The Role of Machine Learning in Thyroid Cancer Diagnosis

Monika D. Kate(✉) and Vijay Kale

Dr. G. Y. Pathrikar College of Computer Science and Information Technology, MGM University, Aurangabad, Maharashtra, India
monu.kate10@gmail.com, vkale@mgmu.ac.in

Abstract. Thyroid cancer occurs in the thyroid gland cells. This is butterfly shaped gland which is situated at lower part of neck. Thyroid cancer might not cause any symptoms initially. But as the cancer grows, it causes pain and swelling in neck. To detect and classify abnormalities of the thyroid gland Ultrasound imaging is mostly used. Computer aided diagnosis (CAD) help healthcare sector to increase the diagnosis accuracy and to reduce biopsy ratio. Machine learning techniques play a vital role in diagnosing diseases from the medical database. Various diseases can be predicted early using machine learning techniques. In this study, many previous research works are reviewed which use machine learning techniques to predict thyroid cancer. Machine learning classification techniques like Decision tree Classification, Random Forest Classification, Naïve Bayes, Kernel SVM, K-Nearest Neighbours, Support vector Machines (SVM), Logistic regression etc. are reviewed from the research papers. A short description about machine learning and segmentation is also presented.

Keywords: Machine learning (ML) · SVM · Segmentation

1 Introduction

Detecting cancerous cell(s) as quickly as possible can potentially save millions of lives. In Thyroid Cancer, the survival rate increases significantly with early detection and treatment of cancer. Early detection is beneficial for both, patients as they do not need to go under expensive treatment of benign tumors and for the doctors as early detection will help them in giving proper treatments to their patients and necessary care would be taken of them. Detection of cancer is done using various techniques such as image processing, deep learning, artificial intelligence etc. Image processing plays an crucial role in the fields like image mining, medical imaging, medical image processing. Image processing technique performs operations on an image to extract information from it or enhance it. In Pre-processing stage unwanted data (noise signals) are removed from ultrasound images. Then as per selected segmentation method, the suspected region is extracted where chances of cancer presence are more. With the help of extracted

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 276–287, 2023.
https://doi.org/10.2991/978-94-6463-136-4_25

features of segmented area, we can detect and classify whether the nodule is malignant or benign. The multiple steps are performed on collected thyroid ultrasound images before detection of disease. At initial stage ultrasound image dataset is collected as an input to the machine learning algorithm. In the next step by using segmentation, image is divided into different segments to zoom in the affected area. After that features are extracted from these segments. Next, the relevant features are selected to build a model. Finally, the appropriate classifier is used to classify the extracted data and make prediction based on this classification. These steps are used in every experiment of machine learning.

2 Methodology

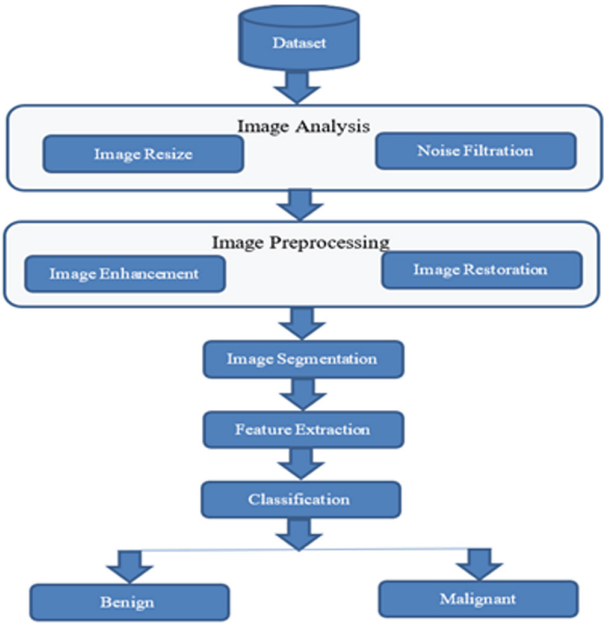

Fig. 1. Diagram shows the methodology which is used to detect thyroid cancer.

Based on literature review, the workflow of our work begins with the data collection. Data collection is the very first step of developing a predictive model. Accuracy of predictive model depends on the quality and quantity of collected dataset. Then we begin with dimensionality reduction. Dimensionality reduction technique is the way of converting higher dimension dataset in to preferred lesser dimension dataset (Fig. 1). In the next stage our total data get divide in to two parts, Training data and Testing data. 80% of our total data utilize for training of the model and remaining 20% of the data get used for testing the accuracy of the model. Moving in next phase, here we choose appropriate model based on our so far processed data and the features that are considered for model building. In the final phase, we train and evaluate the performance of our selected model.

2.1 Reviewed Data Set

Digital database of Thyroid Ultrasound images, which are available from an open-source scientific community. Below are the details of the sources:

1. DDTI: Thyroid Ultrasound Images available on cimalab.unal.edu.co, University of Columbia.
   - Here 400 Ultrasound Images of Thyroid glands available for our research.
2. Thyroid Disease database available on Kaggle.com by Ross Quinlan, Garavan Institute.
   - Here 3700 records of Hypothyroid with multiple attributes are available for research.
3. Hypo/Hyperthyroidism dataset available on Kaggle.com by Dario Madarino.
   - Here 9 datasets about hypo and hyper thyroid test data are available for research.
4. Thyroid Ultrasound Images dataset on UCI's machine learning repository.

2.2 Segmentation

The main aim of image segmentation is to make image more meaningful and easier to analysis. In digital image processing, image segmentation is a method using which we can divide digital image into different subgroups called segment according to their features and properties. In initial step of image analysis, the Image segments are located which has similar attributes. In simple words, Image segmentation is a method of assigning labels to every pixel in image. Common label is assigned to the pixel which belongs to the same category. These labeled data is further used in machine learning algorithm. There are different types of segmentation which are classified as follows:

1. Approach Based Classification
   (a) Similarity Detection (Region Approach)
   (b) Discontinuity Detection (Boundary Approach)
2. Technique Based Classification
   (a) Structural Segmentation Techniques
   (b) Stochastic Segmentation Techniques
   (c) Hybrid Techniques

Image segmentation can be performed by below mentioned approaches:

1. Thresholding Image Segmentation
2. Edge-Based Image Segmentation
3. Region-Based Image Segmentation
4. Watershed Image Segmentation
5. Clustering-Based Segmentation

Thresholding. Thresholding, in this image segmentation method we choose an appropriate constant value called threshold value(T). We compare every pixel value with this threshold value and divide image pixel into different regions. This method is mostly used to separate objects from background. Thresholding is the method which converts multi-level image (grey scale) into binary image.
Different thresholding technique are:

- Simple Thresholding
- Otsu's Binarization
- Adaptive Thresholding

Edge-Based Segmentation. Edge-Based Image Segmentation method detects edges of different segments in an image. It reduces the size of image and filter out unwanted and unnecessary information so that we can focus on important structural properties of an image.
Edge based image segmentation classified into following categories:

- Search-Based Edge Detection
- Zero-Crossing Based Edge Detection

Region-Based Segmentation. Region based Image Segmentation method divides the image into a groups of connected pixels called regions which are having similar properties.
We can classify region-based segmentation into the following categories:

- Region Growing
- Region Splitting and Merging

Watershed Segmentation. Watershed segmentation is region-based method. In image processing, a watershed is a transformation on a gray scale image. A watershed algorithm would handle the image as if it was a topographic map.

Clustering-Based Segmentation Algorithms. A Clustering-Based Segmentation Algorithms, it is the process of grouping image segments into a different clusters which are having similar features. Some of popular clustering algorithms are K-mean clustering, Fuzzy C-Means.

2.3 Classification

Machine learning contribute immensely in the analyzing large amounts of medical data. Machine learning has large number of applications now a days. Computer systems use machine learning to perform a particular task based on trends and specific patterns without relying on a detailed instruction set. It focuses on making predictions and thus producing an output based on that prediction.

Using machine learning classification algorithm, we can predict early-stage diagnosis of disease and provide a solution. An early-stage diagnosis and treatment can save people life as well as it will minimize the death rate because of severe disease. Machine learning is a subpart of artificial intelligence (AI). Machine Learning (ML) is study of algorithms which learn through historical data and improves performance through experience. Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed to do so [28] (Fig. 2).

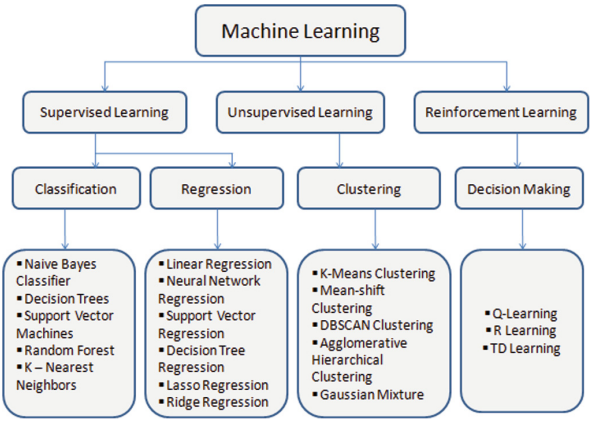

Fig. 2. Following figure shows categories of Machine learning algorithm.

Supervised Learning. Supervised learning is a type of machine learning method which is based on supervision. In this method the model is trained according to the labelled dataset and built a predictive model which predict the output for sample data (unseen data). Below are the two categories of Supervised learning algorithms:

- Classification
- Regression

Unsupervised Learning. Unsupervised learning is a type of machine learning method which is not based on supervision. In this method the model is trained according to the unlabeled dataset and are allow to act on that data without any supervision. This predictive model itself find the hidden pattern and predict the output for sample data (unseen data).
Below are the two categories of Unsupervised learning algorithms:

- Clustering
- Association

Reinforcement Learning. Reinforcement learning method is based on a feedback, in which a reinforcement agent gets a reward for each correct action and gets a penalty for each wrong action. The machine learns automatically with these feedbacks and improves its performance.

3 Literature Review

Prabal Poudel et al. [1]: have compared the three different machine learning techniques (SVM, ANN and RFC) for thyroid texture classification and segmentation. In this research two different dataset were used. Dataset 1 consists of total 675 2D thyroid ultrasound images with image size of 760X500 pixel. The Second dataset has 16 subjects and each subject has 156 to 289 2D thyroid US images. For Classification and segmentation of these 2D thyroid US images SVM (Support Vector Machine), ANN (Artificial Neural Network) and RFC (Random Forest classifier algorithms are used. By comparing the performance on same dataset and then on different dataset, comparison analysis was performed. It is observed that accuracy training of the classifier are similar, ANN perform comparatively better that SVM and RFC.

Gyanendra Chaubey et al. [2]: have compared widely used three algorithms namely logistic regression, decision trees and k nearest neighbor (kNN) algorithms to predict thyroid disease and evaluate their performance in terms of accuracy. In this study it is observed that KNN classifier has 96.875% accuracy.

Pushkar Sathe et al. [3]: have classify tumour into malignant or benign tumour using different features from several cell images. They have fed the machine with categorized data sets of malignant and benign cell images and for better results and have also shuffled these images within themselves and then have successfully trained the machine with malignant and benign data sets after which successful testing is performed with an unknown data set.

K. Shailaja et al. [4]: have reviewed various machine learning algorithms for prediction of various diseases. These algorithms are used for developing decision support for healthcare sector.

D. Selvathi et al. [5]: This study is focused on to develop an automatic system that classify thyroid images and segments the thyroid gland using machine learning algorithm. Thyroid ultrasound images with nodules and without nodules dataset was used. In this method mixed thyroid images (with normal thyroid region and thyroid nodules) Were classified using SVM (Support Vector Machine) and Extreme Learning Machine (ELM). During this experiment, the ELM segmentation method is observed with better accuracy than SVM segmentation method.

Ankita Tyagi et al. [6]: This research is about to predict thyroid disease using machine learning algorithm like SVM, K-NN, and decision tree. It is observed that Neural Network perform better over other techniques. The performance of Support vector machine and decision trees techniques are also observed good.

Vijay Vyas et al. [7]: The aim of this research was to develop a computer aided system to classify thyroid nodule as benign or malignant. Thyroid ultrasound

images of 99 patient were used in this research. The SVM and ANN classification algorithms were used as well as both algorithms were compared based on their accuracy score. From the experimental result it is observed that SVM outperformed ANN, SVM achieved 96% accuracy.

Jamil Ahmed Chandio et al. [8]: This paper addresses the classification problem of Medullary Thyroid Cancer and classifies malignant and non-malignant cells and tissues. The proposed methodology of this paper is based upon three layers. Layer 1: Image preprocessing, Layer 2: Classification Model, Layer 3: Result Visualization.

Shoon Lei Win et al. [9]: This research study is about to predict cancer recurrence from microarray gene expression data. Here machine learning based approach is used. Gene expression data set used as input to machine learning algorithm. The model will predicts whether a specific cancer may reappear within a specific time-frame. In this study three cancer recurrence data set of CNS cancer recurrence, prostate cancer recurrence and breast cancer recurrence were used.

Shaik Razia et al. [10]: In this research various Machine learning algorithms like Support Vector Machine (SVM), Multiple Linear Regression, Naïve Bayes and Decision Trees are used to diagnose thyroid disease. This comparative study results are compared and found that Decision Trees algorithm provided greater accuracy to diagnose thyroid disease.

Sonali Bhadoria et al. [11]: This research Proposed an automatic segmentation method using two tools, Analyze 10.0 and mazda for segmentation of thyroid US images and segmentation by applying specific algorithm. As well as comparative analysis is also done. They also provided a summary of all the results.

Reema Mathew A et al. [12]: In this review paper comparative study of different feature extraction technique is done for detection of brain tumor. From study it is observed that there are several hybrid approaches can be developed which may result in higher accuracy and better results.

Eystratios G. et al. [13]: This study is about to detect thyroid nodule using ultrasound images and videos. Author proposed a computer aided system, named TDN (Thyroid nodule detector). In this research SVM and KNN classification techniques were used, experimental result shows SVM classifier perform better than KNN.

Dhyan Chandra Yadav, Saurabh Pal [14]: The main aim of this research study was to detect thyroid disease using three different classification algorithms like Random Tree, J48 and Hoeffding. In this research dataset of 499 patient were used. Accuracy of individual algorithm is calculated. A new ensembled method were introduce by author and apply again on the same dataset which is used to calculate individual algorithm accuracy. From the experimental result it is observed that ensembled method provides better classification having 99.2% accuracy and 99.36% sensitivity.

Yijun Wu, Ke Rao et al. [15]: The main aim of this study is to develop machine learning based predictive model to detect Central lymph node metastasis (CLNM) which is occur frequently in patients with papillary thyroid cancer (PTC). In this study 1103 PTC patient data were reviewed. Six machine learning algorithm were applied in this study like AdaBoost (adaptive boosting), DT (decision tree), random forest classifier (RFC), ANN, extreme gradient boosting (XGBoost) and gradient boosting decision tree (GBDT). This study conclude GBT is best machine learning based model for the prediction of CLNM in PTC patient.

Jagdeesh saraf and Dr. Kalpana V. [16]: A computer based technique for segmenting and classifying the nodules as malignant or benign is proposed in this research paper. US thyroid image dataset were used in this research. Edge detection techniques were used to segment the image and ANN classifier were used to classify nodule. Thresholding and template matching shape based feature extraction method were used and textural features were extracted from US thyroid image.

Massoud Sokouti, Mohsen Sokouti and Babak Sokouti [17]: In this study, the important roles of detecting and diagnosing cold nodules has been discussed since they are known as high risk of infected by the thyroid cancer. The image preprocessing and processing techniques including image enhancement, image segmentation (thresholding) were applied. These techniques couldn't solely capable of determining the cold nodules in thyroid radioisotope images so, a hill climbing algorithm was applied in order to automatic determination of these regions as it was much simpler than other intelligent systems such as Fuzzy SVM, ANN, SVM and genetic algorithms.

M Kalaiyarasi et al. [18]: This study proposed a different machine learning classification algorithm to classify benign and malignant tumor. In this study breast cancer patient dataset is used. Machine learning algorithm like K-Nearest Neighbor, Logistic Regression, Support Vector Machine are used in this study.

Polepogu Rajesh, Kunduru Umamaheswari [19]: Objective of this paper is to provide a complete solution to diagnosis the suspicious thyroid region in the thyroid gland. An advanced method of completely automatic identification of thyroid nodule is proposed in this paper. Microscopic images undergo automated procedure to recognize a disorder in the thyroid. The proposed method shows good results and identify the thyroid nodule present in the image with good level of accuracy.

Fu-sheng Ouyang [20]: The purpose of this study was to compare the classification performance of linear and nonlinear machine-learning algorithms for the evaluation of thyroid nodules using pathological reports as reference standard. It is observed that the performance of both linear and non-linear machine learning algorithm are almost equal.

Lay Khoon Lee et al. [21]: This review paper gives the overview of segmentation method. It also gives the idea about which segmentation method should applied on digital images like MRI, CT scan, 3D MRI, X-Ray, and US images. It discuss about advantages and disadvantages of various segmentation methods.

Ahmet Akbaş et al. [22]: This research study use machine learning technique having aim to improve performance in diagnosis of thyroid cancer. In this research BayesNet, NaiveBayes, SMO, Ibk and Ran-dom Forest classifiers were used for classification. Using WEKA data mining program author compare these classifiers by using different measures like accuracy, Kappa, Matthews Correlation Coefficient (MCC), and ROC. Experimental result shows that Random Forest classifier performs better.

Yongfeng Wang et al. [23]: This paper compared the classification performance of radiomics and deep learning-based methods using ultrasound images with thyroid nodules. The comparison results show that the deep learning-based model shows better accuracy.

Chandan R et al. [24]: have proposed a machine learning based model to detect thyroid disease. In this research accuracy of individual ML algorithm were calculated, KNN provide 93.84% accuracy, SVM gives 95.38% accuracy, ANN provides 75.38% accuracy, Decision tree provides 92.3% accuracy, with Logistic Regression 96.92% accuracy is obtained. From the experimental reason it is observed that Logistic Regression outperformed other classifiers.

Rebecca Smith-Bindman [25]: This research study is about to find out the risk of thyroid cancer using thyroid ultrasound image. They used three US thyroid imaging attributes 1) Entirely solid composition, 2) microcalcifications, 3) size larger than 2cm on which they decide whether sample need to undergo for biopsy or not. This is helpful to reduce unnecessary biopsies.

K. Sumithra et al. [26]: This research study is about to presents an overview of different image processing techniques. It also provides the idea about image processing tool which researcher can use for image processing. Various image processing techniques has been review such as image recognition, image restoration, image enhancement and image segmentation.

Nikita Singh et al. [27]: In this research author has propose a segmentation method and also performed a comparative study of KNN, SVM and Bayesian classifier. Thyroid ultrasound images were used for this research. MATLAB version 7.7.0 software were used for image processing. Accuracy of individual classifier is calculated, and it is observed that SVM has 84.62% accuracy, KNN has 46.15% accuracy and Bayesian has 38.46% accuracy. From the result obtained from experiment it is observed that SVM shows best accuracy.

4 Conclusion

This review work provides the knowledge about segmentation techniques and machine learning algorithms. Multiple papers are reviewed for this study. Some researchers have compared the different machine learning techniques and its accuracies (KNN provide 93.84% accuracy, SVM gives 95.38% accuracy, ANN provides 75.38% accuracy, Decision tree provides 92.3% accuracy, with Logistic Regression 96.92% accuracy). Neural Network is observed perform better among other techniques. Along with Neural Network, the SVM (Supprot Vector Machine) and Decision tree techniques are also performing well.

This paper presents a systematic review study for prediction of various diseases. Various research papers study emphasize that Machine Learning techniques are very useful to determine the accurate prediction of disease. It also classifies tumor as malignant or non-malignant. In Future work, we plan to develop a predictive model which use machine learning algorithms to detect thyroid cancer at early stage with greater accuracy using US thyroid images.

References

1. Prabal Poudel: Thyroid Ultrasound Texture Classification Using Autoregressive Features in Conjunction with Machine Learning Approaches. IEEE Access vol 7 2019. https://doi.org/10.1109/ACCESS.2019.2923547.
2. Gyanendra Chaubey et al : Thyroid Disease Prediction Using Machine Learning Approaches: Published in Springer Link, https://doi.org/10.1007/s40009-020-00979-z.
3. Pushkar Sathe et al.: Cancer Detection using Machine Learning. In: International Research Journal of Engineering and Technology (IRJET)-Volume: 07 Issue: 09.
4. K. Shailaja et al.: Machine Learning in Healthcare: A Review : Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018). In: IEEE Conference Record 42487; IEEE Xplore ISBN:978-1-5386-0965-1.
5. D. Selvathi et al.: Thyroid classification and segmentation in ultrasound images using machine learning algorithms. In: Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN) 2011.
6. Ankita Tyagi et al.: Interactive Thyroid Disease Prediction System Using Machine Learning Technique. In:5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018),20–22 Dec, 2018, Solan, India. https://doi.org/10.1109/ICSCCN.2011.6024666.
7. Vijay Vyas Vadhiraj et al. Ultrasound Image Classification of Thyroid Nodules Using Machine Learning Techniques. In: Medicina (Kaunas). 2021 Jun; 57(6): 527. Published online 2021 May 24. https://doi.org/10.3390/medicina57060527.
8. Jamil Ahmed Chandio et al.: Decision Support System for Classification Medullary Thyroid Cancer. In: IEEE ACCESS publication. DOI 10.1109/ACCESS.2017.
9. Shoon Lei Win et al.: Cancer Recurrence Prediction using Machine Learning. In: International Journal of Computational Science and Information Technology (IJCSITY) Vol. 2, No. 2, May 2014.
10. Shaik Razia et al.: A Comparative study of machine learning algorithms on thyroid disease prediction: International Journal of Engineering and Technology, 7 (2.8) (2018) 315–319.
11. Sonali Bhadoria et al : Comparison of Segmentation Tools for Multiple Modalities in Medical Imaging -Journal of Advances In Information Technology, Vol. 3, No. 4, November 2012. https://doi.org/10.4304/jait.3.4.197-205.
12. A. R. Matthew, A. Prasad and P. B. Anto, A review on feature extraction techniques for tumor detection and classification from brain MRI, 2017 In: International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017, pp. 1766–1771 https://doi.org/10.1109/ICICICT1.2017.8342838.
13. Eystratios G. et al.: Thyroid Nodule Detection system for Analysis of Ultrasound Images and Videos. In: Journal of Medical Systems 36(3), 1271–81. https://doi.org/10.1007/s10916-010-9588-7.

14. Dhyan Chandra Yadav, Saurabh Pal et al.: Discovery of hidden pattern in thyroid disease by machine learning algorithms. In: International Journal of Pharmaceuticals and Health Care Research (IJPHR),11(1) 61–66, 2020.
15. Yijun Wu et al.: Machine learning algorithms for the prediction of central lymph node metastasis in patients with papillary thyroid cancer. In: Frontiers in endocrinology 11,816, 2020.
16. Jagdeesh saraf and Dr. Kalpana V. et al.: Thyroid Cancer Detection Using Image Processing. In: International Journal of Research and Scientific Innovation (IJRSI), Volume-IV, Issue-VIII, August-2017.
17. Massoud Sokouti, Mohsen Sokouti, Babak Sokouti et al.: Computer Aided Diagnosis of Thyroid Cancer Using Image Processing Technique. In: International Journal of Computer Science and Network Security, Volume 18- No.4, April 2018.
18. M Kalaiyarasi, R Dhanasekar et al.: Classification of Benign or Malignant Tumor Using Machine Learning. In: IOP Conf. Series: Materials Science and Engineering 995 (2020) 012028 https://doi.org/10.1088/1757-899X/995/1/012028.
19. Polepogu Rajesh, Kunduru Umamaheswari et al.: Thyroid Disorder Detection Using Image Segmentation in Medical images. In:(IJSDR), Volume 1, Issue 6, Ver. I (June 2016).
20. Fu-sheng Ouyang et al.: Comparison between linear and nonlinear machine learning algorithms for the classification of thyroid nodules. In: European Journal of Radiology Volume 113, April 2019.
21. Lay Khoon Lee et al.: A Review of Image Segmentation Methodologies in Medical Images. In: Published in Springer. https://doi.org/10.1007/978-3-319-07674-4_99.
22. Ahmet AkbaşPerformance Improvement with Combining Multiple Approaches to Diagnosis of Thyroid Cancer. In: Published in: Scientific Research. https://doi.org/10.4236/eng.2013.510B055.
23. Yongfeng Wang et al.: Comparison Study of Radiomics and Deep Learning-Based Methods for Thyroid Nodules Classification Using Ultrasound Images. In: Published in IEEE Access (Vol: 8).
24. Chandan R et al.: Thyroid Detection Using Machine Learning. In: International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 5, Issue 9, ISSN No. 2455–2143.
25. Rebecca Smith-Bindman et al.: Risk of Thyroid Cancer Based on Thyroid Ultrasound Imaging Characteristics. In: JAMA Internal Medicine October 28 2013, Vol. 173, Number 19.
26. K. Sumithra et al.: A Survey on Various Types of Image Processing Technique. In: International Journal of Engineering Research and Technology (IJERT) ISSN: 2278–0181 Vol. 4 Issue 03, March-2015.
27. Nikita Singh et al.: A Segmentation Method and Comparison of Classification Methods for Thyroid Ultrasound Images. In: International Journal of Computer Applications (0975–8887) Volume 50 - No.11 July 2012.
28. https://en.wikipedia.org/wiki/Machine_learning