

Optimized building of machine learning technique for thyroid monitoring and analysis

¹Dr K Butchi Raju, ²Prasanna Kumar Lakineni, ³K.S.Indrani, ⁴G. Mary Swarna Latha, ⁵K. Saikumar

¹Department of CSE, GRIET, Hyderabad, Telangana, India. butchiraju.katari@gmail.com

²Associate Professor Dadi Institute of Engineering and Technology, Visakhapatnam, India. prasannakumar@gmail.com

³Department of ECE, Institute of Aeronautical Engineering, JNTUH, Hyderabad, India.

sriind.kotam@gmail.com

⁴Department of ECE, Institute of Aeronautical Engineering, JNTUH, Hyderabad, India. conswarnasuresh.tr@gmail.com

⁵JRF, Dept. of ECE, Koneru Lakshmaiah Education Foundation, Guntur, India. saikumarkayam4@ieee.org

Abstract: Many people now a days are suffering from a thyroid disorder. Hence, the earlier detection of thyroid becomes very important in the medical field to cure the patient. Thyroid disorder was caused by the imbalanced state of thyroid hormones and is of two types: overproduction (Hyperthyroidism) and less production (Hypothyroidism). This paper applied machine learning models on the hypothyroid dataset taken from the UCI data repository to partially fulfil thyroid disorder detection. The detection methods are based on pattern recognition, machine learning and data mining. The experimental results have shown that the proposed model will give good results.

Keywords: Machine Learning; Thyroid, Hyperthyroidism; Disease diagnosis; Health

I. INTRODUCTION

One of ten Indian struggle because of thyroid illness. Thyroid illness mainly happens for women between the ages of 17-54. Extreme thyroid level leads to increased blood pressure, maximizes cholesterol levels, cardiovascular complications, decreased fertility, and depression. An Electronic Health Record (EHR) comprises the information stored digitally regarding the health information regarding a specific person, including the opinions, laboratory experiments, reports diagnosis, medications, procedures, patient, predicting information, and allergies [1]. Thyroid hormone is formed by the thyroid gland, which stands as one of the endocrine glands. This hormone's primary role is to speed up the human body's metabolism, burn calories, protein, and restrict the other hormonal gland while there is excessive secretion [2]. The thyroid gland recognizing the thyroid illness from the examined report is a complicated and tedious job that can be determined only by experienced and knowledge. Traditionally, there are two approaches: examining the blood tests by lab technicians, and the other is doctor's diagnosis depending on the indications, symptoms, and checking of the patient physically [3] to predict the thyroid. This is not easier to examine every one report from the large dataset to predict

the result. The main task is to predict thyroid disorder with better accuracy.

The thyroid gland conceals thyroid hormones to regulate the body's metabolic level. Failure of thyroid hormone will tend to thyroid illness. The thyroid gland is an endocrine gland that produces hormones. The thyroid gland discharges thyroxine (T4) and triiodothyronine (T3) inside. The main hormones are contained in the bloodstream. Thyroid hormones have two functions: they regulate metabolism and influence development. A couple of greatest general issues of thyroid sickness or thyroid illness such as Hyperthyroidism - discharges excessive thyroid hormone inside blood because of overactive thyroid and Hypothyroidism - when the thyroid remains not in active status & discharges much less thyroid hormone inside blood[4].

II. LITERATURE

Priyanka Duggal et al. [5] proposed machine learning techniques for feature selection then classification aimed at thyroid disease diagnosis. There are two types of thyroid disorders that imbalance the metabolism rate in humans : hyperthyroidism and Hypothyroidism. The major task is to classify the thyroid disorder. Feature selection is the major problem in pattern recognition. This is included in pre-processing. Univariate filtering, repetitive elimination and tree-based filtering are the feature collection approaches suggested. Naïve Bayes, Help Vector and Random Forest were three strategies of classification. Prerana et al. [6] present a back-propagation technique towards detecting the thyroid. An artificial Neural network has been established to classify the preliminary thyroid prediction by back-propagating a mistake. ANN is then prepared for laboratory study, but not the same sets. The teaching can be performed as supervised schooling and unregulated learning in two forms. Dharamkar B et al. [7] uses machine learning fusing C4.5 and random forest classification technique to classify thyroid. Then the results are compared with other techniques. Parneet Kaur and

Deepak Aggarwal [8] proposed that classification approaches are discussed to predict the class label. This classification of dataset helpful aimed at predicting various diseases from a large volume of patient's dataset. The Diabetic's dataset is used to classify the decision table from the support and confidence to obtain better accuracy. The naïve Bayes and fuzzy KNN are processed together for the medical dataset, which provides better accuracy.

Gupta n et al. [9] increases the thyroid disease detection accuracy by a modified ant lion optimization algorithm (MALO). This method selects the most important features that improve the accuracy for classification and reduce computational time. Thyroid condition is treated with Random Forest, k-nearest Neighbour's, & Decision Tree. Ling Chen et al. [10] use Electronic Health Record to find the last stage patients and proposes SHG-Health is a graph-based semi-supervised learning algorithm (Semi-supervised Heterogeneous Graph on Health). The Cause of Death(COD) database are prepared for the high-risk dataset from the GHE database. This record contains patients' details, metal report and physical report and supervised learning to predict risky patients. Variable VS, Shukla S.[11] proposed hypothyroid study with KNN and SVM. They use K-Nearest Neighbors, SVM, Logistic regression, and Neural Network to find the hyperthyroid disorder stage.

Yadav DC and Pal S [12] identify thyroid symptom for treatment. The secret pattern in the data set is extracted using two methods. Ensemble-I generated by a decision tree, fitting & neural network, then Ensemble-II produced by bagging and boosting system combinations. Ensemble-I vs Ensemble-II is responsible for the proposed model. An ensemble-II produced model is higher than that of another ensemble-I model in the whole experimental set-up. Raisinghani Setak. [13] proposed a method to detect thyroid disease based on patient symptom reports. They applied different models to achieve better accuracy. Sudesh Kumar and Nancy [14 15] proposed a clustering and data mining technique.

METHODOLOGY

The proposed methodology consists of Data collection, Disease detection and assessment, and data pre-processing and model construction. Figure 1 illustrates the real operation[16-23].

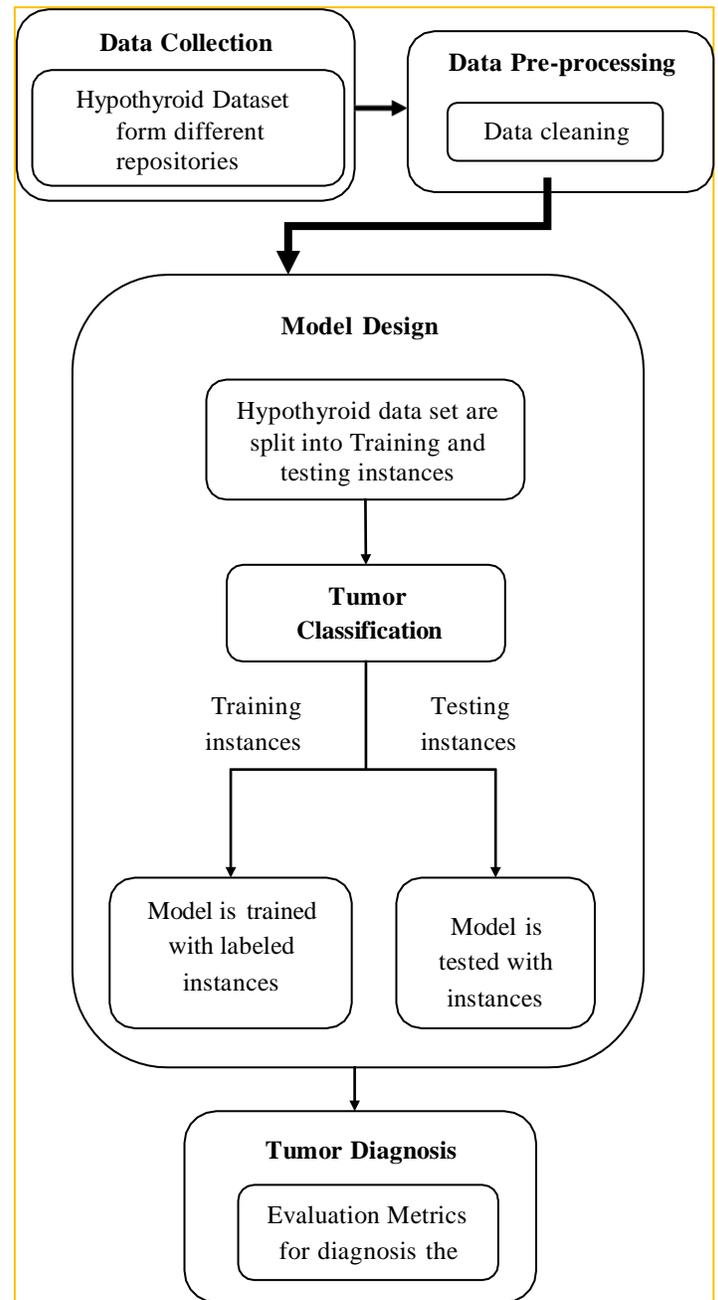


FIG 1. STRUCTURE OF THE MODEL

Data Collection: MR Images are collected from different repositories called UCI, Kaggle, etc.,

Data Pre-processing: Data pre-processing is the most critical and first stage of any machine learning project. During this process, raw data was gathered and transformed into a machine learning model [24-26]. This pre-processing cleans

up the data. All null values in the dataset are deleted during data cleaning.

Model Design: Splitting Data and Classification are the two steps that make up model architecture.

Splitting data: The dataset is split hooked on train and testing instances for preparation and examining the model. During this work, 80% of information was taken as prepared information, and 20% is taken as data for testing.

Classification: As the dataset contains hypothyroid instances, various machine learning practices have functioned on the dataset. This paper considered logistic regression, decision tree, random forest, AdaBoost, and Gradient Boost Classifiers for prediction. Logistic regression [15, 16]: This method specifies the probability of a class like Hypothyroid or Negative. This model makes use of the logistic function to make a binary dependent variable. Let x_1 & x_2 are predictors that are variable with a binary answer Y . Consider there remains a straight line among x_1 & x_2 to log of odds of the event $Y=1$, mathematically this relation is shown in equation 1.

$$l = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \quad (1)$$

Where ℓ is log-odds, b remains the logarithm's foundations, then β_i remain the generated model parameters. The above formula can be improved by the exponentiation of the log-odds shown in equation 2.

$$\frac{p}{1-p} b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \dots \quad (2)$$

By using easy algebraic operation, the probability β_i is by equation 3.

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} + 1} = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} \dots \quad (3)$$

Decision Tree [17, 18]: Decision Trees comes under the category of Supervised Learning. Each problem is solved by using tree representation by this classifier. Each internal node represents the features; branches represent the feature's outcome, then the leaf node represents target class labels. A class's prediction process in decision trees starts from the root with the comparison of tree attribute values with internal nodes until the leaf node with predicted class value reaches. The whole training set is considered as root with predicted class value reached. Data derives in archives of the form: $(X, Y) = (x_1, x_2, \dots, x_n, Y)$

The variable Y remains the target variable. X is a vector of all attributes. Information Gain measures how a given attribute will separate the training inputs according to their required target classification. The measure of change in entropy is known as Information gain. It is calculated using equation 4.

$$\text{Infogain} = - \sum_{i=1}^n p_i \log_2 p_i \dots \quad (4)$$

Information gain is used to find the best attribute at every node towards split the tree. It keeps the tree smaller. The attribute is considered as the best attribute if the information gain for that attribute is high.

Random Forest [19, 20]: Random forest classifier is an ensemble technique. An ensemble is a combination of two or more models. This method splits the dataset into random samples, then applies a decision tree to each sample. The output is obtained by performing majority voting on every output from each decision tree. Consider training set, D , of d tuple. K decision trees are produced for every sample in the following way: For every iteration, i ($i = 1, 2, k$), a training set, D_i , of d tuples remains sampled by replacing D . Let F be the number of features that are obtained at every division of decision treat every node. To prepare a decision tree classifier, M_i accidentally picks, at every node, F attributes as division candidates at any node. The trees are bigger and not clipped. The trees are huge. Forest-RI is termed random forests modelled like this and randomly chosen inputs.

AdaBoost [21,22,23]: It is an Adaptive Boosting algorithmic technique. Given a dataset D with d class labelled tuples like $(X_1, y_1), (X_2, y_2), \dots, (X_d, y_d)$, here i remains the class label of tuple X_i . This method assigns an equivalent weight of $1/d$ to every tuple. K classifiers are produced aimed at every round. In round i , the training set i is formed by sampling the tuples from the dataset. The substitution sample is used – more than once, the same tuple is picked. The ability to pick a tuple depends on its weight. A style classifier, M_i , comes from D_i 's fitness tuples. Test data is formed by calculating its error. The weights of the teaching tuples are then matched to the classification. The tuple weights remain increases when it is incorrectly classified, and the weight decrease if correctly classifies. Classification is difficult based on tuples weight – the heavier the weight, the more misclassified it is. These weights will produce the samples for the next round of classification. The simple concept is to concentrate on the misclassified tuples in the previous round while constructing a classifier. Gradient Boost[24,25]: It is also an ensemble method. This method makes use of stage-wise fashion to build the model. In this method, weak learners are combined into a single strong learner. In the low-square retrieval environment, where the aim remains towards reduce the mean squared error and "teach" model F to predict values of the form $=F(x) = F(x)$, it is simpler to grasp. It is shown in equation 5

$$MSE = \frac{1}{n} \sum_i (\hat{y} - y_i)^2 \dots \quad (5)$$

Now Hypothyroid Diagnosis: In this phase, results are analyzed/diagnosis for disease identification. This also analyzed the model for better understanding and identification.

III. DATA SET

The authors used a data set (kaggle) for their experiments containing 3163 records about persons with thyroid dysfunctions. The classification model has 26 attributes; the class attribute is the target, and it has two possible values: Hypothyroidism and negative. A description of the attributes utilized in the experiments remains shown in Table 1.

SL.NO	Attribute	Description
1.	Class	Hypothyroid/Negative
2.	Age	Numeric
3.	Sex	M/F
4.	on_thyroxine	F/T
5.	query_on_thyroxine	F/T
6.	on_antithyroid_medication	F/T
7.	thyroid_surgery	F/T
8.	query_hypothyroid	F/T
9.	query_hyperthyroid	F/T
10.	pregnant	F/T
11.	sick	F/T
12.	tumor	F/T
13.	lithium	F/T
14.	goitre	F/T
15.	TSH_measured	Y/N
16.	TSH	Numeric
17.	T3_measured	Y/N
18.	T3	Numeric
19.	TT4_measured	Y/N
20.	TT4	Numeric
21.	T4U_measured	Y/N
22.	T4U	Numeric
23.	FTI_measured	Y/N
24.	FTI	Numeric
25.	TBG_measured	Y/N
26.	TBG	Numeric

Preprocess the data by using python Label Encoder from sklearn. Preprocessing module, the actual code is depicted in table 2.

<pre> from sklearn.preprocessing import LabelEncoder encoder = LabelEncoder() data_copy['Unnamed: 0'] = encoder.fit_transform(data_copy['Unnamed: 0']) data_copy['Sex'] = data_copy['Sex'].replace({'M':0, 'F':1}) data_copy = data_copy.replace(to_replace={'f':0, 't':1, 'y':1, 'n':0}) </pre>
--

The dataset's actual data is depicted in the histogram format, shown in Figures 2 (directly) and 3 (NaN values are replaced with Mean & Median. The actual purpose of these figures is which features are important & which ones not. From Fig 2 & 3, it is observed that every feature is important. The Hypothyroid and Negative class comparison is depicted in Fig 4. From Fig 4, it is observed that Negative class number than Hypothyroid class.

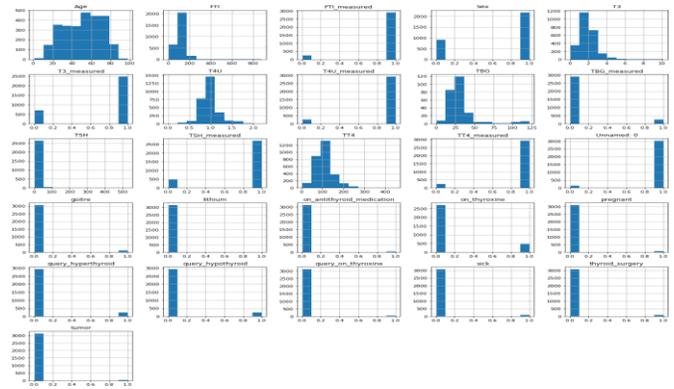


FIG 2 HISTOGRAM OF EVERY FEATURE

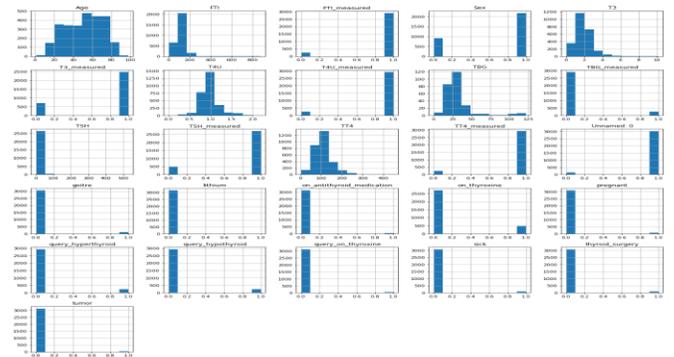


FIG 3 HISTOGRAM OF EVERY FEATURE AFTER NAN VALUES ARE REPLACED WITH MEAN & MEDIAN

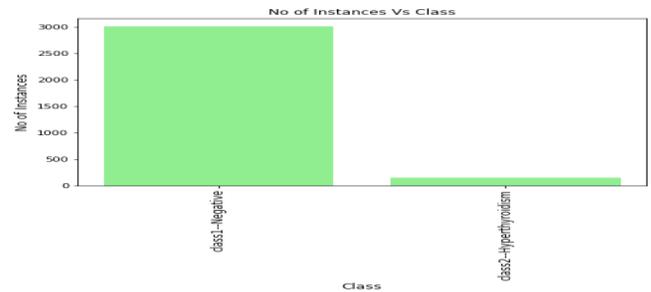


FIG 4 NEGATIVE AND HYPERTHYROIDISM CLASS COMPARISON

IV. RESULTS AND DISCUSSIONS

This paper considered logistic regression, decision tree, random forest, AdaBoost, Gradient Boost Classifiers. This paper considered the hypothyroid dataset, which contains 26 attributes and 3163 instances for testing the models. Sample Instances of the dataset is shown in Fig 5.

Unnamed: 0	Age	Sex	on_thyroxine	query_on_thyroxine	on_antithyroid_medication	thyroid_surgery	query_hypothyroid	que
0	hypothyroid	72	M	f	f	f	f	f
1	hypothyroid	15	F	t	f	f	f	f
2	hypothyroid	24	M	f	f	f	f	f
3	hypothyroid	24	F	f	f	f	f	f
4	hypothyroid	77	M	f	f	f	f	f

FIG 5 HYPOTHYROID DATASET SAMPLE INSTANCES



FIG 6 CONFUSION MATRIX FROM LOGISTIC REGRESSION

The confusion matrix from logistic regression is shown in Fig 6 for the testing dataset. From Fig 6, it remains observed that TP=26, TN=593, FP=3 and FN=11. Accuracy for the testing set is 0.977, and the F1 score is 0.888.



FIG 7 CONFUSION MATRIX FROM DECISION TREE

The confusion matrix from the Decision Tree is shown in Fig 7 for the testing dataset. From Fig 7, it remains observed that TP=31, TN=594, FP=2 and FN=6. Accuracy for the testing set is 0.987, and the F1 score is 0.937.

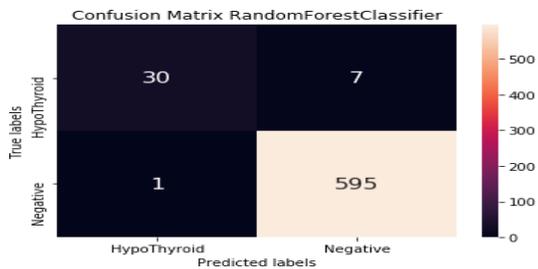


FIG 8 CONFUSION MATRIX FROM RANDOM FOREST

The confusion matrix from Random Forest is shown in Fig 8 for the testing dataset. From Fig 8, it remains observed that TP=30, TN=595, FP=1 and FN=7. Accuracy for the testing set is 0.987, and the F1 score is 0.937.

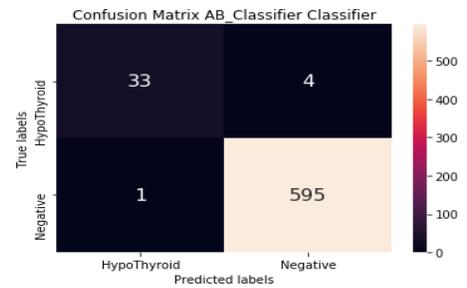


FIG 9 CONFUSION MATRIX FROM ADABOOST

The confusion matrix from AdaBoost is shown in Fig 9 for the testing dataset. From Fig 9, it remains observed that TP=33, TN=595, FP=1 and FN=4. Accuracy for the testing set is 0.992, and the F1 score is 0.962.



FIG 10 CONFUSION MATRIX FROM GRADIENT BOOST

The confusion matrix from Gradient Boost is shown in Fig 10 for the testing dataset. From Fig 10, it remains observed that TP=32, TN=593, FP=3 and FN=5. Accuracy for the testing set is 0.987, and the F1 score is 0.941. From Fig 10, it is concluded that this data set unbalanced, so instead of accuracy, we considered the F1 score for comparison. F1 score for logistic regression, decision tree, random forest, AdaBoost, Gradient Boost Classifiers is shown in Table 3.

	Logistic regression	Decision tree	Random forest	AdaBoost	Gradient Boost
F1 score	0.888	0.939	0.937	0.962	0.941

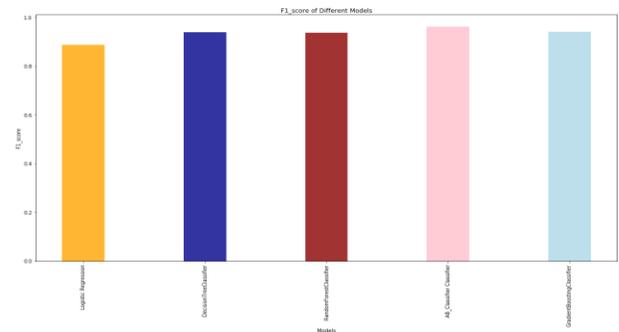


FIGURE:12 F 1 SCORE

The graph is drawn based on the values of Table 3 and depicted in Fig 11., it is observed that AdaBoost has shown the highest F1 score, so it exhibits good accuracy for the dataset.

V. CONCLUSIONS

Among the glands in the endocrine thyroid gland remains the largest gland. In the considered dataset, the target variable remains classified into two values Hypothyroid and Negative. Data mining is applied to classify these two from the whole dataset. Logistic regression, decision tree, random forest, AdaBoost, Gradient Boost techniques are applied to the dataset to predict the hypothyroid disorder. This research consumes investigated thyroid dataset by ranked improved F-score ordering. The work presented in this paper has shown that the F1 score is a reliable one for predicting the classification of thyroid patients. The inference of this that this score is diagnosed correctly and early for more patients. This work is applied in future for heart disease, diabetes and others datasets for validation.

REFERENCES

- [1] Temurtas F. A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications*. 2009 Jan 1;36(1):944-9.
- [2] Sollini M, Cozzi L, Chiti A, Kirienco M. Texture analysis and machine learning characterize suspected thyroid nodules and differentiated thyroid cancer: Where do we stand?. *European journal of radiology*. 2018 Feb 1;99:1-8.
- [3] Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*. 2017 Aug 1;30(4):477-86.
- [4] Chen HL, Yang B, Wang G, Liu J, Chen YD, Liu DY. A three-stage expert system based on support vector machines for thyroid disease diagnosis. *Journal of medical systems*. 2012 Jun 1;36(3):1953-63.
- [5] Duggal P, Shukla S. Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques. In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2020 Jan 29 (pp. 670-675). IEEE.
- [6] Prerana, Parveen Sehgal, Khushboo Taneja, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", published in *International Journal of Research in Management, Science & Technology*, Vol. 3, No. 2, April 2015.
- [7] Dharamkar B, Saurabh P, Prasad R, Mewada P. An Ensemble Approach for Classification of Thyroid Using Machine Learning. In *Progress in Computing, Analytics and Networking 2020* (pp. 13 -22). Springer, Singapore.
- [8] Parneet Kaur, Deepak Aggarwal, "Classification of Medical Dataset using Hybrid Feature Selection & Enhanced Decision Table Classification Approach", *International Journal of Science and Research*, Volume 6 Issue 3, March 2017
- [9] Gupta N, Jain R, Gupta D, Khanna A, Khamparia A. Modified Ant Lion Optimization Algorithm for Improved Diagnosis of Thyroid Disease. In *Cognitive Informatics and Soft Computing 2020* (pp. 599 -610). Springer, Singapore.
- [10] Ling Chen, Xue Li, Quan Z. Sheng, Wen-Chih Peng, "Mining Health Examination Records - A Graph-based Approach", *IEEE Transactions On Knowledge Discovery And Engineering*, 2016.
- [11] Yadav DC, Pal S. To generate an ensemble model for women thyroid prediction using data mining techniques. *Asian Pacific Journal of Cancer Prevention*. 2019 Apr 1;20(4):1275-81.
- [12] Raisinghani S, Shamdassani R, Motwani M, Bahreja A, Lalitha PR. Thyroid Prediction Using Machine Learning Techniques. In *International Conference on Advances in Computing and Data Sciences 2019 Apr 12* (pp. 140-150). Springer, Singapore.
- [13] Sudesh Kumar, Nancy, "Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization", *International Journal on Recent and Innovation Trends in Computing Communication*, Volume: 2 Issue: 10, 2014.
- [14] Kurt I, Ture M, Kurum AT. Comparing logistic regression performances, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert systems with applications*. 2008 Jan 1;34(1):366-74.
- [15] Ohsaki M, Wang P, Matsuda K, Katagiri S, Watanabe H, Ralescu A. Confusion-matrix-based kernel logistic regression for imbalanced data classification. *IEEE Transactions on Knowledge and Data Engineering*. 2017 Mar 15;29(9):1806-19.
- [16] Metwaly, A. F., Rashad, M. Z., Omara, F. A., & Megahed, A. A. (2014). Architecture of multicast centralized key management scheme using quantum key distribution and classical symmetric encryption. *The European Physical Journal Special Topics*, 223(8), 1711-1728
- [17] Ravi Kumar Gupta, " Employment Security and Occupational Satisfaction in India," *Journal of Advanced Research in Dynamical & Control System*, Vol. 10, Issue 10, pp. 244-249, 2018.
- [18] Farouk, A., Zakaria, M., Megahed, A., & Omara, F. A. (2015). A generalized architecture of quantum secure direct communication for N disjointed users with authentication. *Scientific reports*, 5(1), 1-17
- [19] Naseri, M., Raji, M. A., Hantehzadeh, M. R., Farouk, A., Boochani, A., & Solaymani, S. (2015). A scheme for secure quantum communication network with authentication using GHZ-like states and cluster states controlled teleportation. *Quantum Information Processing*, 14(11), 4279-4295
- [20] Wang, M. M., Wang, W., Chen, J. G., & Farouk, A. (2015). Secret sharing of a known arbitrary quantum state with noisy environment. *Quantum Information Processing*, 14(11), 4211-4224
- [21] Ravi Kumar Gupta, " Minimum Wage and Minimum Work Hour in India", *Journal of Advanced Research in Dynamical & Control System*, Vol. 11, 02-Special Issue, pp. 2402-2405, 2019.
- [22] Kankaew, K. (2020). The Evolution of Agribusiness Management Values from Labor to Brain Mechanism that Shape Leadership Style. *E3S Web of Conferences*, Vol. 175, no.13033.
- [23] Leo Willyanto Santoso, Bhopendra Singh, S. Suman Rajest, R. Regin, Karrar Hameed Kadhim (2020), "A Genetic Programming Approach to Binary Classification Problem" *EAI Endorsed Transactions on Energy*, Vol.8, no. 31, pp. 1-8. DOI: 10.4108/eai.13-7-2018.165523
- [24] Kankaew, K. (2020). Mindset development by applying U theory and religious concept in educational system: Thailand as a case. *E3S Web of Conferences*, Vol. 164, no.12002.
- [25] Zhou, N. R., Liang, X. R., Zhou, Z. H., & Farouk, A. (2016). Relay selection scheme for amplify- and forward cooperative communication system with artificial noise. *Security and Communication Networks*, 9(11), 1398-1404.
- [26] R. Arulmurugan and H. Anandakumar, " Region-based seed point cell segmentation and detection for biomedical image analysis," *International Journal of Biomedical Engineering and Technology*, vol. 27, no. 4, p. 273, 2018.