International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Decision Tree Ensemble Techniques To Predict Thyroid Disease

Dhyan Chandra Yadav , Saurabh Pal

Abstract*: Decision tree provides help in making decision for very complex and large dataset. Decision tree techniques are used for gathering knowledge. Classification tree algorithms predict the experimental values of women thyroid dataset. The objective of this research paper observation is to determine hyperthyroidism, hypothyroidism and euthyroidism participation in hormones can be good predictor of the final result of laboratories and to examination whether the propose ensemble approach can be similar accuracy to other single classification algorithm. In the proposed experiment real data from 499 thyroid patients were used classifications algorithms in predicting whether thyroid detected or not detected on the basis of T3, T4 and TSH experimental values. The results show that the expectation of maximization classification tree algorithms in those of the best classification algorithm especially when using only a group of selected attributes. Finally we predict batch size, tree confidential factor, min number of observation, num folds, seed, accuracy and time build model with different classes of thyroid sickness. Different classification algorithms are analyzed using thyroid dataset. The results obtained by individual classification algorithms like J48, Random Tree and Hoeffding gives accuracy 99.12%, 97.59% and 92.37 respectively. Then we developed a new ensemble method and apply again on the same dataset, which gives a better accuracy of 99.2% and sensitivity of 99.36%. This new proposed ensemble method can be used for better classification of thyroid patients.

***Keywords* : J48, Random Tree, Hoeffding, Prediction, T3, T4, TSH, hypothyroidism, hyperthyroidism, euthyroidism and ensemble model**

I. INTRODUCTION

Hormones play major role in blood stream to maintain metabolism in human. The production of high hormones and low hormones both are dangerous. The general objective of thyroid gland is to produce thyroid hormones. The main objective of thyroid gland is to maintain bloodstream through the regulation of metabolism. If thyroid gland produces more hormones then it will be hyperthyroidism and if thyroid gland produces less hormones then it will be hypothyroidism [1].

The three hormones tri-iodothyronine (T3), L-thyroxin and TSH regulate the metabolic functions of human body. These hormones utilize proteins and manage fats in human body. In pituitary gland thyrotrophic stimulating is released if require more hormones. The pituitary gland control and manage production of hormones in the blood stream [2].

Thyroid disease is a different thing about all other diagnosis system. It's visibility and treatments are different. Thyroid hormone has many symptoms in initial to final stage. It is generally arises from disorder life style and foods after it increasing and decreasing hormones production finally make health system [3].

The paper analysis is organized on batch size, confidential factor, num decimal places, num folds, seed and accuracy of a model in decision making through J48, Hoeffding and Random Tree. The discuss about all dataset of thyroid in multiple way of classification tree algorithm and finally measure the evaluation accuracy increases with time built model training set. Classification decision tree provide many types help in identification of thyroid disease. It provides better help in dataset classification as a tree model in which attributes represents root ,nodes and leaf nodes. Analyst easily analysis all the functions of related dataset [4].

By the help of proposed three algorithms easily classify hyperthyroidism, hypothyroidism and euthyroidism. We easily indentify as a tree path and finally reach on decision node to leaf nodes. Present paper discuss to all the symptoms of thyroid patients and declare types of problem. It is very difficult to identify hyperthyroidism and hypothyroidism problems in another way [5].

II. RELATED WORK

Ahmad et.al discussed about thyroid endocrine gland in blood issue and function of the body. They discussed by feature selection, Fuzzy rule, maximal absolute difference Linguistic Hedge and total serum thyroxin. They provided classification accuracy 98.604% and achieved different testing phase of clustering one, two, three and four clusters for each class and 12 fuzzy rules. The generated 88.372%, 90.6977%, 91.6744%, and 97.6744% cluster size during training phase [6].

Tahani et.al discussed about clustering ensemble model and how combines multiple clustering models. They analyzed adaptive clustering ensemble model. Adaptive algorithm measured and transformed initial clusters into binary representation aggregation to produce final clusters. They used co- association, k-means, similarity measurement, machine learning and data mining [7].

Xiyu et.al discussed about new class of tissue system .They analyzed traditional tissue P systems to new class of tissue system.

Manuscript published on 30 September 2019
* Correspondence Author
Dhyan Chandra Yadav*, Department of Computer Applications, Veer Bahadur Singh Purvanchal University, Jaunpur, India. Email: dc9532105114@gmail.com
Saurabh Pal, Department of Computer Applications, Veer Bahadur Singh Purvanchal University, Jaunpur, India. Email: drsaurabhpal@yahoo.co.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

They used thyroid disease analysis, tissue P system, membranes structure and clustering algorithm. They analyzed thyroid disease for classification [8].

Amrollahi et.al discussed about effect on thyroid gland of human bodies. They analyzed how thyroid function managed and balanced the metabolism. They used expert systems Bio-chemistry. They used fuzzy rules to system provided help in non expert who are suspicions of their thyroid function and provided help in expert for their diagnosis [9].

Ahmad et.al discussed about thyroid hormones production from thyroid gland. They analyzed compression hard and fuzzy clustering for thyroid disease and find optimal number of clusters. They used thyroid disease K-means, K-model clustering fuzzy C-means. They improved actual number of clusters present in thyroid data set and find clustering performance is much better to compare to other [10].

Saiti et.al discussed about thyroid cancer by different classifier algorithms. They increase accuracy of thyroid cancer dataset. They generated an ensemble model to predict thyroid cancer and provided much accuracy compare to other previous prediction [11].

Vikas et.al discussed about medical dataset by different machine learning algorithms. They used data mining different algorithms for taking decision as a decision tree and regression tree. After all the prediction find 93% classification accuracy. They suggested to boost algorithm for prediction [12].

Vikas et.al discussed about breast cancer in women by some different machine learning algorithms. They used supporting key as like: Breast, Data Mining, Naïve Bayes and RBF Network. After all the prediction they find Naïve Bayes give the highest accuracy 97.36% [13].

Bridget et al. discussed about circulation serum FT4 level in pregnant women. In circulation FT4 perchlorate exposure is negatively associated with circulating levels in third trimester pregnant women. They used Perchlorate , Iodine, Pregnancy, birth and Weight[14].

Awasthi and Anil Antony discussed about classification and diagnosis of thyroid disease . they used KNN, Support Vector Machine, T3, T4 and TSH for diagnosis. They find some values missing while the user entering the values. They used K-Nearest Neighbor algorithm in thyroid diagnosis for approximating the missing values in the user input [15].

This analysis paper analyze classification tree algorithms: J48, Random tree and Hoeffding for predicting where thyroid hormones as like T3, T4 and TSH experimental values. The results show that the expectation maximization classification tree algorithms in those of the best classification algorithm especially when using only a group of selected attributes. Summaries all the analysis paper predict batch size, tree confidential factor, min number of observation, num folds, seed, accuracy and time build model with different classes of thyroid sickness.

III. METHODOLOGY

Collected thyroid dataset is from github uci and pathology. The use of these data sets for only experimental purposes. We have used all methodology are used in four stages:

A- Data Description
B- Algorithm description
C- Proposed Method

A. Data Description

In this experiment we select data from Rahul thyroid diagnosis center and github uci. All the dependable variables have definition with his explanation as like: Hyperthyroidism: Too much hormone production, Hypothyroidism: To little hormone production, Euthyroidism: The state of normal thyroid function. All the hormones (T3, T4 and TSH) define in with his evolution ranges in ng/dl, µg/dl and µl/ml. The values of T3, T4 and TSH mentioned in above table.1.

Table.1 Thyroid Dataset variables representation

Source	Rahul thyroid diagnosis center, https://github.com/mikeizbicki/datasets/blob/master/csv/uci/new-thyroid.names		References
Sample Size	499= Total: 228 =Hyperthyroidism, 237= Euthyroid State and 34= Hypothyroidism		
Dependent Variables			
Hyperthyroidism	Too much hormone production		
Hypothyroidism	To little hormone production		[3],[16]
Euthyroidism	The state of normal thyroid function		
Independent Variables			
T3	(60-200)ng/dl	Triiodothyronine Stimulates the metabolism	
T4	(4.5-12.0)µg/dl	Thyroxin produced by thyroid gland	[3],[16],[17]
TSH	(6.3-5.5)µl/ml	Thyroid Stimulating Hormone pituitary hormone	

B. Algorithms Description

In this analysis developed model provides support to doctor in treatment. Proposed model is for consulting only doctors but final decision follow by doctors in treatment. Three classification algorithms J48, Random Tree and Hoeffding. Generate and combine model with carrying the majority by voting algorithms. Different varieties of seeds and portioned into different classes which is based on many features**.**

RANDOM TREE: Random tree provide a platform for merging the individual learners. It constructs a random field of data for constructing decision tree. Every nodes of generated tree behaves as like best split for all variables and randomly choose best node. If we use random tree as a group of tree then it will be tree predictors or forest but all the mechanism follows the random trees and the outputs the class level has received the majority of votes. Random tree improve the performance of single decision tree and conscience more way of randomization [17], [18],[19],[20].

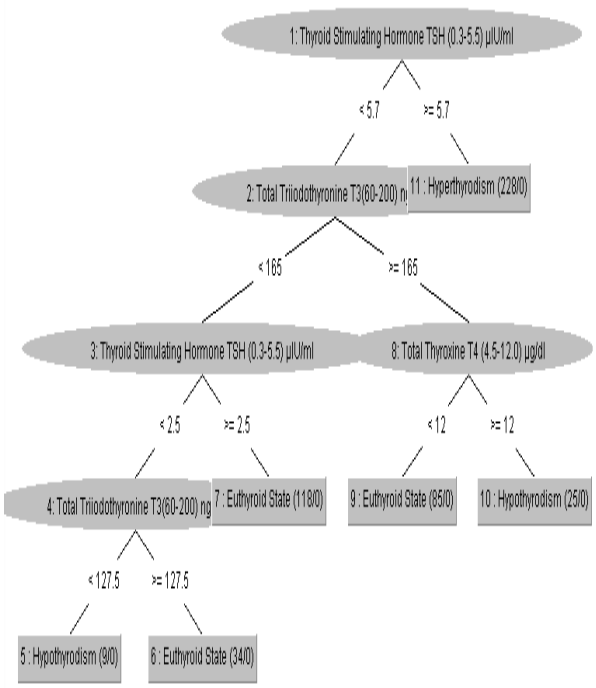


Fig.1. Random tree representation for thyroid dataset

J48: The generate the building model of thyroid data set classes by J48 algorithm from a set of records that contain class level. The decision tree find out the way of attribute direction behaves for all thyroid instances. By the help of this algorithm we will generate the rules for prediction for all target variables. The main objective of this algorithm in decision generation more progressive, decision tree and gain more accurate result by decision tree [21].

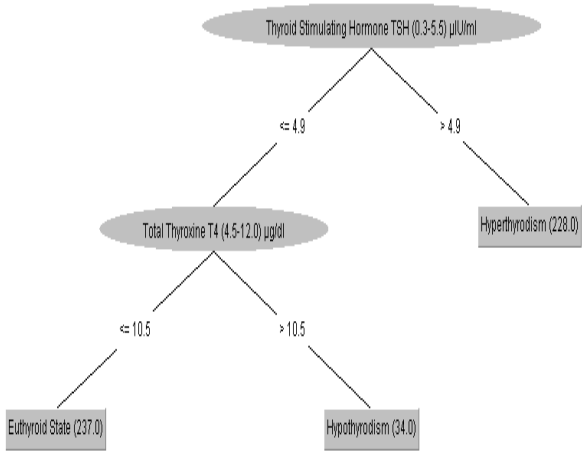


Fig.2. J48 representation for thyroid dataset

***HOEFFDING*:** Hoeffding decision tree algorithm support in generating stream. Now generate an incremental generating stream that will not be change over time [22].

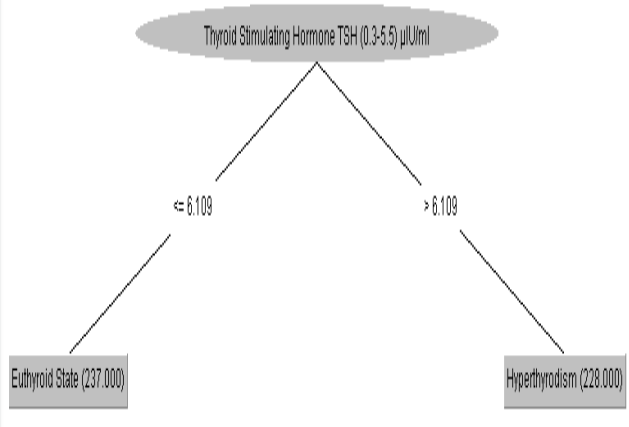


Fig.3. Hoeffding representation for thyroid dataset

C. Proposed Method

This analysis uses algorithm for classification and prediction, by the help of these decision trees easily organize form of tree structure of thyroid dataset. Nodes of the tree shows the attributes of the dataset and edges will be used for represent the value of these attributes and finally find the leaf nodes as decision nodes. In propose model select women thyroid dataset from laboratories and after the preprocessing of missing values evaluate the correlation of attribute and take the values of T3, T4 and TSH then apply the algorithm in many faces of tree. Classify all thyroid dataset with different iteration of attributes and compare with tree ensemble model and finally evaluate the majority of voting for different classes of thyroid as like hyperthyroidism, hypothyroidism and euthyroidism.

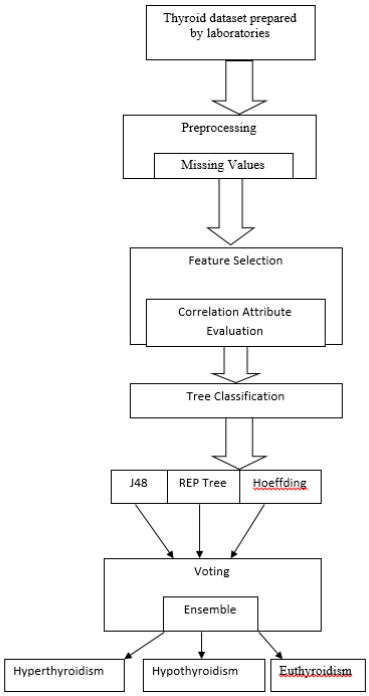


Fig.4. Proposed ensemble model for thyroid dataset

IV. RESULTS AND DISCUSSION

Thyroid dataset in which 499 cases included as a record in csv file. The classification model has four attributes T3, T4, TSH and the target variables as a class level observation have three type of classes that are: hypothyroidism, hyperthyroidism and euthyroidism.

Some issues overcome in random split thyroid dataset in training and testing set. The random splits find contradiction between getting results and realistic results. In this paper, discuss and compare three tree algorithms with an ensemble model:

Experiment-I:

The role of J48 tree algorithm in thyroid dataset for different number of samples as like 100, 200 and 300, by the help of these batch prediction easily perform instances process. In this analysis observe confidential factor (0.25, 0.50 and 0.75) pruning and find minimum number of branches of instances in thyroid dataset experiment are 2, 3 and 4,. Analyze and discuss about controlling the sequences by number and reduce error pruning for randomize dataset.

Table.2.Computational table for thyroid dataset using J48

Iterations	Batch Size	Confidential Factor	Min NumObj	Num Folds	Seed	Accuracy	Time (in Second)
1	100	0.25	2	3	1	97.32	0.04
2	200	0.50	3	5	2	97.34	0.04
3	300	0.75	4	10	3	99.12	0.03

By the experiment it is clear that the default value of seed is 1, but select different sequence of random attributes by changing the seed 2 and 3. In this experiment find, if increase batch size (300), MinNumObj (4), Number of fold (10) and seed (3) then find highest accuracy (99.12%) with less time build model 0.03 seconds.

Experiment-II:

The role of random tree algorithm in thyroid dataset for different number of samples. In this analysis observe confidential factor, minimum number of branches, Num Folds and seed with different increasing accuracy.

Table.3. Computational table for thyroid dataset using Random tree

Iterations	Batch Size	Confidential Factor	Min NumObj	Num Folds	Seed	Accuracy	Time (in Second)
1	100	0.25	2	3	1	93.67	0.02
2	200	0.50	3	5	2	93.67	0.02
3	300	0.75	4	10	3	97.59	0.02

In this experiment find, if increase batch size (300), MinNumObj (4), Number of fold (10) and seed (3) then find highest accuracy (97.59%) with less time build model 0.02 seconds.

Experiment-III:

The role of hoeffding tree algorithm in thyroid dataset for different number of samples. In this analysis observe confidential factor, minimum number of branches, Num Folds and seed with different increasing accuracy.

Table.4 Computational table for thyroid dataset using Hoeffding

Iterations	Batch Size	Confidential Factor	Min NumObj	Num Folds	Seed	Accuracy	Time (in Second)
1	100	0.25	2	3	1	89.31	0.07
2	200	0.50	3	5	2	89.22	0.07
3	300	0.75	4	10	3	92.37	0.05

In this experiment find, if increase batch size (300), MinNumObj (4), Number of fold (10) and seed (3) then find highest accuracy (92.37%) hoeffding tree with less time build model 0.05 seconds.

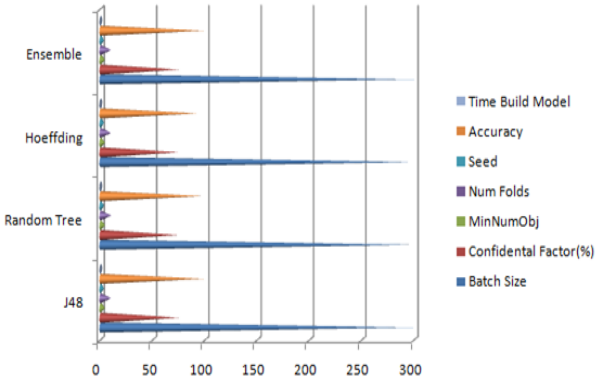
Experiment-IV:

The role of ensemble model in thyroid dataset for different number of samples. In this analysis observe confidential factor, minimum number of branches, Num Folds and seed with different increasing accuracy.

Table.5. Computational for thyroid dataset using Ensemble Model

Iterations	Batch Size	Confidential Factor	Min NumObj	Num Folds	Seed	Accuracy	Time (in Second)
1	100	0.25	2	3	1	98.43	0.05
2	200	0.50	3	5	2	98.43	0.05
3	300	0.75	4	10	3	99.20	0.05

In this experiment find, if increase batch size (300), MinNumObj (4), Number of fold (10) and seed (3) then find highest accuracy (99.20%) with less time build model 0.05 seconds. In the above all analysis of experiment I,II,III and IV, find the highest batch size(300), confidential factor(75%), minimum number of objects(4), number of folds (10) and seed(4) claculated heighest classification accuracy (99.2%).



	Ensemble	J48	Random Tree	Hoeffding
Time Build Model	0.05	0.03	0.02	0.05
Accuracy	99.2	99.12	97.59	92.37
Seed	4	4	4	4
Num Folds	10	10	10	10
MinNumObj	4	4	4	4
Confidental Factor(%)	75	75	75	75
Batch Size	300	300	300	300

Fig.5. Computational figure with table of thyroid dataset for comparing

It is generated by Ensemble model of all given tree algorithms so majority of voting find that ensemble model of these tree algorithm is best. There is not major difference between generated time build models.

V. CONCLUSION

Hormones disorders in thyroid are a major problem in human. Various researchers work continues on this thyroid field and they used classification based data mining techniques. In this analysis used J48, Hoeffding and Random tree on thyroid dataset and identify more accurately model of decision tree on all possible experiments. In experimental study collect data from the values of T3, T4, TSH, Thyroid gland, Hyperthyroid, Hypothyroid and Euthyroidism at various levels and find (99.2%) classification accuracy in thyroid dataset. It is more accurate result of ensemble model compares the all other used tree algorithms with (0.05 seconds) time built model. Summaries all the experimental results and implement decision tree using more helpful data mining technique. In this analysis the ensemble classification technique improved evaluates accuracy and test thyroid dataset. In future work observe the identification of different affected factors of thyroid dataset and test more different and large dataset for diabetes, heart disease etc.

ACKNOWLEDGEMENTS

The author is grateful to Veer Bahadur Singh Purvanchal University Jaunpur, Uttar Pradesh, for providing financial support to work as Post Doctoral Research Fellowship.

REFERENCES

1. Ozyilmaz, Lale, and Tulay Yildirim. "Diagnosis of thyroid disease using artificial neural network methods" Neural Information Processing,. Proceedings of the 9th International Conference on. Vol. 4. IEEE, 2002.
2. http://www.emedicinehealth.com/thyroid_faqs/ar ticle_em.htm.
3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3 169866/ (accessed dec 2015).
4. J.W.F.Elte, J.K.Bussemaker, K.A.Haa, “The natural history of euthyroid multinodul argoiter” BMJ, Postgrad Med J, 66 (1990)186-190.
5. G.Zhang,L.V.Berardi, “An investigation of neural networks in thyroid function diagnosis”, HealthCare Management Science, (1998) 29-37.
6. Ahmad Taher Azar Ella Hassanien and Tai ham kom “Expert system based on Neural Fuzzy Rules for Thyroid Disenes Diagnosis", International confrence on Bio- Science and Bio-Technology. BSBT 2012, mulgrab 2012, IUrc 2012: computer application for Bio-Technology, multimedia and Ubiquitous city PP94-105.
7. Tahani Alqurashi and Wenjia Wang "Clustering ensemble method", International Journal of machine learning and cybernetics PP-1-20. 16 January 2018.
8. Xiyu Liu and Alice Xue "The thyroid analysis by a class of tissue P system", Information Tecnology in Medicine and Education (IT ME), 2012, International Symposium on Voiume 2.
9. S. Amrollahi Biyniki, I.B Turksen and M.H. Fozel Zarandi "Fuzzy rule based on expert system for diagnosis of thyroid disense", 2015 IEEE confrence on computational Intelligence in Bio information and computational Biology (CIBCB) 12--15 Aug 2015.
10. Ahmad Taher Azar, Shaima Ahmad El-Said and Abonl Ella hassanien "Fuzzy and hard clustering Analysis for thyroid disense", Computer method and programe in Biomedicine III(2013) I-I6 www.intl.elsevierhealth.com/journals/cmpp.
11. F. Saiti, A. A. Naini, M. A. Shoorehdeli and M. Teshneh-lab, “Thyroid Disease Diagnosis Based on Genetic Algo-rithms Using PNN and SV M”, The International Bioin-formatics and Biomedical Engineering (ICBBE), Beijing, 11-13 June 2009, pp. 1-4.
12. Vikas Chauraisa, Saurabh pal and Tiwari “Chronic Kidney Disease: A Predictive model using Decision tree”, International Journal of engineering Research and technology, Vol.11, Issue 11. November 2018, ISSN -0974-3154.
13. Vikas Chauraisa, Saurabh pal and Tiwari “Prediction of begin and benign and malignant breast cancer using data mining techniques” Journal of algorithms and Computational Technology, Vol. 12(2), 2018.
14. Bridget A. Knight, Beverley M. Shields, Xuemei He, Elizabeth N. Pearce, Lewis E. Braverman, Rachel Sturley and Bijay Vaidya “Effect of perchlorate and thiocyanate exposure on thyroid function of pregnant women from South-West England: a cohort study” Thyroid Research (2018) 11: 9, Springer Nature Switzerland AG 2018.
15. A K Awasthi and Anil Antony “An Intelligent System for Thyroid Disease Classification and Diagnosis” 2018 Second international conference on Invetive Communication and Computational Technologies, IEEE Xplore: 27 September 2018
16. http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease 2013.
17. http://www.irdindia.in/journal_ijaece/pdf/vol3_iss4/2.pdf
18. N. Landwehr, M. Hall, and E. Frank, “Logistic model trees”, Mach. Learn., vol. 59, no. 12, pp.161-205, 2005.
19. Breiman Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32.
20. Liaw, Andy (16 October 2012). “Documentation for R package random forest”. Retrieved 15 March 2013.
21. Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE.
22. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/Hoeffding Tree.html.
23. Verma AK, Pal S, Kumar S. Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study. Applied biochemistry and biotechnology. 2019 Jul 27:1-9.
24. Yadav DC, Pal S. To Generate an Ensemble Model for Women Thyroid Prediction Using Data Mining Techniques. Asian Pacific Journal of Cancer Prevention. 2019 Apr 1;20(4):1275-81.
25. Verma AK, Pal S, Kumar S. Classification of Skin Disease using Ensemble Data Mining Techniques. Asian Pacific Journal of Cancer Prevention. 1887;20(6).

AUTHORS PROFILE

Dhyan Chandra Yadav received his MCA (Computer Application) from VBSPU Jaunpur, U.P., India (2008) and obtained his Ph.D degree from the SVU, Gajraulla, Amroha, U.P. (2016). He then joined the VBSPU Jaunpur as Post Doctoral Fellow (2018). His research interests include Data Mining, Compiler and Knowledge Discovery.

Saurabh Pal received his M.Sc. (Computer Science) from Allahabad University, UP, India (1996) and obtained his Ph.D. degree from the Dr. R. M. L. Awadh University, Faizabad (2002). He then joined the Dept. of Computer Applications, VBS Purvanchal University, Jaunpur as Lecturer. At present, he is working as Head and Associate Professor at Department of Computer Applications. Saurabh Pal has authored a commendable number of research papers in international/national Conference/journals and also guides research scholars in Computer Science/Applications. He is an active member of IACSIT, CSI, Society of Statistics and Computer Applications and working as Reviewer/Editorial Board Member for more than 50 international journals. His research interests include Image Processing, Data Mining, Grid Computing and Artificial Intelligence