JES
Journal of Engineering Sciences

Vol 13, Issue 06, JUNE/2022

ISSN NO:0377-9254

THYROID DISEASE DETECTION USING MACHINE LEARNING TECHNIQUES

1Mr.M.Ram Chandra, Assistant Professor, BTech, Department of CSE, ramchandram@sreenidhi.edu.in

2Nunchu Ayushi Yadav, BTech, Department of CSE, ayushiyadav585@gmail.com

3Gadey Yeshwanth Reddy, BTech, Department of CSE, yeshureddy26@gmail.com

4S.Sai Abhinav, BTech, Department of CSE, abhi888601@gmail.com

ABSTRACT: Thyroid disorder leading cause of medical diagnosis and prediction development, which medical science is a complicated axiom. The thyroid gland is one of our body's main organs. Thyroid hormone secretions are responsible for regulating metabolism. Hyperthyroidism and hypothyroidism are the two prominent thyroid disorders that produce thyroid hormones for the control of body metabolism. Machine learning is critical in the disease prediction process and in the study and classification models used for thyroid disease on the basis of data obtained from hospital datasets. A decent knowledge base must be ensured, built, and used as a hybrid model to solve dynamic learning tasks like medical diagnosis and prediction of tasks. Basic techniques of machine learning are used for the identification and inhibition of the thyroid. The data set is trained by using algorithms such as Random Forest Classifier, XG Boost, KNN Classifier, Logistic Regression. The Random Forest Classifier is used to predict the Thyroid of the patient. The dataset is trained by the algorithm to get the accuracy and data cleaning is done to improve the accuracy. If the patient has a risk of getting thyroid our system has to give suggestions like recommending Foods to eat and Foods to avoid, medication etc.

Keywords- *Random Forest Classifier, XG Boost, KNN Classifier, Logistic Regression.*

1. INTRODUCTION

The evolvement computational biology is used in healthcare industry. It allows collection of stored patient data for the prediction of the disease. There are prediction algorithms which are available for the diagnosis of the disease at early stages. The medical information systems are rich of datasets but there are only few intelligent systems which can easily analysis the disease. Over a period of time, the machine learning algorithms have started playing a crucial role in resolving the complex and non-linear problems in the developing model. In any disease prediction models are used to override the features that can be selected from different datasets which can be used in classification in healthy patient as accurate as possible. If this is not done, misclassification can lead to a healthy patient getting unnecessary treatment. The Thyroid gland is an endocrine gland present in the human neck beneath the Adam's apple which help in secretion of thyroid hormone that influence the rate of metabolism and protein synthesis. The thyroid hormones are useful in counting how briskly the heart beats and how fast we burn calories. The thyroid secretes two types of active hormones called levothyroxine (T4) and triiodothyronine (T3). These hormones help in regulating the body temperature. These also aid in energy-bearing and transmission in every part of the body and decisive in protein management. Iodine is considered as the main building block of the thyroid gland. It's prostrated in few specific problems. Undersupply of these hormones can lead to hyperthyroidism. There are many originations related to hyperthyroidism and underactive thyroids. There are various kinds of medications like thyroid surgery is liable to ionizing radiation, continual tenderness of the thyroid, deficiency of iodine and lack of enzyme to make thyroid hormones.

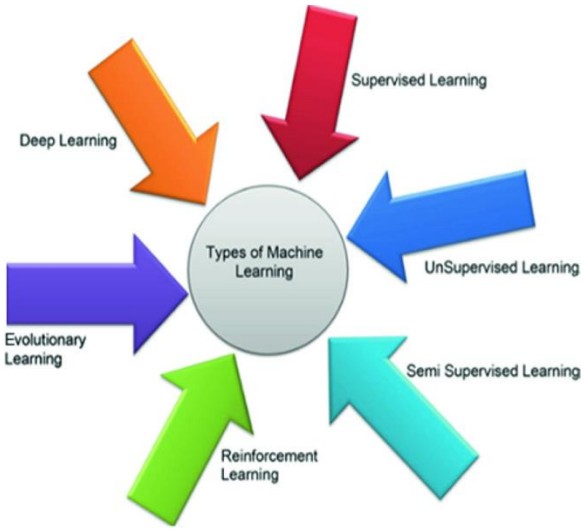


Fig.1: Machine learning techniques for Thyroid detection

Dosing the levothyroxine is not an easy task since the treatments can vary greatly and strongly depend on the amount of residual thyroid function of the patient, the body weight, and thyroid-stimulating hormone levels [12]. For this reason, the dose of levothyroxine should be administered over the patients lifetime and adjusted based on the physiological changes (a.e., weight or hormonal changes) throughout life and concomitant medical conditions (a.e., pregnant women). This requires continuous monitoring of the patients status based on clinical and laboratory assessment and appropriate adjustment of their levothyroxine therapy. Therefore, the prediction of treatment trends could represent an useful support to the endocrinologist and can improve the quality of life of the patient. The use of machine learning techniques can effectively support endocrinologists while monitoring patients. Recent studies have been successfully applied to classify and predict and, for this reason, have been widely used in the diagnosis of many different problems such as heart disease [17], diabetes [14] and Parkinson's disease [3, 4], reducing the time and costs required for the treatment of a patient. This study proposes an approach based on machine learning techniques exploiting hormonal parameters related to the thyroid and other clinical data concerning the patient, to predict if the patient's treatment needs to be increased, decreased, or remain unchanged.

2. LITERATURE REVIEW

2.1 Interactive Thyroid Disease Prediction System using Machine Learning Techniques:

Thyroid disease is a major cause of formation in medical diagnosis and in the prediction, onset to which it is a difficult axiom in the medical research. Thyroid gland is one of the most important organs in our body. The secretions of thyroid hormones are culpable in controlling the metabolism. Hyperthyroidism and hypothyroidism are one of the two common diseases of the thyroid that releases thyroid hormones in regulating the rate of body's metabolism. Data cleansing techniques were applied to make the data primitive enough for performing analytics to show the risk of patients obtaining thyroid. The machine learning plays a decisive role in the process of disease prediction and this paper handles the analysis and classification models that are being used in the thyroid disease based on the information gathered from the dataset taken from UCI machine learning repository. It is important to ensure a decent knowledge base that can be entrenched and used as a hybrid model in solving complex learning task, such as in medical diagnosis and prognostic tasks. In this paper, we also proposed different machine learning techniques and diagnosis for the prevention of thyroid. Machine Learning Algorithms, support vector machine (SVM), K-NN, Decision Trees were used to predict the estimated risk on a patient's chance of obtaining thyroid disease.

2.2 Comparison Study of Radiomics and Deep-Learning Based Methods for Thyroid Nodules Classification using Ultrasound Images:

Thyroid nodules have a high prevalence and a small percentage is malignant. Many non-invasive methods have been developed with the help of the Internet of Things to improve the detection rate of malignant nodules. These methods can be roughly categorized into two classes: radiomics based and deep learning based approaches. In general, convolutional neural networks based deep learning methods have achieved promising performance in many medical image analysis and classification applications; however, no existing comparison has been done between radiomics based and deep learning based approaches. Therefore, in this paper, we aim to compare the performance of radiomics and deep learning based methods for the classification of thyroid nodules from ultrasound images. On one hand, we developed a radiomics based method, which consists of extracting high throughput 302-dimensional statistical features from pre-processed images. Then dimension reduction was performed using mutual information and linear discriminant analysis respectively to achieve the final classification. On the other hand, a deep learning based method was also developed and

tested by pre-training a VGG16 model with fine-tuning. Ultrasound images including 3120 images (1841 benign nodules and 1393 malignant nodules) from 1040 cases were retrospectively collected. The dataset was divided into 80% training and 20% testing data. The highest accuracies yielded on the testing data for radiomics and deep learning based methods were 66.81% and 74.69%, respectively. A comparison result demonstrated that the deep learning based method can achieve a better performance than using radiomics..

2.3 Prediction of Thyroid Disease Using Machine Learning Techniques:

The paper presents several methods of feature selection and classification for thyroid disease diagnosis, related to the machine learning classification problems. Two common diseases of the thyroid gland, which releases thyroid hormones for regulating the rate of body's metabolism, are hyperthyroidism and hypothyroidism. Classification of these thyroid diseases is a considerable task. An important problem of pattern recognition is to extract or select feature set, which is included in the pre-processing stage. The proposed methods of feature selection are Univariate Selection, Recursive Feature Elimination and Tree Based Feature Selection. Three classification techniques have been used namely Naïve Bayes, Support vector machines and Random Forest. Results shows that the Support Vector Machines are the most accurate technique and hence this was used as a classifier to separate the symptoms of thyroid diseases into 4 classes namely Hypothyroid, Hyperthyroid, Sick Euthyroid and Euthyroid (negative).

2.4 Segmentation of Thyroid Gland in Ultrasound image using Neural Network:

The thyroid gland is highly vascular organ, and lies in the anterior part of the neck just below the thyroid cartilage. Ultrasound imaging is most commonly used to detect and classify abnormalities of the thyroid gland. Other modalities (CT/MRI) are also used. There is a challenge to segment ultrasound medical image which is often blurred and consists of noise as other modalities like CT contains ionizing radiations and expensive. Thus, there is a need to apply a method to automated segment well the objects for future analysis without any assumptions about the object's topology are made. Various methods or techniques are used for automatic segmention of thyroid gand but the application of neural network in image processing provides a better solution to segmentation problem. In this paper we use Feedforward neural network to classify the region using feature extraction and then segment it. Experiment and results are shown.

2.5 Diagnosis of thyroid disease using artificial neural network methods:

Proper interpretation of the thyroid gland functional data is an important issue on the diagnosis of thyroid disease. The primary role of the thyroid gland is to help regulation of the body's metabolism. Thyroid hormone produced by the thyroid gland provides this. Production of too little thyroid hormone (hypo-thyroidism) or production of too much thyroid hormone (hyper-thyroidism) definites the type of thyroid disease. In this work, various neural network methods have been used to help diagnosis of thyroid disease.

2.6 A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis:

Proper interpretation of the thyroid gland functional data is an important issue in the diagnosis of thyroid disease. The primary role of the thyroid gland is to help regulation of the body’s metabolism. Thyroid hormone produced by the thyroid gland provides this. Production of too little thyroid hormone (hypothyroidism) or production of too much thyroid hormone (hyperthyroidism) defines the type of thyroid disease. Artificial immune systems (AISs) is a new but effective branch of artificial intelligence. Among the systems proposed in this field so far, artificial immune recognition system (AIRS), which was proposed by A. Watkins, has shown an effective and intriguing performance on the problems it was applied. This study aims at diagnosing thyroid disease with a new hybrid machine learning method including this classification system. By hybridizing AIRS with a developed Fuzzy weighted pre-processing, a method is obtained to solve this diagnosis problem via classifying. The robustness of this method with regard to sampling variations is examined using a cross-validation method. We used thyroid disease dataset which is taken from UCI machine learning respiratory. We obtained a classification accuracy of 85%, which is the highest one reached so far. The classification accuracy was obtained via a 10-fold cross-validation.

3. IMPLEMENTATION

In the prediction process, machine learning plays a key role, and paper research and the classifications of

models used in thyroid disease detection. The data set is taken from Kaggle Website. We also proposed different approaches for machine learning and thyroid diagnosis. Machine Learning Algorithms: Support Vector Machine, XG Boost, Random forest, and K-NN classifier were used to calculate an estimated probability of a patient having thyroid disease. We will also suggest what foods to eat and foods to avoid. We will use the Random Forest Classifier algorithm for the front end. The back end and Front end is connected by Flask. Data cleaning is done to improve the accuracy.

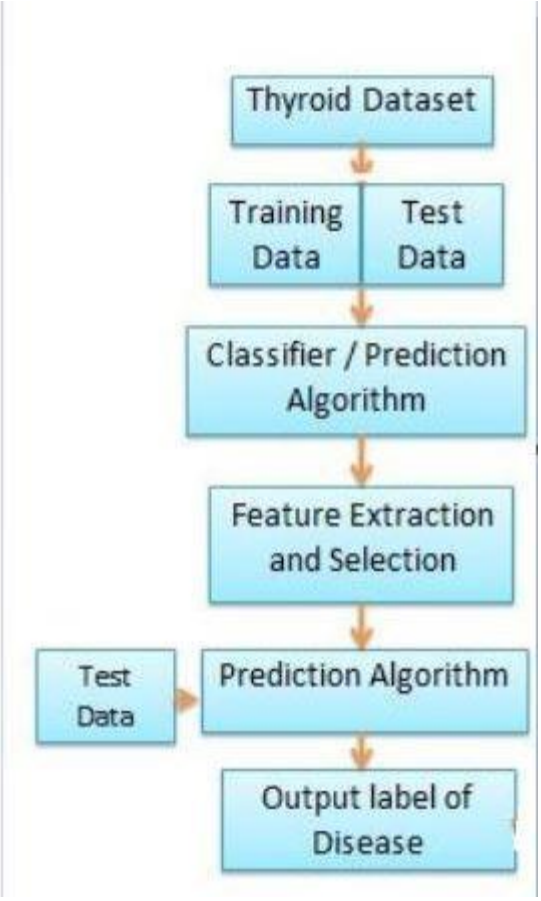


Fig.2: Workflow diagram

The technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

For predicting Thyroid disease analyzing blood report is required to analyze and predict disease. Thyroid blood test data set analysis will be conducted using various supervised machine learning classifier techniques. Based on the accuracy of different algorithm, best accuracy algorithm will be chosen to fetch the result. For first part, thyroid data set is taken from UCI repository. The dataset of hyperthyroidism and hypothyroidism is used where hyper and hypo are the two labels. These data set need to be checked before feeding it to training. There may be presence of null data or unnecessary data, this should undergo data cleaning to remove such data. Cleaned data is used as training data and test data, which is fed as input to the algorithm. The algorithm extracts the features from different dataset to classify the data according to the labels. To check the accuracy of the prediction, test data is fed to the algorithm. Based on the feature extracted, probability will be generated for test data by comparing the features of both. Highest probability value will be classified to that particular label whether it is hyperthyroidism or hypothyroidism.

We have used four algorithms such as Random forest classifier, XG Boost, Logistic Regression, K-Nearest Neighbor out of which Random Forest and XG boost has achieved higher accuracies such as 98 and 99 respectively, as random forest is a robust algorithm. It has been used for the front end implementation Our main aim of the project is to detect the thyroid disease at early stages and with minimum of parameters with accurate results. The website is designed in such a way that even a layman can use it easily. The user has to login into the website with his unique id and password which is already stored in database.

After the login, the user needs to enter the values like his age, Gender and there are some series of questions in which he has to respond with yes or no. The questions are whether the patient is sick or whether he has undergone any thyroid surgery previously, does he have thyroxine, Do you have Goitre, Hypothyroid according to reports and Hyperthyroid according to reports. The most important parameters are TSH(Thyroid Stimulating Hormone), T3(Triiodothyronine), TT4(Total Thyroxine), FTI(Free Thyroxine Index). If there is an increase in FTI value, the patient is having more risk of getting Thyroid disease. We even suggest that foods to eat and foods to avoid according to the amount of thyroid present in the person We also suggest medicines to the patients for a speed recovery.

4. ALGORITHMS

SVM:

A **support vector machine** (SVM) is machine learning algorithm that analyzes data for

classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is 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.

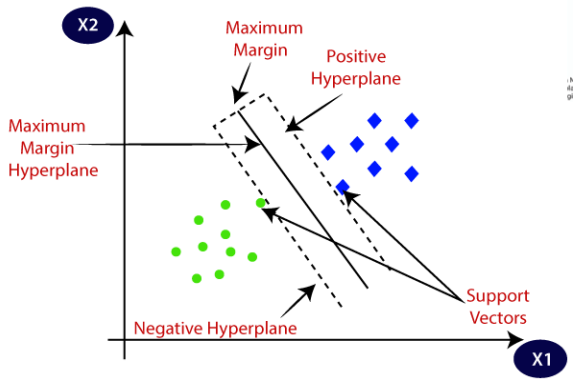


Fig.3: SVM model

RANDOM FOREST CLASSIFIER:

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. Regression will take all the mean, median of the output. It depends on the distribution of the output how the decision tree is given. Random forest is also known as random decision forest which belongs to the category of ensembled methods. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

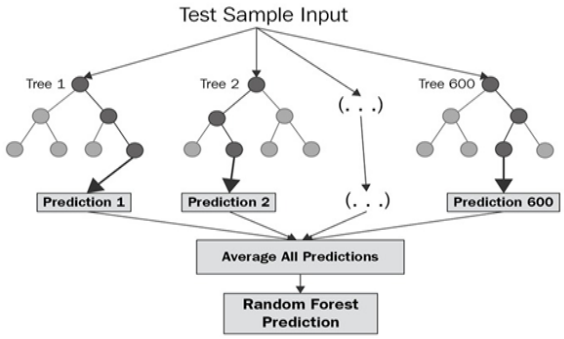


Fig.4: Random forest model

XGBOOST:

XGBoost provides a wrapper class to allow models to be treated like classifiers or regressors in the scikit-learn framework. The XGBoost model for classification is called XGBClassifier. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

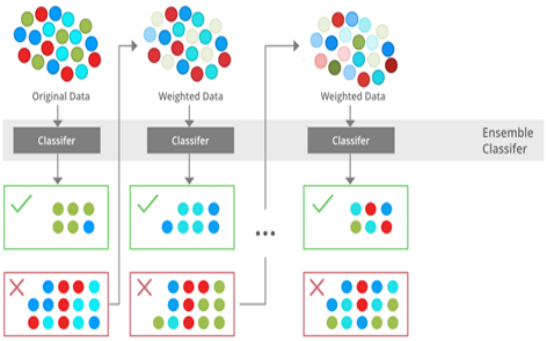


Fig.5:Xgboost model

KNN CLASSIFIER:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithmK-NN algorithm can be used for Regression as well as for Classification.

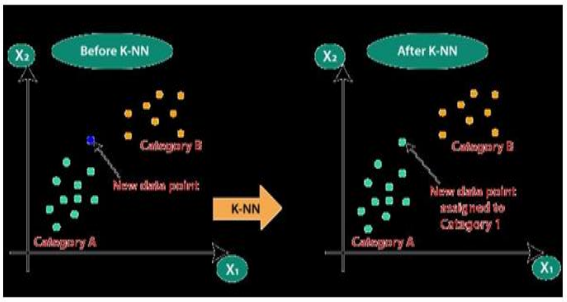
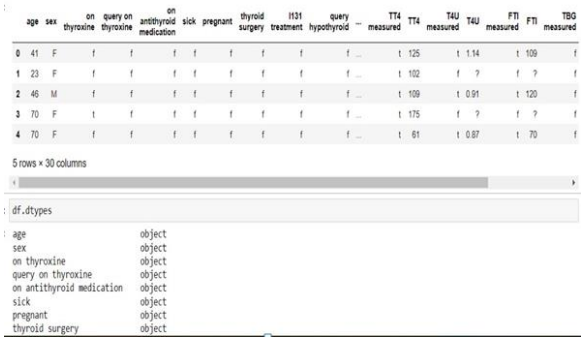


Fig.6: KNN model

5. EXPERIMENTAL RESULTS



	age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	I131 treatment	query hypothyroid	...	TT4 measured	TT4	T4U measured	T4U	FTI measured	FTI	TBG measured
0	41	F	f	f	f	f	f	f	f	f	...	t	125	t	1.14	t	109	f
1	23	F	f	f	f	f	f	f	f	f	...	t	102	f	?	f	?	f
2	46	M	f	f	f	f	f	f	f	f	...	t	109	t	0.91	t	120	f
3	70	F	t	f	f	f	f	f	f	f	...	t	175	f	?	f	?	f
4	70	F	f	f	f	f	f	f	f	f	...	t	61	t	0.87	t	70	f

5 rows × 30 columns

df.dtypes

age	object
sex	object
on thyroxine	object
query on thyroxine	object
on antithyroid medication	object
sick	object
pregnant	object
thyroid surgery	object

Fig.7: Dataset

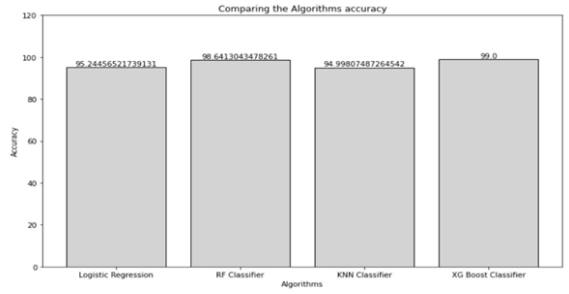


Fig.8: Data visualization

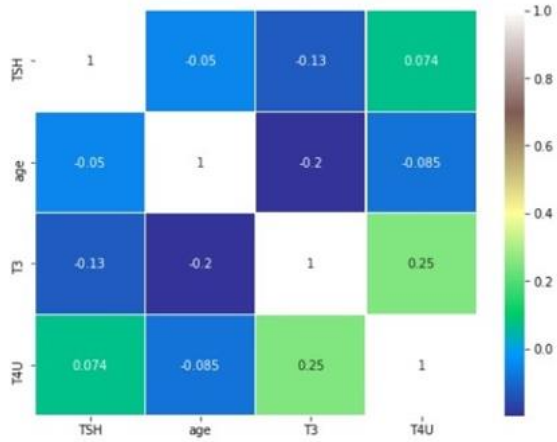


Fig.9: Correlation matrix

Correlation map is a 4*4 matrix which is in symmetric manner which indicates that all the diagonal values will have the same values and left half of the diagonal is a mirror reflection of the right half. Terrain defines the colour of the map and annotation indicates that the values are True and visible. Gcf function helps to plot the values In an order.

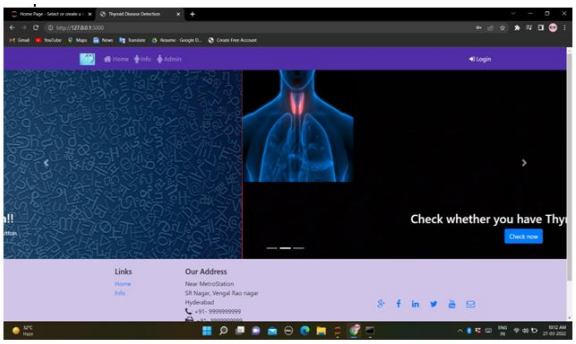


Fig.10: Home screen

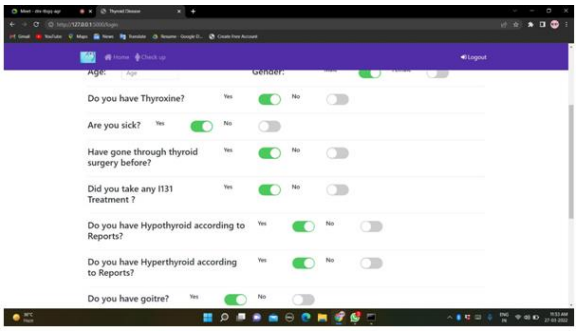


Fig.11: User report

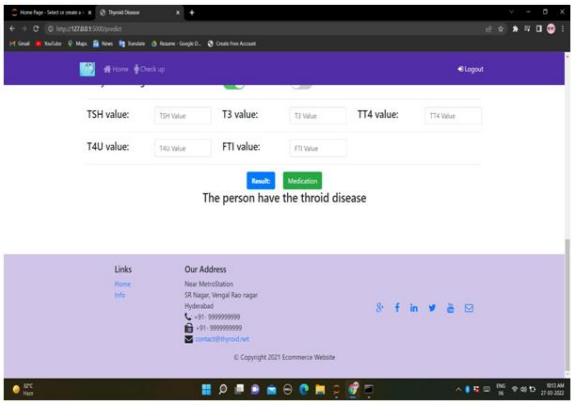


Fig.12: Disease detection screen

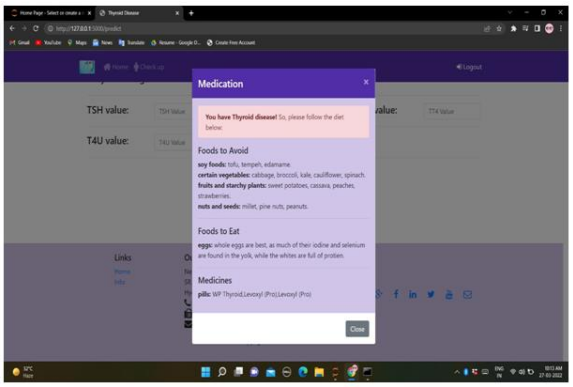


Fig.13: Medication screen

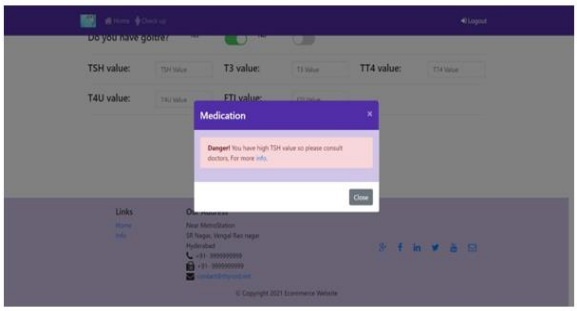


Fig.14: Medication output

6. CONCLUSION

Thyroid Detection using Machine Learning is a project idea that aims a smart and precise way to predict thyroid disease. We have made use of logistic regression algorithm to train our dataset and to predict thyroid disease with more accuracy. Here the machine is trained to detect whether the person normal, hyperhypothyroidism based on the user's input. So when user enters data in web app the data will be processed in backend (model) and the result will be displayed on the screen. Our objective was to give society an efficient and precise way of machine learning which can be used in applications aiming to perform disease detection..

7. FUTURE SCOPE

Further development can be do by using image processing of ultrasonic scanning of thyroid images to predict thyroid nodules and cancer, which cannot be recognized in blood test report. By combining both the results, thyroid disease prediction can cover all thyroid related diseases..

REFERENCES

[1]Ankita Tyagi and Ritika Mehra. (2018).“Interactive Thyroid Disease Prediction System using Machine Learning Techniques” published on ResearchGate.

[2] YongFeng Wang,(2020). “Comparison Study of Radiomics and Deep-Learning Based Methods for Thyroid Nodules Classification using Ultrasound Images” published on IEEEAccess.

[3] Sunila Godara,(2018). “Prediction of Thyroid Disease Using Machine Learning Techniques” published on IJEE.

[4] Hitesh Garg,(2013). “Segmentation of Thyroid Gland in Ultrasound image using Neural Network” published on IEEE.

[5] L. Ozyılmaz and T. Yıldırım,(2002). “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, pp. 2033–2036).

[6] K. Polat, S. Sahan and S. Gunes,(2007) “A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis,” Expert Systems with Applications,(vol. 32, pp. 1141-1147).

[7] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab,(2009) “Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM,” in 3rd International Conference on Bioinformatics and BiomedicalEngineering. ICBBE 2009.

[8] G. Zhang, L.V. Berardi,(2007) “An investigation of neural networks in thyroid function diagnosis,” Health Care Management Science,1998, (pp. 29-37.)

[9] V. Vapnik,(2012).Estimation of Dependences Based on Empirical Data, Springer, New York.

[10] Obermeyer Z,(2016). Emanuel EJ. Predicting the future— big data, machine learning, and clinical medicine. N Engl ; (375:12161219).

[11] Breiman L.(2001) StatisticalModeling: the two cultures.Stat Sci. ;16:199-231..

[12] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. (2017) Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. ; 9:245-250

[13] S. Godara and R. Singh,(2016) "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, (Vol. 910).

[14] Sunila, Rishipal Singh and Sanjeev Kumar.(2016) "A Novel Weighted Class based Clustering for Medical Diagnostic Interface." Indian Journal of Science and Technology (Vol 9).