Turkish Journal of Computer and Mathematics Education Vol.11 No.03 (2020), 2135-2141
DOI: https://doi.org/10.17762/turcomat.v11i3.13611
Research Article

An Approach to Diagnosis of Thyroid using Data Mining Techniques

Preeti Chaudhary
Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002,

Abstract
The occurrence of thyroid as disease tends to affect the physical health of people across the world. The disease is expected to be present in the endocrine glands of a human body and appears to be in the anterior position with respect to the neck. An imbalanced production of hormones from such glands results into a dysfunction of thyroid glands due to insufficient generation of such hormones. Hence, the glands begin to swell and might lead to malignant tumours. One of the available treatments to for such a disease is being done using sodium levothyroxine; commonly referred to as LT4. It is known to be hormone for treating thyroid and its respective disorders. In addition to the treatments being made; it is necessary to predict and identify the occurrence of the disease in a human body so that it can be cured at the right time in the early stages. For this to occur; endocrinologists must detect the imbalance being created with respect to the generation of thyroid hormones. Hence, the proposed research work presents the detection and identification of the same by utilizing the concepts of data mining techniques. A database consisting of 857 patients is acquired from a repository and pre-processing stages are applied to it to refine and filter the collected data. Four data mining algorithms are further used to evaluate the accuracy thus produced and the algorithm with highest generating accuracy is thus declared as the optimised model. For the implementation of the proposed research paper; the execution of random forests generated highest accuracy of 81.26 percent.

Keywords: data mining, diagnosis, thyroid, LT4, random forests

Introduction

The occurrence of thyroid is observed to be present in the endocrine glands while functioning around the anterior position of the neck. The hormones responsible to cause a misbalance in such a case are referred to as FT3 and FT4. Such hormones are not generated in sufficient quantities and thereby lead to release in the blood flow which causes the area around the neck to swell. The release of such hormones in the human body might also trigger other medical ailments such as; it might fluctuates BP levels, pulse rates, alter the human metabolism and increase the temperature of the body. In addition to this; the nutrients of the human body are also absorbed and does not leads to the organs for further nourishment. The occurrence of this ailment might also sometimes result into hyperthyroidism or hypothyroidism. A hyperthyroidism is scenario; wherein large number of such hormones are produced which might in turn lead to an increase in weight of the human body [1]. On the other hand, presence of hypothyroidism is triggered by less production of such hormones which can eventually lead to other medical ailments and disorders such as inflammation and swelling. If the occurrence of the disease is not detected at the right stage or through the early stages while the significant symptoms of thyroid may appear; the disease might even lead to malignant tumours. Thereby; making the entire process of disease diagnosis; a critical issue.

One of the commonly used treatments to cure thyroid is referred to as LT4; which is often used as an indication so as to comprehend the levels of hormonal imbalance caused in the human body due to secretion of hormones from the endocrine glands. However; following are the major cause that might result into the occurrence of the disease:

- Goiter
- Relapse of goitre or its partial removal from the endocrine glands
- Therapy used to replace hypothyroidism
- Consumption of anti-thyroid drugs
- Inflammation and swelling caused around the neck
- Detection of malignant tumour
- Lack of creation of hormones

Identifying the root cause of the disease and further using an LT4 might prevent the occurrence of thyroid; but in majority of the cases the relapse of thyroid occurrence is majorly dependant on the amount of residuals thus present in the human body and the capacity of the human body to fight back the residuals. This must be done so

as to balance the hormonal levels of thyroid secretion from the endocrine glands. In order to overcome this; an appropriate amount of LT4 dose is given to the patient and his physiological change are therefore observed. A continuous monitoring of hormonal changes is monitored and their LT4 doses are altered accordingly [2]. This leads to adjustments in the dosage which are supported by endocrinologist so that the patient's health can be improved with every dosage.

However, through the conduction of research studies by multiple scholars and medical experts it has been observed that the detection and diagnosis of medical ailments such as that of heart disease, diabetes, lung cancer etc. the overall time and complexities required are comparatively more. Such complexities might result into a delay or lag in the diagnosis of the disease and might worsen the condition of the patient. Hence, the primary aim of the work thus presented; is to detect the early occurrence of thyroid in a human body so that his physical parameters could be monitored from the right stage and a proper screening of his clinical data can be performed. This is also done so the appropriate level of LT4 dosage adjustments can be executed in accordance to the level of hormonal release in the human body of a thyroid patient. The research paper; is thus based on the collection of database of 857 thyroid and healthy patients and further classify them as thyroid positive or negative. For this reason; the stages of data pre-processing, filtering, removing and extraction are performed so that only respective features are thereby extracted. In the next stage; four data mining algorithms are executed in order to determine their respective accuracies. On obtaining their precision factors; a comparative analysis amongst the implemented algorithms is performed and the model with the highest generating accuracy is declared to be an optimised model.

Literature Survey

Multiple research works have been performed in this domain; wherein the authors have contributed their work in order to detect and diagnose the occurrence of thyroid in a human body at the right stage. For this purpose; they have taken into account various physical parameters of the body along with historical data of the patient. Authors [3] used machine learning models to detect the same by taking into consideration various hormonal factors of a human body. For this purpose; they obtained the dataset from the UCI repository consisting of 1124 thyroid patients. The dataset had two csv files; for train and test each. The aim of the research scholars was to classify thyroid positive patients from that of thyroid negative patients. A collective decision of classifying the obtained dataset as thyroid infected and thyroid healthy patients was taken. The author also used data mining techniques to diagnose the same. The algorithms thus used involved the implementation of KNN, Naïve Bayes, SVM and random forests. On conduction of the experimental analysis it was observed that the model with the execution of SVM generated an accuracy of 86.36 percent and was therefore declared as the optimised model.

In another research work by authors in [4] they used the implementation of neural networks to detect the same. The implementation was followed by MLP wherein the neurons involved in the hidden layers were used to diagnose the same. For this purpose; the authors used the dataset obtained from UCI repository with a sample size of 563 thyroid positive instances. The usage of neural networks helped to assign respective weights to the neurons which were further assisted by assigning respective biases. The adjustments were however made; and the neuron was made to trigger using feed forward neural network. The trigger caused by weight adjustments led to the generation alterations in the activation function. Adam was used as the optimiser and ReLu as the activation function. In addition the implementation of MLP; various data mining algorithms such as KNN, SVM and decision trees was also implemented. A comparative analysis of the overall system model was thus performed and their accuracies were recorded. It was observed that the implementation of MLP as a feed forward based neural network generated the highest accuracy of 89.63 percent.

In a different work proposed by authors in [5] they used multiple data mining techniques to diagnose the same. However their implementation was observed to focus on establishing a relationship between TSH, T3 and T4 hormonal levels in the human body. Based on the characteristic's thus observed from these levels; the author proposed to classify various types of thyroid such as hypo and hyper. For this reason; a dataset from Kaggle repository was obtained and three neural networks and four data mining algorithms were used. The dataset consisted of 865 total instances thereby comprising of thyroid positive and negative events. The two files of train and test were present in the Kaggle repository as csv files. The patient files comprised of their health information with regards to their pathological data such as hormonal levels, BP levels, pulse rate, BMI etc. On

the basis of such parameters various decision were taken by clinical experts in the initial stages. The obtained results were further combined and integrated with the dataset and further sent for pre-processing stages. In this stage; unnecessary involved data was henceforth removed and discarded and the final dataset was sent for training and testing phase. SVM, KNN, decision trees and random forests were used as data mining techniques and LSTM, CNN, RNN were used as neural network based algorithms. On conduction of experimental analysis; it was observed that the execution of CNN generated the highest accuracy of 91.63 percent.

Methodologies Used

This section of the research paper summarizes the data mining algorithms thus used to implement the proposed work.

A. Decision Trees

The concept of decision trees, which are utilised most frequently in the processes of classification and regression, is typically represented visually as a tree form. This tree-like structure provides a description of all of the instances that are there, and it does so by using the attributes of the executing model. The structure of the decision tree is composed of leaf nodes, which are distinct from single nodes. In the event that the image sample of the running model normally corresponds to the same class, the node that is connected to it will become the leaf. In later phases, it will be necessary to make decisions based on the properties of individual nodes, which will each correspond to a distinct branch of the decision tree. In this particular investigation, the primary objective of the decision tree is to anticipate the target class by making use of the judgements obtained from earlier branches. This is accomplished through the usage of internodes as well as nodes [6].

At each level of implementation, the decision tree chooses a node by doing a cost-benefit analysis on the information gain associated with each of its feature attributes. One of the most fundamental and frequently employed approaches to machine learning is the utilisation of decision trees. By merging the ideas of classification and regression, this method is largely utilised to solve issues pertaining to the classification process. On the other hand, decision trees are examples of supervised learning. They have a structure similar to a tree and consist of a root node that represents the entire population of samples, internal nodes that show dataset properties, branches that illuminate algorithmic rules and leaf nodes that reflect the result. The node that is connected to the node that eventually becomes the leaf is determined by whether or not the running model's image sample usually falls into the same class. Further phases of the process make use of the attributes of the nodes, which each individually represent a branch on the decision tree, to inform their conclusions. The primary purpose of the decision tree is to make an estimate of the target class by applying the outcomes of decisions made in earlier branches. This is accomplished with the help of internodes and nodes. So, this classifier may be seen as a prediction model that is based on machine learning that constructs and displays a relationship between the dataset, the values, and the features. There are two distinct kinds of nodes that make up the structure of the decision tree. These are the leaf nodes and the single nodes.

B. Random Forest

An example of an ensemble algorithm is the Random Forest Algorithm, which is essentially a collection of decision trees. When these several decision trees are put together, it's possible to draw a conclusion. The base of the decision tree is represented by a leaf node. It's possible for a decision tree to have a lot of leaf nodes. The action to take is determined by the decision tree according to the options available at each leaf node. On the other hand, if there are multiple decision trees to choose from, the leaf nodes of each tree are grouped together as seen in the image to the right. This hybrid method is eventually referred to as a "random forest classifier" because it is also utilised to produce a result. This phrase is used to define the hybrid approach. The Random Forest Classifier has an inherent advantage over other classification algorithms due to the fact that errors in one component of the algorithm do not impact the performance of the other components [7]. Thus, by utilising this strategy, all of the findings are evaluated before the design and production of the finished product. As the name suggests, Random Forests are composed of a network of interconnected trees that, when brought together, form a single model framework. The developing trees are initially the most important component due to the fact that they are laboured on and then collected by making the most of its capabilities. When additional data are added after the integration of all of the trees has been completed, the performance of the final output improves. At the later phases, when all of the decision trees have been combined, the categorisation component moves in much

closer to the object it is attempting to classify. When all of the substitutions are taken into account, the phase of training and testing that comes after the execution seems to have a greater degree of precision. As more data from the testing phase becomes available, there will be an increase in the total number of trees in the forest. After then, the test-train split methodology is used to the entire dataset due to the fact that it makes use of a wide variety of forest trees in this way.

C. SVM

The implementation of a Support Vector Machine (SVM), which is primarily used for the purposes of classification and regression and is therefore carried out in accordance with the purpose of implementing the respective system model, is one of the machine learning algorithms that is commonly used. SVM is primarily used for the purposes of classification and regression. However, the majority of the time it is utilised for the purpose of classification problems. This involves charting each vector that is obtained and assigning it a position in an n-dimensional space so that related characteristics may be recovered from the dataset repository [8]. So, the plane on which the corresponding curves and vector values are plotted has a significant impact on how SVM is implemented. The implementation of a traditional SVM, on the other hand, entails a segregation point where two classes are further segregated and the planes are connected to one another via corresponding lines drawn in between the plotting objects, which are acquired from the values so calculated.

D. KNN

K-Nearest Neighbor is commonly referred to by the acronym KNN, and it is used to create predictions using the dataset that is thus made available. Based on how near the generated clusters are to the instances so acquired, the dataset's predictions are made. Yet, the occurrences of thyroid incidence are identified by averaging the variables so calculated using the results of variable classification [9]. By varying the values and clustering them with the occurrences that have the highest likelihood of being relevant, this variable classification is achieved. All relevant examples are used on the training dataset in the following stage, and the new distance is then calculated [10].

Proposed Methodology

The primary aim of the study is to diagnose the presence of thyroid amongst patients and further classify them as thyroid positive or thyroid negative. For this purpose, a dataset was initially obtained and underwent the stage of pre-processing. A stage of normalizing the images of thyroid instances occurred in this step wherein the images were further augmented so as to increase the sample size of the dataset. The process of augmentation was however performed in two stages of training and testing. The obtained model was finally executed on four data mining algorithms and compared with the execution of AlexNet. In the pre-processing stage; the criteria of filtering the raw data was performed and redundant data was discarded. In addition to this; missing and NULL values were replaced with their respective binary values obtained through normalization. An entire column with such values was filled and data visualization was done in order to determine the total number of positive and negative instances of thyroid occurrence. A total of 857 patient instances was worked upon and further categorised as positive or negative by labelling them as 0 or 1. The data samples were finally trained on the respective algorithms and evaluated for accuracy. The diagram below depicts the architectural flow of the same:

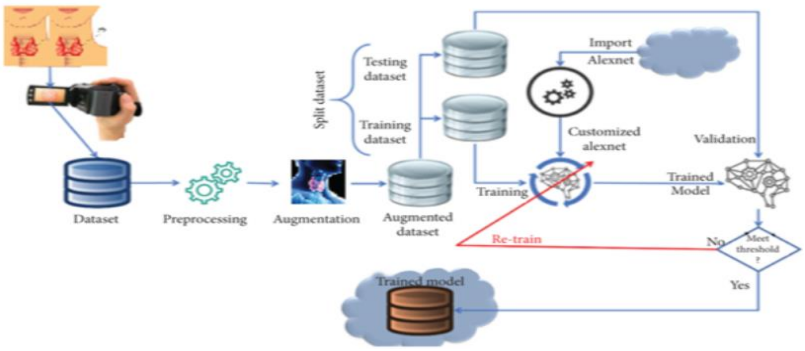

Figure 1: Architectural flow of the study

After the occurrence of pre-processing the dataset; the data is split into a ratio of 70 and 30 wherein 70 percent of the data is used for training purpose and 30 percent of the data is used for testing purpose. After the process

of binarization wherein the instances are labelled as 0 and 1; the samples of thyroid are classified as either thyroid positive or negative. A total of 293 instances were labelled as thyroid positive and 564 instances were labelled as thyroid negative. A visualization of this data is represented in diagram below:

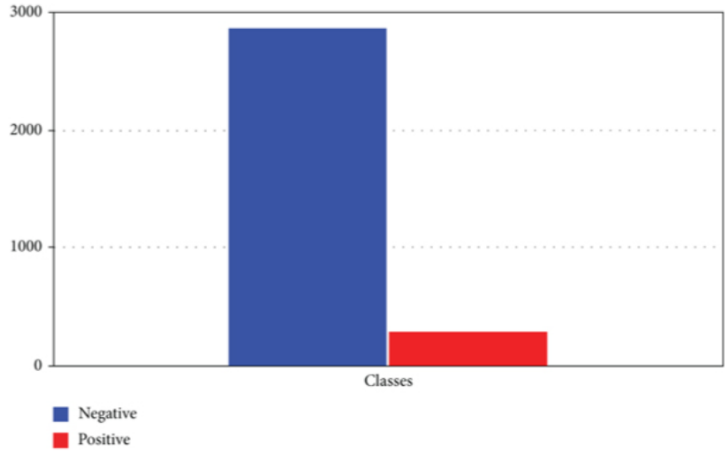

Figure 2: Data Visualization of thyroid negative and positive instances

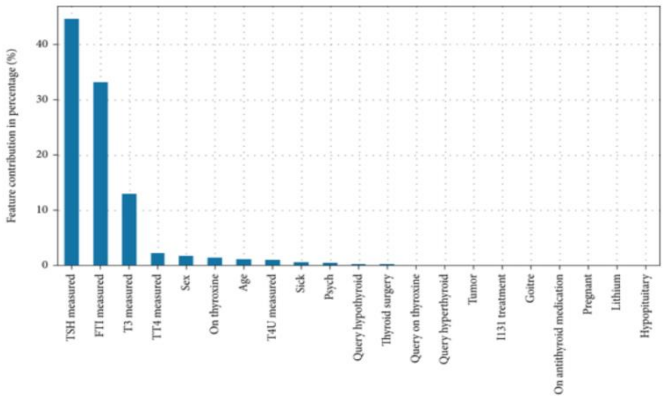

Figure 3: Feature Extraction from the dataset

Results

Once the dataset is trained and tested; it is further sent for evaluation. This is done in order to determine which algorithm generated the highest accuracy. For the evaluation purpose of the proposed research paper; sensitivity and specificity; along with values of true positive and negative rates are calculated. The figure below represents a classification of accuracies thus obtained through the implementation of the study:

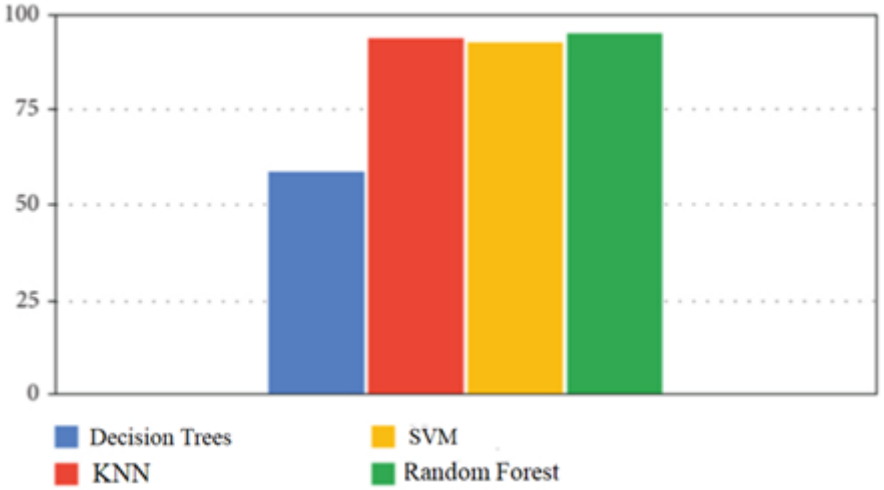

Figure 4: Performance of Classifiers

The table below depicts the values of sensitivity and specificity thus obtained:

Table1: Comparison of classifiers

Classifier	Sensitivity	Specificity
(1) Decision Trees	59%	91%
(2) KNN	94%	81%
(3) SVM	93%	78%
(4) Random forest	94.8%	91%

Conclusions

The implementation of the study thus implies that the usage of data mining techniques can thus be used in order to diagnose diseases in the medical field. On evaluation of the proposed research work, it was observed that the execution of random forests thus generate the highest accuracy in detecting the occurrence of thyroid in patients. A total of 857 thyroid instances were calculated and further labelled as thyroid positive or negative.

References

[1] A. Shrivas and P. Ambastha, "An ensemble approach for classification of thyroid disease with feature optimization," International Education and Research Journal, vol. 3, no. 5, pp. 1–4, 2019

[2] A. Dewangan, A. Shrivas, and P. Kumar, "Classification of thyroid disease with feature selection technique," International Journal of Engineering & Technology, vol. 2, no. 3, pp. 128–133, 2016

[3] A. Begum and A. Parkavi, "Prediction of thyroid disease using data mining techniques," in International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 342–345, Coimbatore, India, 2019

[4] J. H. Moon and S. Steinhubl, "Digital medicine in thyroidology: a new era of managing thyroid disease," Endocrinology and Metabolism, vol. 34, no. 2, pp. 124–131, 2019

[5] C. Ma, J. Guan, W. Zhao, and C. Wang, "An efficient diagnosis system for Thyroid disease based on enhanced Kernelized Extreme Learning Machine Approach," in International Conference on Cognitive Computing, pp. 86–101, Cham, 2018

[6] P. Poudel, A. Illanes, M. Sadeghi, and M. Friebe, "Patch based texture classification of thyroid ultrasound images using convolutional neural network," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5828–5831, Berlin, Germany, 2019

[7] Efficient Research on the Relationship Standard Mining Calculations in Data Mining SN Popat, YP Singh Journal of Advances in Science and Technology| Science & Technology 14 (2)

[8] Privacy Conflicts Detection and Resolution in Online Social Networks S Nilesh International Journal of Innovative Research In Computer and Communication Engineering

[9] K. Chandel, S. Veenita Kunwar, T. C. Sabitha, and S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and naive Bayes classification techniques," CSI Transactions on ICT, vol. 4, no. 2-4, pp. 313–319, 2016

[10] Ioniţă, Irina, and Liviu Ioniţă. "Prediction of thyroid disease using data mining techniques." BRAIN. Broad Research in Artificial Intelligence and Neuroscience 7.3 (2016): 115-124