IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 11, November 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

Thyroid Disease Prediction with KNN, Random Forest and Decision Tree: A Study

Er. Surbhi Jain[1], Dr. Simmi Dutta[2], Er. Sheetal Gandotra[3]

P.G. Student, Dept. of Computer Engineering, Govt. College of Engineering &Technology, Jammu, India[1]

Head of Dept, Dept of Computer Engineering, Govt. College of Engineering & Technology, Jammu, India[2]

Assistant Professor, Dept of Computer Engineering, Govt. College of Engineering & Technology, Jammu, India [3]

ABSTRACT: This research provides the in-depth analysis of mechanisms used to predict thyroid related diseases. Machine learning based mechanisms are most common towards prediction of thyroid related diseases. Empirical study consists of phases including pre-processing, feature extraction, training, and testing. Pre-processing used to remove the noise from the dataset. The noise handling mechanisms includes median filtering and gaussian filtering. The feature extraction mechanisms are employed to extract the attributes necessary for the training phase. The training phase performed to make the model learn about the symptoms and diseases present within the dataset. The last phase consists of classification. Classification phase generally consists of KNN, Random Forest and decision tree. Validation is accomplished with the help of metrics including classification accuracy, specificity, sensitivity, and F-Score. These metrics must be improved to generate the accurate result about the disease.

KEYWORDS: Thyroid related diseases, machine learning, KNN, random forest, decision tree

I. INTRODUCTION

Thyroid is a disease that affect the human body and causes a formation issue with thyroid gland that is important part of human body. Metabolism corresponding to body is adversely affected with thyroid related disease. Issues with thyroid hormone secretion causing two common types of disease including Hyperthyroidism and hypothyroidism. These two diseases cause imbalance in the regulation rate of body's metabolism. Machine learning based mechanism plays a critical role in the process of thyroid prediction. This paper discussed the models that takes the information from UCI machine learning repository. The dataset then fed into the system for training. After the training process, finalized testing is performed(U Sidiq, 2019).

Thyroid disease can affect the human body and severe issues can be a result related to metabolism as discussed in (Jha et al., 2022). The effective classification of disease using the machine learning approaches including decision tree is discussed by (I Ionita, 2016). The decision tree-based mechanism generally operates on dataset. Dataset can be derived from benchmarked websites including UCI machine learning repository, Kaggle and many more(Yasir HusseinShakir, 2020). The attributes of the dataset must be pre-processed to remove noise that ultimately increase the classification accuracy. Mode based approach as followed in (Chaubey et al., 2020a). In this approach, highest value from the attribute is replaced with the missing value. After this phase, segmentation is performed. Segmentation process divides the overall dataset into critical and non-critical parts(SM Gorade, 2017). The last phase is classification that actually predict the disease. Classification process includes neural network-based approach, decision tree, KNN, random forest and many more as discussed in (Chaubey et al., 2020b). The approaches followed and corresponding results are discussed in section 4. Machine learning based methodology used for the data collection of thyroidrelated disease is given in figure 1. This is a critical phase as rest of the phases depends upon this phase(Priyadharsini&Sasikala, 2022). Normalization process is used to bring the result in certain range for improving the process of calculations.

Rest of the paper is organized as under. Section 2 provides the in-depth study of approaches used for Thyroid related diseases. Section 3 gives the problem formulation that ultimately result in better techniques for prediction of thyroid related diseases. Section 4 gives the empirical study of the existing approaches and last section gives the conclusion.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

II. RELATED WORK

The work has been done towards detecting the disease caused by thyroid. This section is discussed in multiple parts. At first place, role of thyroid chemicals has been discussed and in the next section data mining mechanism used for prediction are discussed and at the end, comparative analysis of the techniques have been presented.

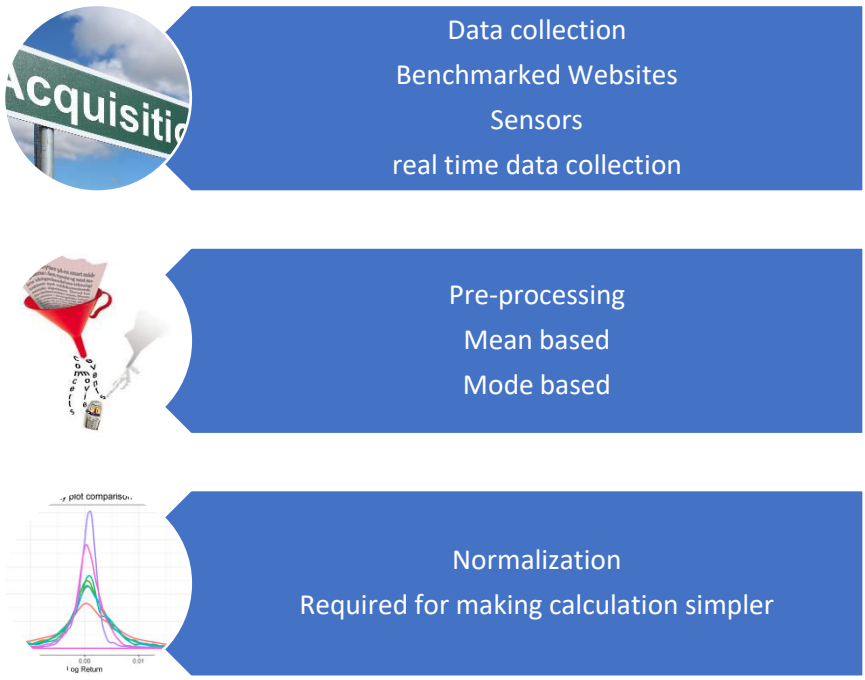

Figure 1: Data collection and pre-processing-based methodology

A. Role of Thyroid chemicals

It is widely believed that both hypothyroidism and hyperthyroidism can cause changes in the way that lipoproteins are organized as discussed in (TL Fonseca, 2013). Since a few key lipoprotein transport catalysts are controlled by thyroid hormones, it is not surprising that thyroid disease frequently causes a worsening of lipoprotein transport(B Gereben, 2015). By suppressing the rate-limiting protein of cholesterol amalgamation (hydroxymethylglutaryl coenzyme A [HMG CoA] reductase), intracellular free cholesterol stifles endogenous cholesterol production, enabling the phone to govern its own cholesterol level through a nearby input control framework. By activating the enzyme HMG CoA reductase, which catalyses the conversion of HMG CoA to mevalonate, thyroid chemical animates the hepatic cholesterol union once again. This results in a better cholesterol union in hyperthyroidism and a decreased one in hypothyroidism. (Chakraborty I 2010)

Nevertheless, because thyroid chemicals may also regulate the union and oxidation of LDL cholesterol, blood cholesterol levels are alternately increased in hypothyroidism and decreased in hyperthyroidism. Additionally, about 3% of thyroxine (T4) is linked to lipoproteins, mostly HDL (92%), and less so LDL (6.7%). (A Hood, 2000). The LDL receptor detects the T4-LDL complex, which functions as a credible tool for T4 entry into cells. Finally, a thyroid hormone activates the LDL receptor, increasing the fragmented catabolic speed of apoB without changing the rate at which it is produced. Plasma HDL concentrations have been described as normal or, conversely, decreased in hyperthyroidism and normal or even raised in severe hypothyroidism(AA Arriagada, 2015).

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

B. Data mining techniques used for Thyroid prediction

There are number of machine learning mechanisms that are used for the prediction of thyroid. The methodology corresponding to thyroid related disease prediction with machine learning is presented in figure 2

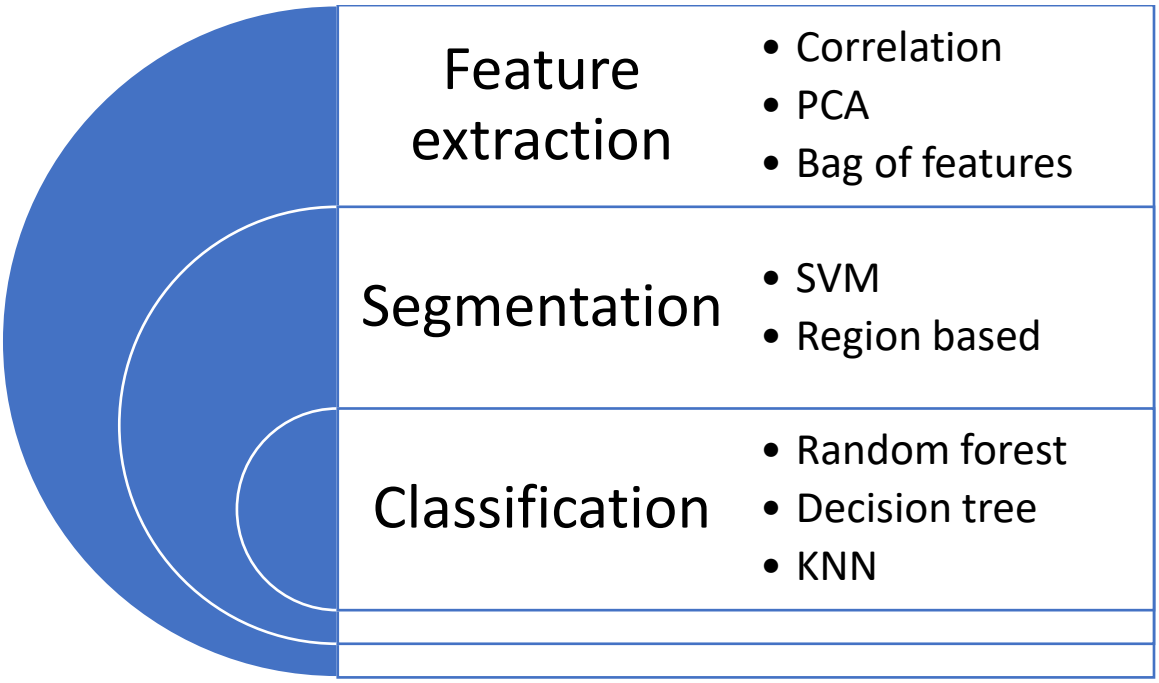

Figure 2: Data mining phases for the Thyroid prediction

Feature extraction is critical in the prediction of thyroid related disease. Correlation based feature extraction as discussed in (Snekhalatha&Gomathy, 2018a). Using correlation-based mechanism, highest correlated features or attributes from the ultrasound images are extracted. Correlation based mechanism as discussed in (Vanderpump, 2011) reduce the dimension to increase the execution speed of overall operation. The performance gain of almost 5% was observed in terms of execution speed through correlation-basedapproach(Yu et al., 2022). Furthermore, Principal component analysis reduces the dimensionality of overall operation(Mugasa et al., 2020). PCA concentrates on extracting the features from the dataset that can contribute towards the actual output. Using PCA may reduce the overall classification accuracy.

Bag of features is another important feature extraction mechanism that is graph based(Han et al., 2022). This mechanism allows effective extraction of features from the image even if feature is not that correlated with the result. Large and complex images may take large amount of time for feature extraction. To avoid the issue, back propagation mechanism with feed forward approach can be used(Snekhalatha&Gomathy, 2018b).

Segmentation is the next phase in the classification of disease. The segmentation process can be achieved with support vector machine. This mechanism is based upon the hyperplanes(Kumar, 2020). The support vector machine divides the dataset into critical and non-critical parts. The critical parts are evaluated for having thyroid disease. Each hyperplane within SVM is attached with class either 0 or 1. 0 means disease is absent and 1 indicates presence of the disease. SVM is commonly collaborated with convolution neural network for the Thyroid image segmentation as done in (Ma et al., 2017). The issue with this approach is limited classes available for prediction. The performance gain in terms of segmentation accuracy is achieved with limited classes using SVM. Another important segmentation mechanism is region based(Poudel et al., 2018). Region based approach identifies the critical regions to be retained that may be adversely affected with the thyroid disease. The issue with this approach is high degree of misclassification.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

Last phase is classification process. Classification can be accomplished with the help of KNN, Random Forest, and decision tree. KNN is the K nearest approach that has static K values(Bai et al., 2020). KNN approach is based upon Euclidean distance calculated to determine the nearest class for prediction. The issue is of static K values that must be dynamic for better prediction accuracy. Random forest-based approach for classification is another approach for thyroid related disease(Prochazka et al., 2019). This approach is much faster as compared to KNN approach for classification with the difference that classification depends upon hit and trial method. The classification accuracy fluctuates in this case and hence may not be suitable for complex and large datasets. Decision tree-based classification model is most common and simple in the prediction of Thyroid related diseases(Yadav & Pal, 2020a). Decision tree forms the branches(parts of disease) corresponding to the main tree(disease) and perform the classification by skipping the branches that do not correlate. The classification accuracy of decision tree corresponding to thyroid related diseases is better as compared to KNN and random forest-basedapproach.

C. Comparative analysis

This section presents the comparative analysis of different techniques used for thyroid related diseases. The comparative analysis is presented within table 1

Reference	Technique	Description	Accuracy
(Akbaş et al., 2013)	Multiple approaches including random forest, svm and knn	Ensemble based approach produced better classification accuracy in the prediction of thyroid	AdaBoostM1 + classifier: RandomForest: 99.8%
(Yadav & Pal, 2019)	Ensemble based approach in the detection of Thyroid within women	High classification accuracy along with F-score was achieved.	The seed value 35 and num-fold value 10 have found the highest accuracy using bagging ensemble techniques. DTC: 98%,RFC:99%,ETC:93%
(Liu et al., 2019)	CNN based mechanism for thyroid related disease: thyroid nodules	Binary classification using CNN	97.1 %
(Jia et al., 2020)	Correlation based thyroid disease: thyroid cancer prediction	High classification accuracy is achieved.	98%
(Zhu et al., 2021)	Generic deep learning based mechanism in the prediction of Thyroid related disease : cancer	Accuracy is high but execution speed is high.	89%

Table 1: Comparative analysis of techniques used for Thyroid disease prediction

III. PROBLEM FORMULATION

The mechanisms including support vector machine, KNN and random forest predict the thyroid however it is not possible to predict the after effect of the Thyroid. This means it is not possible to predict the future impact associated with the thyroid related diseases. Furthermore, correlation analysis can be introduced within convolution neural network to improve the process of the prediction. Pre-processing process also have the issue as the noise handling mechanism must accommodate background subtraction and contrast enhancement that is not considered in earlier work (Prasad et al., 2016)(Azar et al., 2013). The issue can be resolved using mode-based mechanism within the existing median filtering for handling salt and pepper noise along with contrast enhancement for improving the overall

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

classification process. Furthermore, classification phase can be modified using ensemble-based approach for achieving performance gain.

IV. EMPIRICAL STUDY

The empirical study is based upon the result obtained corresponding to thyroid disease prediction using random forest, KNN and decision tree-based approaches. The result corresponding to these approaches is given within table 2

References	Approach	Dataset	Classification accuracy
(Huang et al, 2020) (Yadav & Pal, 2020b)	Diagnostic algorithm with decision tree	Thyroid dataset with csv format	95%
(Li et al., 2020)	Deep learning with random forest	PLCO	94%
(Dov et al., 2021)	Supervised learning based prediction model using KNN	ODDS	91%

Table 2: Results corresponding to KNN, Decision tree and random forest

The result corresponding to decision tree is better as indicated within table 2. The classification accuracy however depends greatly upon the feature selection which is random within random forest approach. This means random forest based approach can serve better in certain situation due to fluctuating nature of this approach. The plot corresponding to table 2 is given in figure 3

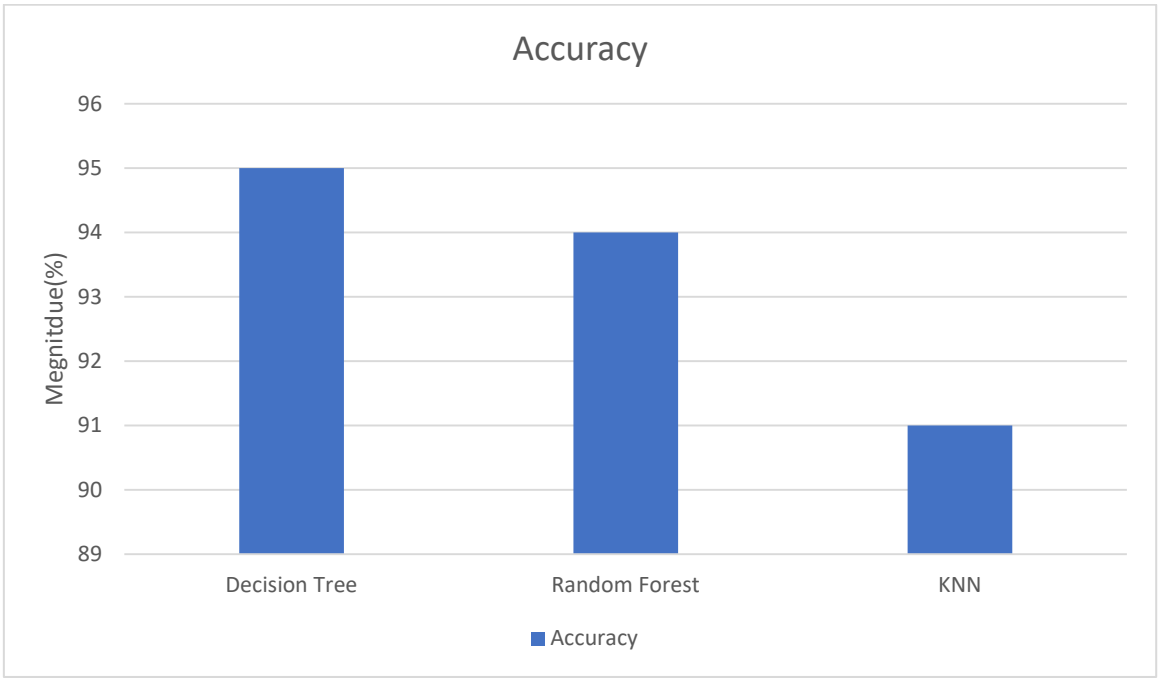

Figure 3: plots of classification accuracy

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

V. CONCLUSION

Thyroid can impact the immune system of the body adversely. Early detection and treatment is critical in this regard. Modern methodologies corresponding to the thyroid detection are effective but further improvement is desired. The Initial work highlights the certain aspects including pre-processing and classification that can be modified to achieve highest possible classification accuracy. The analyzed result corresponding to decision tree produced better classification accuracy in the range of 95% but with the complex dataset this might reduce. To increase the stability, convolution neural network-based mechanism having mode based pre-processing mechanism can reduce the degree of misclassification further. Furthermore, process of normalization can be introduced within the CNN to reduce complexity and improve convergence process.

REFERENCES

1. A Hood, C. K. (2000). Differential effects of microsomal enzyme inducers on in vitro thyroxine (T(4)) and triiodothyronine (T(3)) glucuronidation. *ToxicolSci*, *55*, 78–84.
2. AA Arriagada, E. A. M. O. (2015). Excess iodide induces an acute inhibition of the sodium/iodide symporter in thyroid male rat cells by increasing reactive oxygen species. *Endocrinol.*,*156*, 1540–1551.
3. Akbaş, A., Turhal, U., Babur, S., &Avci, C. (2013). Performance improvement with combining multiple approaches to diagnosis of thyroid cancer. *Engineering.*,*5*(10), 264–267. https://doi.org/10.4236/eng.2013.510b055
4. Azar, A. T., El-Said, S. A., &Hassanien, A. E. (2013). Fuzzy and hard clustering analysis for thyroid disease. *Comput Methods Prog Biomed*, *111*(1), 1–6. https://doi.org/10.1016/j.cmpb.2013.01.002
5. B Gereben, E. M. M. R. (2015). Scope and limitations of iodothyronine deiodinases in hypothyroidism. *Nat Rev Endocrinol*, *11*, 642–652.
6. Bai, Y., Kakudo, K., & Jung, C. K. (2020). Updates in the Pathologic Classification of Thyroid Neoplasms: A Review of the World Health Organization Classification. *Endocrinology and Metabolism*, *35*(4), 696. https://doi.org/10.3803/ENM.2020.807
7. Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2020a). Thyroid Disease Prediction Using Machine Learning Approaches. *National Academy Science Letters 2020 44:3*, *44*(3), 233–238. https://doi.org/10.1007/S40009-020-00979-Z
8. Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2020b). Thyroid Disease Prediction Using Machine Learning Approaches. *National Academy Science Letters 2020 44:3*, *44*(3), 233–238. https://doi.org/10.1007/S40009-020-00979-Z
9. Dov, D., Kovalsky, S. Z., Assaad, S., Cohen, J., Range, D. E., Pendse, A. A., Henao, R., & Carin, L. (2021). Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. *Medical Image Analysis*, *67*. https://doi.org/10.1016/J.MEDIA.2020.101814
10. Han, P., Guo, J., Lai, H., & Song, Q. (2022). Construction method of knowledge graph under machine learning. *International Journal of Grid and Utility Computing*, *13*(1), 11–20. https://doi.org/10.1504/IJGUC.2022.121423
11. Huang, B. L., Chabot, J. A., Lee, J. A., &Kuo, J. H. (2020). A stepwise analysis of the diagnostic algorithm for the prediction of malignancy in thyroid nodules. *Surgery (United States)*, *167*(1), 28–33. https://doi.org/10.1016/J.SURG.2019.05.079
12. I Ionita, L. I. (2016). Prediction of thyroid disease using data mining techniques. *Broad Res ArtifIntellNeurosci*, *7*(3), 115–124.
13. I Ioniţă, L. I. (2016). Prediction of thyroid disease using data mining techniques. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *7*(3), 115–124.
14. Jha, R., Bhattacharjee, V., &Mustafi, A. (2022). Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society. *Wireless Personal Communications*, *122*(2), 1921–1938. https://doi.org/10.1007/S11277-021-08974-3/FIGURES/7
15. Jia, M., Li, Z., Pan, M., Tao, M., Lu, X., & Liu, Y. (2020). Evaluation of immune infiltrating of thyroid cancer based on the intrinsic correlation between pair-wise immune genes. *Life Sciences*, *259*. https://doi.org/10.1016/J.LFS.2020.118248

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

16. Kumar, H. H. S. (2020). A novel approach of SVM based classification on thyroid disease stage detection. *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, 836–841. https://doi.org/10.1109/ICSSIT48917.2020.9214180
17. Li, Y., Chen, P., Li, Z., Su, H., Yang, L., &Zhong, D. (2020). Rule-based automatic diagnosis of thyroid nodules from intraoperative frozen sections using deep learning. *Artificial Intelligence in Medicine*, *108*. https://doi.org/10.1016/J.ARTMED.2020.101918
18. Liu, T., Guo, Q., Lian, C., Ren, X., Liang, S., Yu, J., Niu, L., Sun, W., & Shen, D. (2019). Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. *Medical Image Analysis*, *58*. https://doi.org/10.1016/J.MEDIA.2019.101555
19. Ma, J., Wu, F., Jiang, T., Zhao, Q., & Kong, D. (2017). Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery 2017 12:11*, *12*(11), 1895–1910. https://doi.org/10.1007/S11548-017-1649-7
20. Mugasa, H., Dua, S., Koh, J. E. W., Hagiwara, Y., Lih, O. S., Madla, C., Kongmebhol, P., Ng, K. H., & Acharya, U. R. (2020). An adaptive feature extraction model for classification of thyroid lesions in ultrasound images. *Pattern Recognition Letters*, *131*, 463–473. https://doi.org/10.1016/J.PATREC.2020.02.009
21. Poudel, P., Illanes, A., Sheet, D., &Friebe, M. (2018). Evaluation of Commonly Used Algorithms for Thyroid Ultrasound Images Segmentation and Improvement Using Machine Learning Approaches. *Journal of Healthcare Engineering*, *2018*. https://doi.org/10.1155/2018/8087624
22. Prasad, V., Rao, T. S., &Babu, M. S. P. (2016). Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms. *Soft Comput*, *20*(3), 1179–1189. https://doi.org/10.1007/s00500-014-1581-5
23. Priyadharsini, D., &Sasikala, S. (2022). Efficient Thyroid Disease Prediction using Features Selection and Meta-Classifiers. *Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022*, 1236–1243. https://doi.org/10.1109/ICCMC53470.2022.9753986
24. Prochazka, A., Gulati, S., Holinka, S., &Smutek, D. (2019). Patch-based classification of thyroid nodules in ultrasound images using direction independent features extracted by two-threshold binary decomposition. *Computerized Medical Imaging and Graphics*, *71*, 9–18. https://doi.org/10.1016/J.COMPMEDIMAG.2018.10.001
25. SM Gorade, A. D. P. P. (2017). A study of some data mining classification technique. *Int Res J EngTechnol*, *4*(4), 3112–3115.
26. Snekhalatha, U., &Gomathy, V. (2018a). Ultrasound Thyroid Image Segmentation, Feature Extraction, and Classification of Disease Using Feed Forward Back Propagation Network. *Advances in Intelligent Systems and Computing*, *563*, 89–98. https://doi.org/10.1007/978-981-10-6872-0_9/COVER
27. Snekhalatha, U., &Gomathy, V. (2018b). Ultrasound Thyroid Image Segmentation, Feature Extraction, and Classification of Disease Using Feed Forward Back Propagation Network. *Advances in Intelligent Systems and Computing*, *563*, 89–98. https://doi.org/10.1007/978-981-10-6872-0_9
28. TL Fonseca, M. C.-M. M. C. (2013). Coordination of hypothalamic and pituitary T3 production regulates TSH expression. *J ClinInvestig*, *123*, 1492–1500.
29. U Sidiq, S. A. R. K. (2019). Diagnosis of various thyroid ailments using data mining classification techniques. *Int J Sci Res ComputSciEngInfTechnol*, *5*(1), 2456–3307.
30. Vanderpump, M. P. J. (2011). The epidemiology of thyroid disease. *British Medical Bulletin*, *99*(1), 39–51. https://doi.org/10.1093/BMB/LDR030
31. Yadav, D. C., & Pal, S. (2019). To generate an ensemble model for women thyroid prediction using data mining techniques. *Asian Pac J Cancer Prev*, *20*(4), 1275–1281. https://doi.org/10.31557/apjcp.2019.20.4.1275
32. Yadav, D. C., & Pal, S. (2020a). Prediction of thyroid disease using decision tree ensemble method. *Human-Intelligent Systems Integration 2020 2:1*, *2*(1), 89–95. https://doi.org/10.1007/S42454-020-00006-Y
33. Yadav, D. C., & Pal, S. (2020b). Prediction of thyroid disease using decision tree ensemble method. *Human-Intelligent Systems Integration*, *2*(1–4), 89–95. https://doi.org/10.1007/S42454-020-00006-Y
34. YASIR HUSSEIN SHAKIR. (2020). *Thyroid Disease Data Set | Kaggle*. Kaggle. https://www.kaggle.com/datasets/yasserhessein/thyroid-disease-data-set

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 11, November 2022||

| DOI:10.15680/IJIRSET.2022.1111014|

35. Yu, R., Tian, Y., Gao, J., Liu, Z., Wei, X., Jiang, H., Huang, Y., & Li, X. (2022). Feature discretization-based deep clustering for thyroid ultrasound image feature extraction. *Computers in Biology and Medicine*, *146*. https://doi.org/10.1016/J.COMPBIOMED.2022.105600
36. Zhu, Y. C., AlZoubi, A., Jassim, S., Jiang, Q., Zhang, Y., Wang, Y. B., Ye, X. de, & DU, H. (2021). A generic deep learning framework to classify thyroid and breast lesions in ultrasound images. *Ultrasonics*, *110*. https://doi.org/10.1016/J.ULTRAS.2020.106300

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details