.97	0.97	0.97	1623

Figure 5: Classification report for Decision Tree of the dataset with feature selection

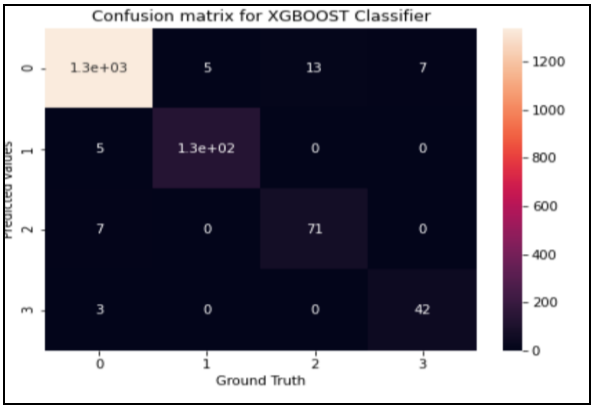

Figure 6: Confusion matrix for Decision Tree of the dataset without feature selection

classification report:

	precision	recall	f1-score	support
0	0.99	0.98	0.99	1363
1	0.96	0.96	0.96	137
2	0.85	0.91	0.88	78
3	0.86	0.93	0.89	45
accuracy			0.98	1623
macro avg	0.91	0.95	0.93	1623
weighted avg	0.98	0.98	0.98	1623

Figure 7: Classification report for Xgboost of the dataset without feature selection

By comparing all the algorithms on the thyroid dataset, the highest accuracy I got to know Random Forest is working best for both multi-class target and continuous. Hence, I conclude that Random Forest performed well compared to Xgboost and Decision tree for continuous data and for categorical data compared to all algorithms for the thyroid0387 dataset.

5. Conclusions

World health is severely affected by diseases that are spreading and increasing every day. The main challenge in the healthcare sector is the death rate caused by non-communicable and thyroid comparatively high to other factors. Lack of treatment & doing delay in diagnosis are the crucial factors for the death of patients. This study tries to work on different Machine Learning supervised algorithms and strategies which are very useful for the diagnosis of disease in an earlier stage.

In this, three classification supervised techniques are studied for predicting the disease and for training dataset different model is also generated, and a part of dataset is tested in the prediction stage. Here comparison between classifiers is based on the accuracy and confusion matrix and then effective classifier is identified by the performances. The dataset was tested using a classification algorithm using a python environment. From the experiment we can observed that the classifier known as Random Forest has provided best accuracy in a term of Performance, with respect to decision tree &xgboost.

For this review the objective is to help and provide a new set of guidelines and dimensions to the people involved in the healthcare field regarding Machine Learning techniques

References

1. G. Rasitha Banu, “Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique”. Communications on Applied Electronics (CAE) – ISSN: 2394-4714 Foundation of Computer Science FCS, New York, USA Volume 4– No12, January 2016.
2. SunilaGodara and Sanjeev Kumar, “Prediction of Thyroid Disease Using Machine Learning Techniques”. International Journal of Electronics Engineering (ISSN: 0973-7383) Volume 10 June 2018.
3. Dr. D Anitha, Mrs.S.SathyaPriya, “Performance Improvement with Multiple Approaches to Predict Disorders Caused by Thyroid Disease”. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727 PP 25-29 www.iosrjournals.org.
4. Pushpanathan G, Gowthami Singh, and Anil Kumar, “Comparative Analysis of Thyroid Disease based on Hormone Level using Data Mining Techniques”. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org NCETESFT – 2020.

5. Suresh Kumar Kashyap, Dr. Neelam Sahu, "A Comparative Study Of Machine Learning Based Model For Thyroid Disease Prediction". International Journal of Creative Research Thoughts © 2021 IJCRT | Volume 9, Issue 4 April 2021 | ISSN: 2320-2882.
6. Marissa Lourdes De Ataide1, AmitaDessai, "Thyroid Disease Detection using Soft Computing Techniques", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072.
7. SayyadRasheeduddin, KurraRajasekhar Rao, "Constructing a System for Analysis of Machine Learning Techniques for Early Detection of Thyroid". International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-2, July 2019.
8. SuwarnaGothane et al, "Data Mining Classification on Hypo Thyroids Detection: Association Women Outnumber Men". International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020.
9. Mir Saleem, S JahangeerSidiq. Akhtar Rasool Malik, "Diagnosis and Classification of Thyroid Disorder using Machine Learning - A Systematic Review". © 2019 JETIR January 2019, Volume 6, Issue 1 www.jetir.org (ISSN-2349-5162).
10. Makas, Hasan, and NejatYumusak. "A comprehensive study on thyroid diagnosis byneural networks and swarm intelligence." Electronics, Computer and Computation (ICECCO), 2013 International Conference on. IEEE, 2013.
11. M. R. NazariKousarrizi, F.Seiti, and M. Teshnehlab An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification.
12. L. Ozyılmaz and T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods," in: Proceedings of ICONIP'02 9th international conference on neural information processing (Singapore: Orchid Country Club, 2002) pp. 2033–2036).
13. http://www.worldscientific.com/worldscibooks/10.1142/6604.
14. Dhyan Chandra Yadav, Saurabh Pal, "Decision Tree Ensemble Techniques To Predict Thyroid Disease", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019..
15. Anisha P R , Kishor Kumar Reddy C and Nguyen Gia Nhu, "Blockchain Technology: A Boon at the Pandemic Times – A Solution for Global Economy Upliftment with AI

and IoT", EAI/Springer Innovations in Communication and Computing, 2022.

16. Kishor Kumar Reddy C, Anisha P R, Shastry R, Ramana Murthy B V, "Comparative Study on Internet of Things: Enablers and Constraints", Advances in Intelligent Systems and Computing, 2021
17. Kishor Kumar Reddy C, Anisha P R, Apoorva K, "Early Prediction of Pneumonia using Convolutional Neural Network and X-Ray Images", Smart Innovation, Systems and Technologies, 2021
18. Kishor Kumar Reddy C and Vijaya Babu B, "ISPM: Improved Snow Prediction Model to Nowcast the Presence of Snow/No-Snow", International Review on Computers and Software, 2015
19. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "SLGAS: Supervised Learning using Gain Ratio as Attribute Selection Measure to Nowcast Snow/No-Snow", International Review on Computers and Software, 2015
20. Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "A Pragmatic Methodology to Predict the Presence of Snow/No-Snow using Supervised Learning Methodologies", International Journal of Applied Engineering Research, 2014.