International Journal of Business Analytics
Volume 9 • Issue 3

Developing an Explainable Machine Learning-Based Thyroid Disease Prediction Model

Siddhartha Kumar Arjaria, Rajkiya Engineering College, Banda, India*
Abhishek Singh Rathore, Shri Vaishnav Vidyapeeth Vishwavidyalaya, India
https://orcid.org/0000-0002-5513-2639
Gyanendra Chaubey, Rajkiya Engineering College, Banda, India

ABSTRACT

Healthcare and medicine are key areas where machine learning algorithms are widely used. The medical decision support systems thus created are accurate enough; however, they suffer from the lack of transparency in decision making and shows a black box behavior. However, transparency and trust are significant in the field of health and medicine, and hence, a black box system is sub optimal in terms of widespread applicability and reach. Hence, the explainablility of the research makes the system reliable and understandable, thereby enhancing its social acceptability. The presented work explores a thyroid disease diagnosis system. SHAP, a popular method based on coalition game theory, is used for interpretability of results. The work explains the system behavior both locally and globally and shows how machine leaning can be used to ascertain the causality of the disease and support doctors to suggest the most effective treatment of the disease. The work not only demonstrates the results of machine learning algorithms but also explains related feature importance and model insights.

KEYWORDS

Explainable AI, Features, Healthcare, Interpretability, Logistic Regression, Machine Learning, SHAP, Thyroid Disease

INTRODUCTION

Accurate decision-making for a given situation serves as a benchmark for human intelligence and combined with critical reasoning catalyzes social change. In the current era of human-machine interaction, machine learning algorithms are being used for decision-making by computational machines (Piano, 2020). These decision support systems effectively make accurate choices in many domains. Although these machine algorithm-based decision support systems are very accurate, they suffer from a lack of interpretability as important structure and relationship-related information of data and models are hidden in these systems, effectively making these systems into black boxes. These black-box systems fail to answer important questions like the effect of each feature in the final decision and model behavior issue. Since the process of result generation is shrouded in mystery, it

DOI: 10.4018/IJBAN.292058 *Corresponding Author

remains unclear if the optimization was done locally or globally. Thus, the outcomes generated by these systems may be stand-alone in terms of results and systems are unable to give the correct basis behind this outcome.

To deal with these issues, it is necessary to make the systems explainable. These explainable systems are white boxes because they transparently explore various aspects of decision making like the importance of features, model, and the predictions of decision support systems (Felzmann, Villaronga, Lutz, & Larrieux, 2020). This is especially useful in the field of medicine, where machine learning algorithms are being used for the effective diagnosis of disease. Artificial Intelligence plays a supportive role in this scenario to find out the signed quantitative impact of each feature on the result (Rong, Mendez, Assi, Zhao, & Sawan, 2020). Transparency in the system can provide additional information that may assist doctors in making the best decision for the treatment of disease. This would enhance the understanding and acceptability of AI by making it explainable as is the case in the scenario where AI is used to must make practical decisions (autonomous car driving) (Samek & Muller, 2019). It covers medical, bioinformatics, banking, insurance, cognitive, space, education, psychology, chemical, and many more domains.

In machine learning, linear systems are simple enough to be transparent. The linear relationship between the cause (Features) and effect(output) can be understood easily. But the scenario becomes complicated when the complex and highly nonlinear environment persists. The complexity of the problems increases with an increase in the number of features and parameters taken into consideration (Schmidt, Marques, Botti, & Marques, 2019).

Thyroid disease is one of the most prominent diseases in the world. More than 12 percent of the U.S. population is affected by thyroid problems throughout their life (World Thyroid Day is Heralded by International Thyroid Societies, 2015). The main cause of the disease is the lack of iodine. Thyroid disproportionately affects women between the age of 17-54. In extreme cases, complications associated with thyroid also cause cardiovascular problems, blood pressure, high cholesterol levels, depression, anxiety, and decreased fertility.

The thyroid gland produces two active hormones: Total Serum Thyroxin (T4) and Total Serum triiodothyronine (T3). These are important to maintain thyroid metabolism in the body and any misbalance in these hormones causes thyroid disease. The thyroid gland is responsible for controlling the metabolism of the body (Teixeira, Santos, & Pazos-Moura, 2020). Thyroid activity in our body is divided into three different sects as euthyroidism, hyperthyroidism, and hypothyroidism. Euthyroidism represents the normal case while hyper and hypo represent the abnormal hormonal situation. Hyperthyroidism is the situation when body cells produce an extra amount of thyroid than needed while the hypothyroid situation is created in case of deficiency. The results produced by machine learning algorithms in differentiating between these cases are very good, but present algorithms are not able to explain these findings (Temurtas, 2009).

The presented work first uses a logistic regression model to diagnose thyroid disease. The work uses the benchmark data set of thyroid disease available at UCI KDD. To explain the Decision support system SHAP (SHapley Additive explanations) method is used. SHAP is the model agnostic way to explain the system globally (Messalas, Kanellopoulos, & Makris, 2019). The portability of the system is very high. The method can be used for explaining almost all the machine learning algorithm performance. The SHAP method is based on the coalition game theory. Each feature of the data set act as a player and the result act the outcome of the game. The SHAP values of each feature indicate the signed impact of each feature on overall system performance (Chakraborty, Awolusi, & Gutierrez, 2021). This paper finds out the importance of each feature in deciding thyroid disease. In addition to this, the conditional dependence of the features is studied instance-wise, and this will help the medical practitioners to understand the hidden facts of the disease. The local methods like decision plots, force plots, and waterfall models demonstrate the role of each feature in disease prediction. The global methods like the feature summary plot demonstrate the overall features in terms of importance globally.

The rest of the paper is divided into five different parts. Section two presents the related work that has been done for explainable AI in the healthcare domain. Section three presents the methodology proposed in the paper. Section four explores the results in detail using different graphs and tables. Section five presents the conclusion based on the findings and results of section four along with the future aspects.

LITERATURE REVIEW

It was a challenge for every researcher to understand the black box behind the prediction of machine learning algorithms. To solve this problem, explainable AI came into existence. This novel field has yet to be used extensively used in explaining the models used in Thyroid Disease prediction. We'll begin our discussion by reviewing the seminal works that have been undertaken to understand the black box behind the prediction of Thyroid Disease and healthcare problems.

Islam et al. (Islam, Barua, Begum, & Ahmed, 2019) discussed case-based reasoning and domain-specific ontology of explaining AI-based models for the diagnosis of hypothyroid. Using this case-based reasoning they were able to attain an accuracy of 95% in the diagnosis of hypothyroid. Chaubey et al. (Choubey, Bisen, Arjaria, & Yadav, 2021) presented a study to predict thyroid-prone patients using three prominent machine learning techniques like Logistic Regression, Decision Tree, and k-Nearest Neighbor (KNN) These three techniques have been compared in terms of their accuracy. Amongst the studied techniques, KNN shows the highest accuracy of 96.987% using T3 and T4 as prominent features. In 2020, Abadi and Berrada (Adadi & M, 2020) explained and interpreted the black box behind artificial intelligence (AI) and discuss its importance in healthcare. They have also discussed the methods, characteristics, and future of explainable AI in healthcare.

Pawar et al. (Pawar, O'Shea, Rea, & O'Reily, Explainable AI in Healthcare, 2020) used XAI in 2020 as a technique to model accountability, transparency in feature selection, result monitoring, and model improvement of AI techniques in the domain of healthcare. Pawar et al. (Pawar, O'Shea, Rea, & O'Reilly, Incorporating Explainable Artificial Intelligence (XAI) to aid Understanding of Machine Learning in the Healthcare Domain, 2020) used explainable AI in 2021 to provide an aid for understanding Machine Learning models in the healthcare domain. In this work, a powerful solution for increasing the understandability of medical practitioners towards AI-based systems is given. This work mainly focuses on the explainability of feature selection in ML models. Yang et al. (Yang, Ye, & Xia, 2021) presented a mini-review to unbox the black box of medical explainable AI via multimodal and multi-data center fusion. They have used quantitative and qualitative analysis to prove the efficacy of AI-based systems from which a broader range of clinical questions can be envisaged.

Moradi and Samwald (Moradi & Samwald, 2021) presented a study on post hoc explanation of black box behind the machine learning classifiers using confident itemset. They introduced the confident itemset explanation (CIE) method to explain the black box and achieved a higher accuracy of 9.3% and higher interpretability of 8.8%. Kenny and Keane (Kenny & Keane, 2019) compared results of black box Artificial Neural Networks and an explainable AI method case-based reasoning to find the feature weighting. They introduced a new method of evaluation called as contribution-based method (COLE) and found it to perform best. Thomas (Thomas & Haertling, 2020) has presented a similarity-based model to detect Thyroid nodules in patients. This work has been conducted using ultrasound images of Thyroid nodules and attained an accuracy of 81.5%.

Hacker et al. (Hacker, Krestel, Grundmann, & Naumann, 2020) talked about the ethical behavior of explainable AI and proposed a tradeoff between the explainability and accuracy of the models. They used two types of case studies as medical and corporate mergers and demonstrated the tradeoff and its effect in the technical case on the spam classification model. Fu et al. (Fu, Chang, Liu, & Yang, 2019) proposed a data-driven group decision method to diagnose thyroid nodules. In the group decision method, they provided a system that depends on the group of models to make the prediction rather than an individual model. Liu et al. (Liu, et al., 2019) proposed a Derived Mean Complete Local Binary

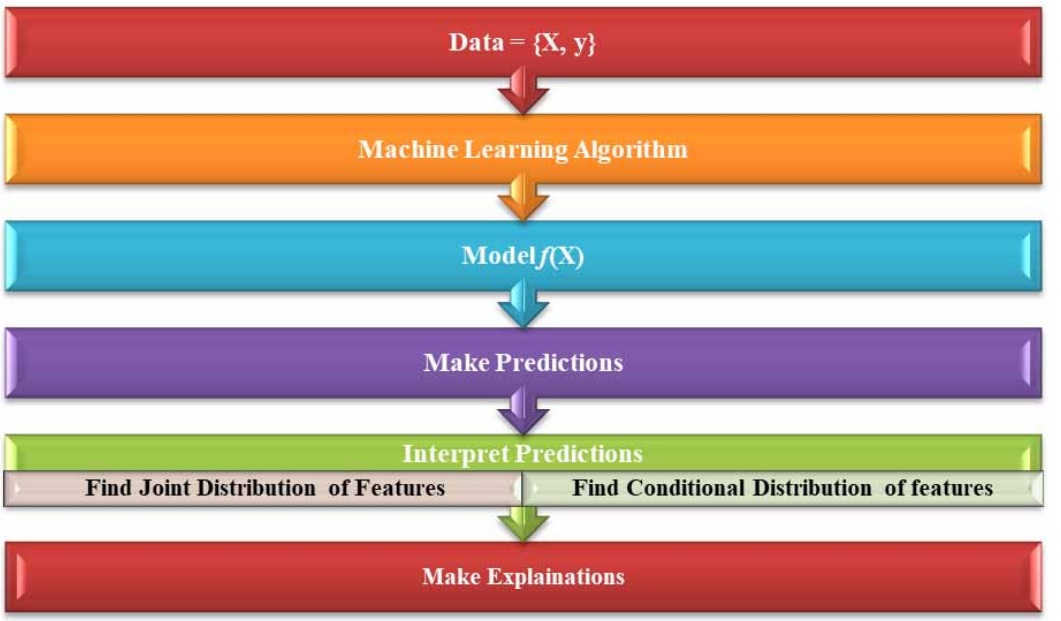
Pattern algorithm to diagnose thyroid disease by extracting features of Thyroid MR Imaging. They used Local Binary Pattern (LBP) and Derived Mean Complete Binary Pattern (DMLBP) algorithm and found DMLBP as most efficient with 94.4%. Chandel et al. (Chandel, Kunwar, Sabitha, Choudhary, & Mukherjee, 2017) proposed a comparative study on thyroid disease classification using KNN and Naïve Bayes. They used the TSH, T4U, and goiter as parameters to analyze the models. KNN gave the highest accuracy of 93.44%.

Even though machine learning and AI have been extensively used in the field of medicine and diagnostics, we found that there is still a gap that exists in the prediction of Thyroid Disease using explainable machine learning. Hence, in this work, we aim to bridge the gap by predicting the accuracy of decision tree algorithms with an explanation of important features, models' accuracy, and making the model more accountable to show the reason behind every decision.

PROPOSED METHODOLOGY

Figure 1 presents the methodology used in the presented work. Logistic regression has been used for thyroid disease prediction. It has good accuracy in predicting thyroid disease, but the model is unable to state the contribution of each feature in the final output. This scenario makes the prediction not fit for deciding the treatment. To increase the trust in the model and provide more information, a model agnostic post hoc Shapley value-based approach is then applied for model interpretation.

Figure 1. Proposed Methodology



The approach provides the signed impact value of each feature on the final output. The feature importance proves beneficial in the disease treatment. In the present work, the Benchmark Thyroid Disease Dataset is taken from the UCI ML Repository.

DATASET DESCRIPTION

Dataset has been taken from Graven Institute in Sydney, Austria from UCI-KDD[1]. This repository contains many datasets from which the "new-thyroid" dataset has been used in the present study. The original dataset contains 5 attributes and 215 instances with 3 classifications, "Normal", "hyper" and "hypo" as given in Table 1.

Table 1. The dataset description with classes.

Classes	No. of instances
Normal as "1"	150
hyper as "2"	35
hypo as "3"	30

Due to a relatively small number of instances in the hyper and hypo class of the dataset, the classification has been simplified into two classes: "normal" as "1" and "having thyroid" as "0" class. The class distribution of the modified dataset is given in Table 2.

The six different feature/subjects of the dataset are:

Table 2. Modified dataset class distribution

Class	No. of instances
Normal as "1"	150
thyroid as "0"	65

1) Classes: Output Class
2) T3-resin uptake: It is a blood to calculate thyroid function. Normal Value: 24-37%
3) Total Serum Thyroxine: Values of more than 11.7 mcg/dL indicates acute thyroid
4) Total serum triiodothyronine: Typically ranges between 80-180 ng/dl
5) Basal thyroid-stimulating hormone (TSH): Normal range is 0.4 to 4.0 milli-international units per liter
6) Maximal absolute difference of TSH value after injection of 200 μg of the thyrotropin-releasing hormone as compared to the basal value [26]

Analysis of this dataset revealed that "Total Serum Thyroxin" (T4) and "Total serum triiodothyronine" (T3) are important features. Table 3 presents the statistics of the Thyroid dataset. Different features are compared to mean, standard deviation, and percentiles values on both the classes are shown in Table 3.

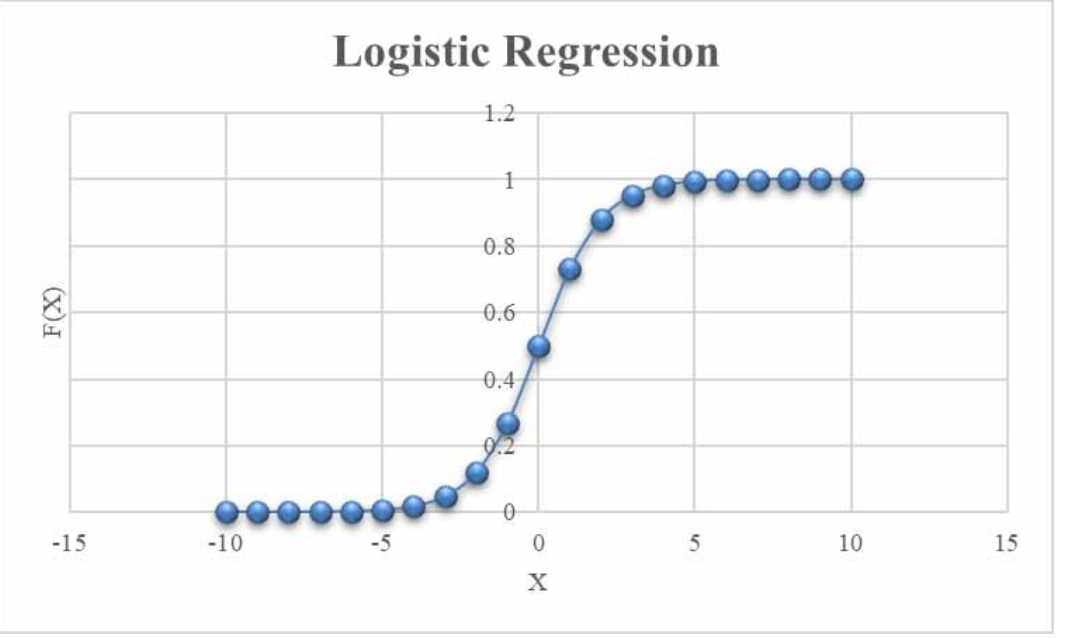
PREDICTIONS THROUGH MACHINE LEARNING

Logistic regression is used for the prediction of thyroid disease. The logistic regression is named after logistic function (sigmoidal functions), commonly used as a binary classifier for non-linear separable data as shown in Figure 2. Mathematically it is represented as in Eq. 1.

Table 3. Dataset Statistics

Features	T3-resin uptake test		Total Serum thyroxin		Total serum triiodothyronine		basal thyroid-stimulating hormone (TSH)		Maximal absolute difference of TSH	
Class	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Count	65	150	65	150	65	150	65	150	65	150
Mean	107.3	110.5	11.14	9.21	2.7621	1.734	6.41	1.31	7.97	2.52
Std	20.31	8.09	7.77	2.044	2.3190	0.474	10.21	0.49	13.58	1.979
Min	65	90	0.5	4.2	0.2	0.4	0.1	0.3	-0.6	-0.7
25%	92.5	105	3.725	7.8	1.1	1.4	0.925	1	-0.075	1.2
50%	109	110	11.55	9.2	1.85	1.7	1.4	1.3	0.35	2.2
75%	120.75	116	17.35	10.4	3.775	2	9.025	1.6	9.325	3.7
Max	144	133	25.3	16.1	10	3.1	56.4	3.7	56.3	13.7

Figure 2. Logistic Regression function as a non-linear binary classifier



$$y = f\left(X\right) = \frac{1}{1 + \exp\left(-X\right)} \tag{1}$$

Where y is a function of X, and X is the set of independent variables.

A linear model treats outputs as a continuous variable and interpolates the values using the best fit hyperplane. This hyperplane is unable to model the probability, hence a linear model is extended

using a logistic function. So, the logistic functions introduce a non-linearity in the model and limit the output in (0,1). The higher the value of *y*, the more will be probable the event becomes. A linear model as shown in Eq. 2 is modeled in logistic regression as shown in Eq. 3, where β are the parameters.

$$y = \beta^T X \tag{2}$$

$$y = \frac{1}{1 + \exp\left(-\beta^T X\right)} \tag{3}$$

If the value of *X* increases let's say by 1 unit, the value of *y* will be affected by the *exp*(- β) factor. The decision-making is directly proportional to exponential factors as shown in Eq. 4.

$$\frac{P\left(y=1\right)}{1-P\left(y=1\right)} = \exp\left(\beta^T X\right) \tag{4}$$

The testing and validation accuracy of the model is 90.77% and 87.33% respectively. The classification report of Logistic Regression Classifier is given in Table 4

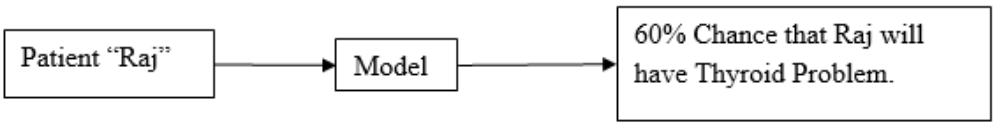
Table 4. Classification Report of Logistic Regression Classifier

	Precision	Recall	F1	Support
0	1.00	0.68	0.81	19
1	0.88	1.00	0.94	46
Accuracy	-	-	0.91	65
Macro Average	0.94	0.84	0.88	65
Weighted Average	0.92	0.91	0.90	65

INTERPRETATION OF PREDICTION

The system up to this point is unable to answer the role of each feature in the result. Hence, it lacks interpretability. Interpretability of the machine learning models refers to the understanding of 'how the system gets this result'. Many different types of approaches like LIME and Partial Dependency

Figure 3. Basic prediction model in AI



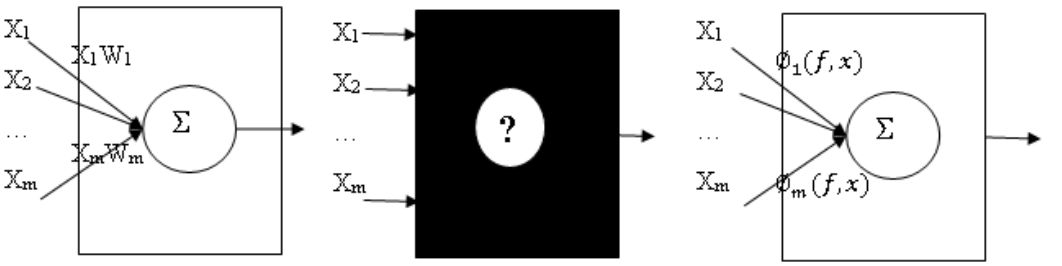
plot can be used for model interpretation. This work uses Model agnostic, Post hoc, a coalition Game theory-based Shapley value-based approach for interpretation of results.

As an example, suppose we use the data from a new patient "Raj" with the trained machine learning model, which leads to a prediction that there is a 60% chance that "Raj" may have a Thyroid problem as shown in Figure 3. But the model is not able to identify the factors that drive the risk of Thyroid. Again, there is a tradeoff between accuracy and interpretability. Complex AI models are less interpretable but more accurate whereas simpler models are interpretable but less accurate.

For the healthcare sector, accuracy is important, but the interpretability must be there as this information directly translates into the understanding of the major risk factors or causes of the disease and can support the selection of necessary therapeutic interventions.

One way to achieve interpretability is to create an accurate model based on simple linear methods, but this is not always feasible. Hence, it is important to find ways of increasing the interpretability of the complex model. Explaining the whole model at a time is very difficult, instead, a single prediction can be used at a time, which will lower down the complexity of the model. Thus, the healthcare practitioner can explain to the patient, what is going on that affected him.

Figure 4. A simple mode (left); Complex model as a black box (middle); Interpretability of complex model (right)



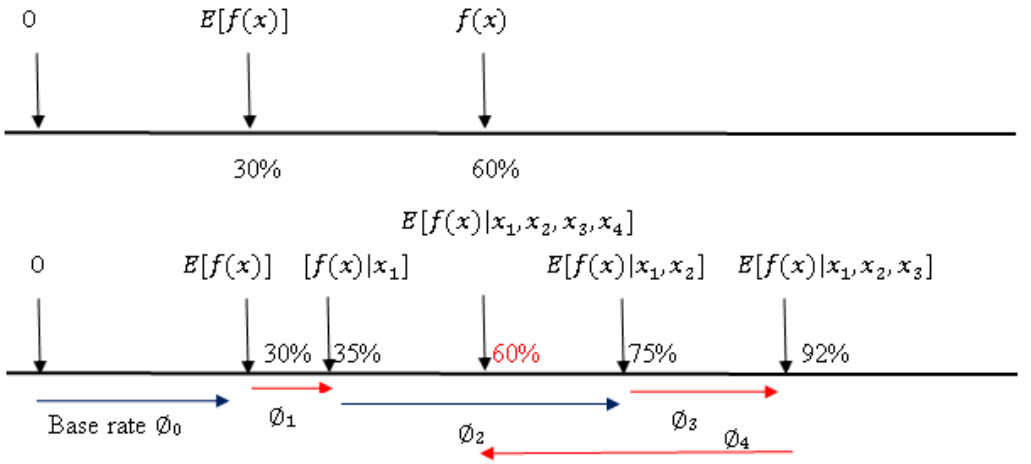
Consider the simple model as shown in Figure 4 (left), one can easily interpret the contribution of each feature using the weights. But the complex models (like neural networks, random forests, etc.) are usually black boxes as shown in Figure 4 (middle) where lots of tensors and their products are available. To explain the decision, instead of focusing on the link weights, SHAP (Lundenberg & Lee, 2017) used $\varnothing\left(f,x\right)$, a function of model *f* and current input *x* to interpret the contribution of feature as shown in Figure 4 (right).

Let us start with the patient "Raj" prediction model as shown in Figure 5(a). Let the base rate or expectation of the model, which signifies how often patients suffered from Thyroid on average, is 30%. Prediction by the model for "Raj" is 60%. To understand how "Raj" is so special, how he got from base rate to current prediction, 30% variability must be explained. To interpret the predictions, let us observe the conditional expectations. Suppose that data has 4 features, and since the order matters, there are 4! ways to allot the features. Calculating the model expectation, the joint distribution of all the features comes into play, for simplicity assume that features are independent.

The E(*f*(*x*)] is calculated by averaging the output of the data set. To calculate the contribution of feature x_1, plug all the values of features and take the average as shown in Eq. 5:

$$\varnothing_1\left(f,x_1\right)=E[f\left(x\right)\mid x_1] \tag{5}$$

Figure 5. Model output with "Raj's Prediction" (upper); Conditional expectation of model with features (lower)



In a similar fashion, $\varnothing_2$, the contribution of x_2 is calculated by averaging the model's prediction conditioned on x_1 with given x_2 as shown in Eq. 6.

$$\varnothing_2\left(f,x_2\right)=E[f\left(x\right)\mid x_1,x_2] \tag{6}$$

Similarly, the contribution of all the features $\varnothing_i$ are calculated. So, $\varnothing_1$ increase the risk of Thyroid from a base rate of 30% to 35%. $\varnothing_2$ increase the risk when symptoms x_1 are considered, to 75%. $\varnothing_3$ increase the risk when symptoms x_1 and x_2 are considered to, 92%. $\varnothing_4$ decrease the risk when symptoms x_1,x_2 and x_3 are considered, to 60%.

Figure 6. mean SHAP Feature Importance

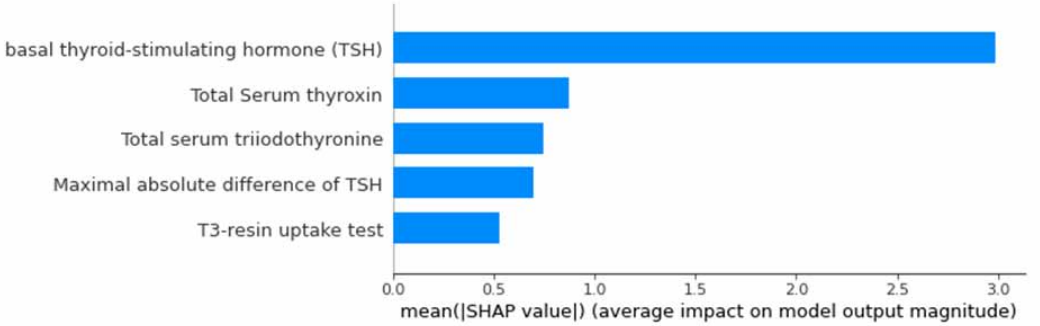
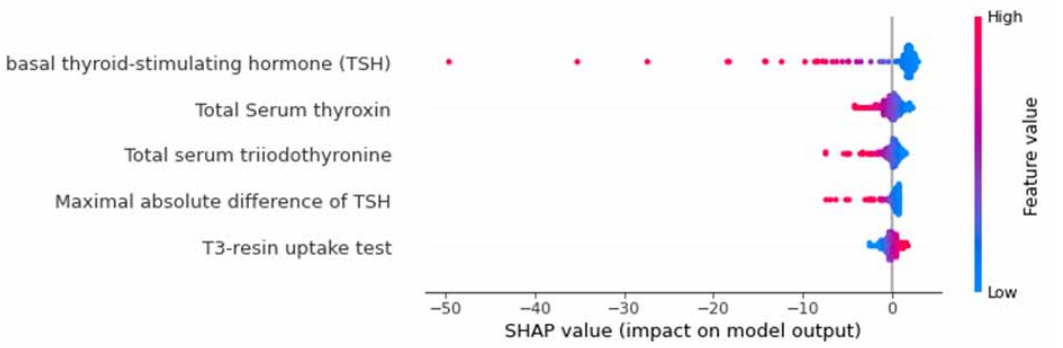


Figure 7. Summary plots with the impact of features on model output

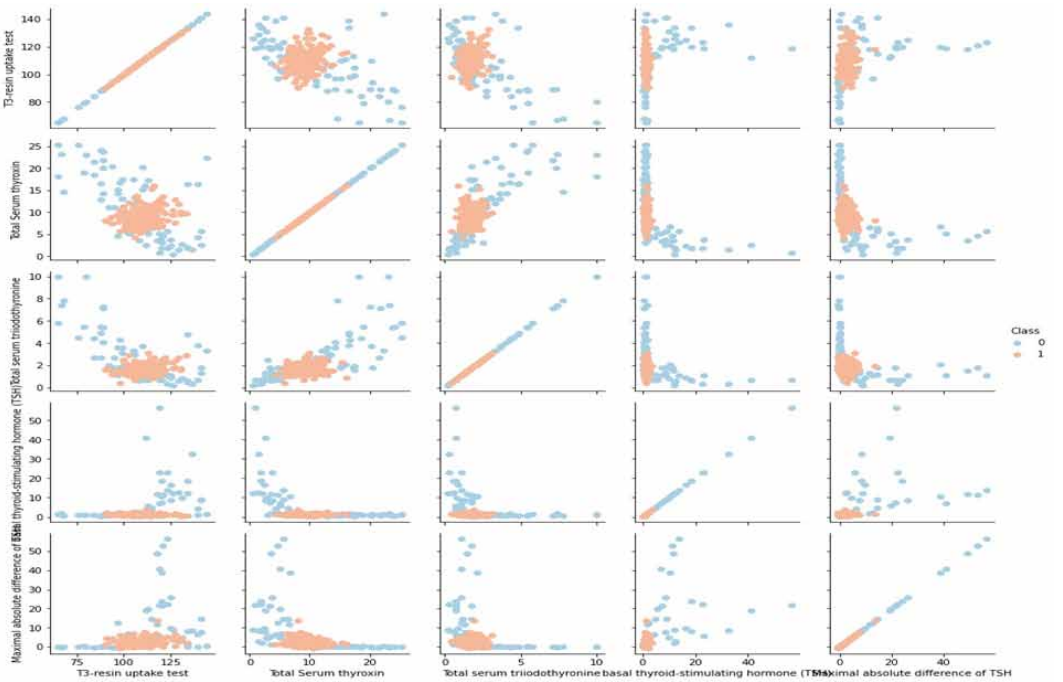


RESULTS AND ANALYSIS:

SHAP values are used in this paper for the local and global interpretation of the model. Figure 6 represents the SHAP feature importance for the logistic regression. The higher the SHAP values of the feature, the higher its importance in prediction. Note that the features are sorted in descending order. In this study, it is found that the basal "basal thyroid-stimulating hormone (TSH)" is the most important feature, changing the mean absolute predicted value to 3.0

Figure 7 plots the SHAP values used to plot the summary of features. It shows the impact of important features in an ordered way by combining feature significance with feature effect. Each point in the plot reflects the instance value of each feature. The location on the vertical axis is determined by the feature and the Shapley value of each instance is represented on the horizontal axis. The plots

Figure 8. Scatter Plot of the features



signify the impact of feature values on the prediction of the machine learning model. The summary plot indicates that low TSH increases the risk of Thyroid whereas high TSH value lower downs the risk of Thyroid. The deviation on the y-axis represents the overlapping values of the instances.

Figure 8 presents the correlation between features. The figure suggests that the features are independent of each other.

A waterfall plot is used to understand how feature affects the learning model prediction from prior probability using unknown distribution of data to the joint probability distribution of all the features. Here, figure 5 analyzes the model's prediction on the 214th instance. The vertical axis of the plot encodes features and indicates the values of instance number 214 of the dataset. The horizontal axis of the plot indicates the range of responses. The expected value of the model indicated at the bottom of the figure is -0.208. This value results from the null model. The "maximum absolute difference" feature lowers down the model prediction by 0.2 with instance value 102. Similarly, the "T3 resin uptake test" feature has more impact on the prediction and lowers down the expectation of the model by 0.47 with instance value 102. It is interesting that "basal thyroid-stimulating hormone (TSH)" with value 1.3 improves the expectation of the model by 1.86.

Figure 9. Waterfall plot for instance number 214

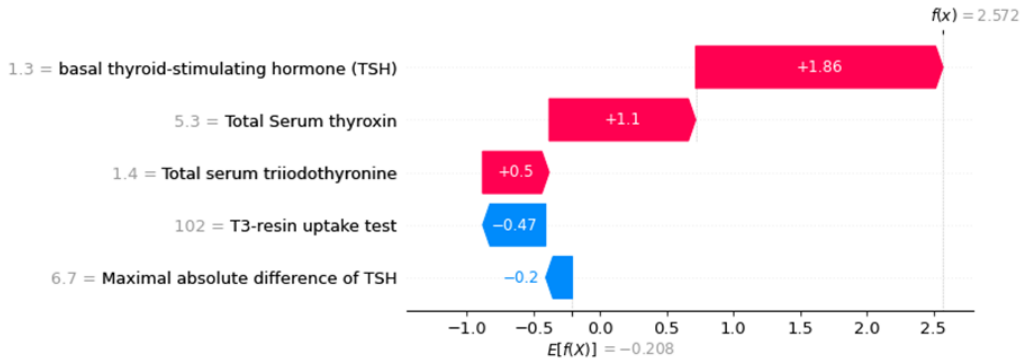
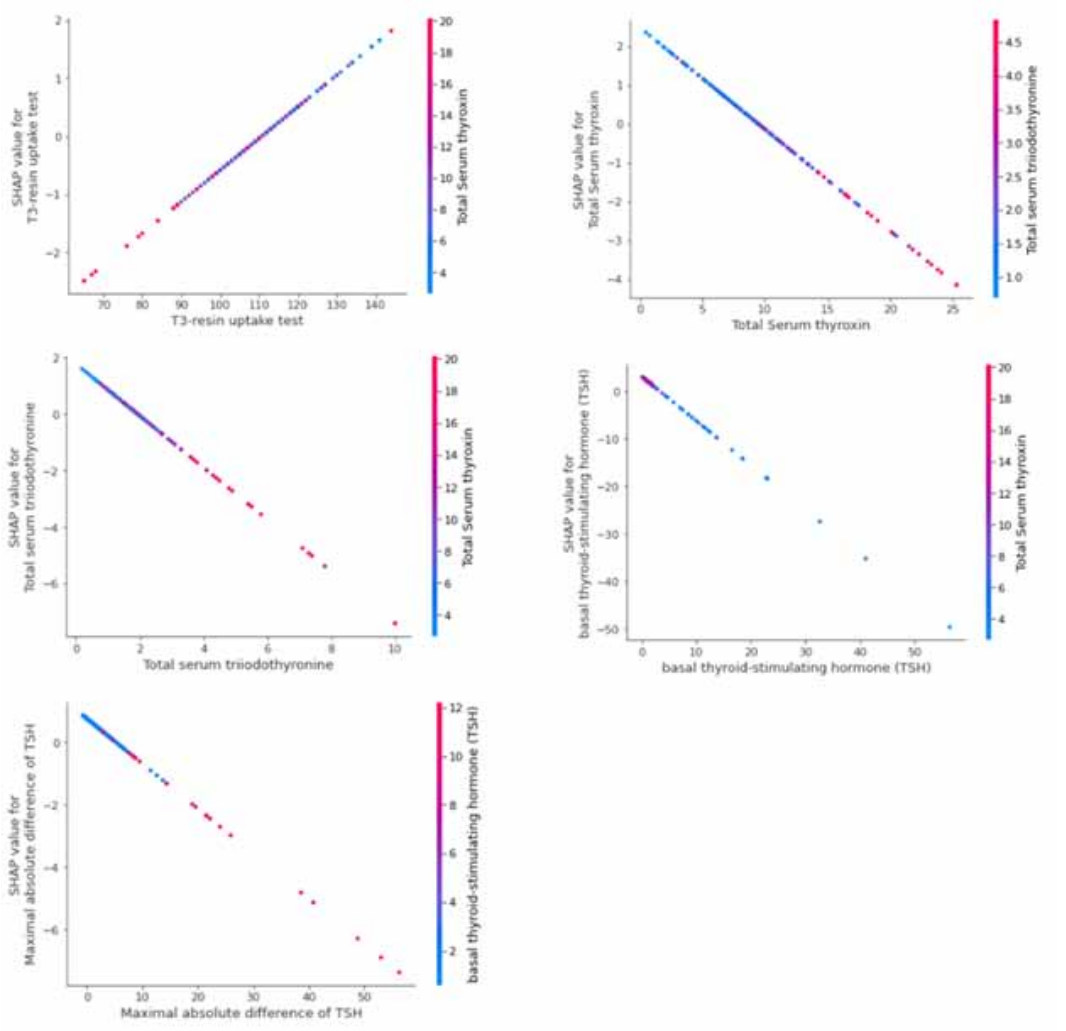


Figure 10 shows the multiple dependency plot. This is the method used for the global interpretation of the model. This plot has discussed the impact of single features on the model prediction. These are very helpful and informative in real-world cases. SHAP dependence plots are an alternative to partial dependency plots (PDP). PDP plots the average model predictions for diverse values of the given feature. SHAP dependence also demonstrates the variance on the y-axis. Consider the multiple dependency plot shown in Fig 6. Each point represents an instance of the data sets. The x-coordinate represents the actual value whereas the y-coordinate represents the effect of instance on the disease prediction.

The slope upwards suggests that the higher the value of the feature, the higher will be the model prediction whereas the slope downwards suggests higher the value of the feature, the lower will be the model prediction. Figure 10(a) suggests that the higher the value of the "T3 resin uptake test", the higher the chances of "Total Serum Thyroxin". Similarly, Figure 10(b) suggests that the higher the value of "Total Serum Thyroxin", the lower the chance of having "Total Serum Triiodothyronine". Since a regression model is used, the model produces a plot in a line instead of having a spread. In such a case, color mixture coding helps to understand the interaction with other features. In Figure 10(a), the overlapped color codes suggest that the "T3 resin uptake test" immersed into interaction with other features. Similarly, from Figure 10(b-e) "Total Serum Thyroxin" immersed into interaction with "Total Serum Triiodothyronine", "basal thyroid-stimulating hormone (TSH)" immersed into

interaction with) "Total Serum Thyroxin", and "maximal absolute difference of TSH" are immersed into interaction with "basal thyroid-stimulating hormone (TSH)".

Figure 10. Multiple dependency graph of different features



Usually, machine learning models are black-box models and many of them are complex. Another thing that comes in the interpretability of models is how the models are making decisions especially when models are complex. Figures 11-13 demonstrate the decision plots to understand the decision process. Figure 11 shows the decision plot on a single instance of the dataset. The plot is centered on the x-axis that represents the model's prediction, while the y-axis represents different features of the dataset. From the bottom of the plot, start with the model base value and how a feature contributes to decision-making for each instance. In Figure 11, the plot starts with model base value 0.45 and ends with the value closer to 1, representing the instance belongs to class 1. "Maximal absolute difference of TSH" with value 6.7 and "T3 resin uptake test" with value 102 leading the model decision towards class 0, but due to the high values of remaining feature, leads the decision towards class 1.

Figure 11. Decision Plot for a single instance

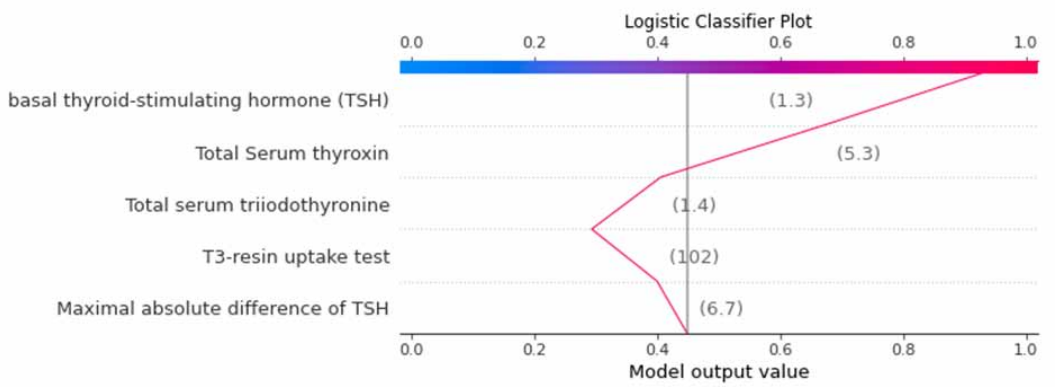


Figure 12. Decision plot for all instances

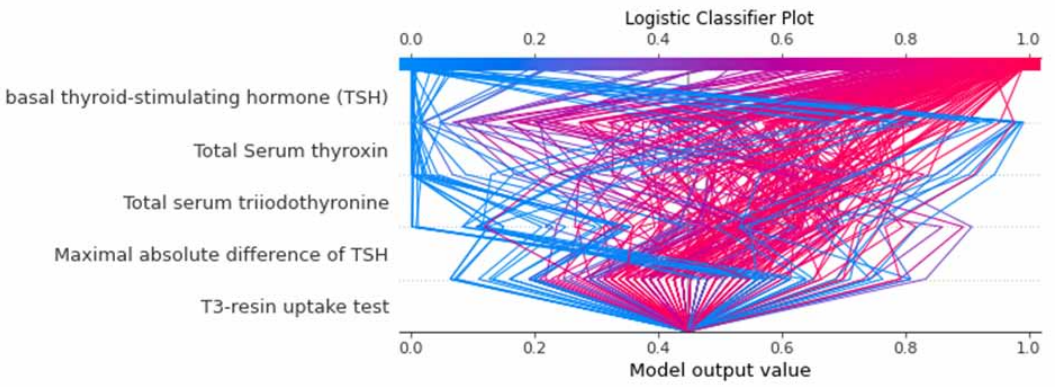


Figure 12 represents the decision plots with all instances, where blue lines represent prediction with an inclination towards class 0, whereas red presents the inclination of decision towards class 1. There might be chances of dilemma that the same instance values leading to both the classes. Such instances can be easily identified and represented in Figure 13.

Figure 14 represents the force plot to interpret the prediction made by the machine learning model on instances from each class. In Figure 14 (upper), the basal TSH and maximal absolute difference of TSH are offset by total serum triiodothyronine and total serum thyroxin. It decreases the model prediction from its base value of 0.4097 to 0.19, thus reducing the chances of the thyroid. Whereas in Figure 14 (lower), the higher values of the basal TSH with total serum triiodothyronine and total serum thyroxin increase the chance of thyroid in the patient.

CONCLUSION

The presented work focuses on the importance of the use of machine learning models for decision support in the medical field. The paper demonstrates a thyroid disease prediction model. Even though machine learning models are often cost-effective, the explainability associated with the models results in a lack of trust and the social acceptability of the model. The paper demonstrates the model-agnostic,

Figure 13. Decision plot for single instances for two different classes

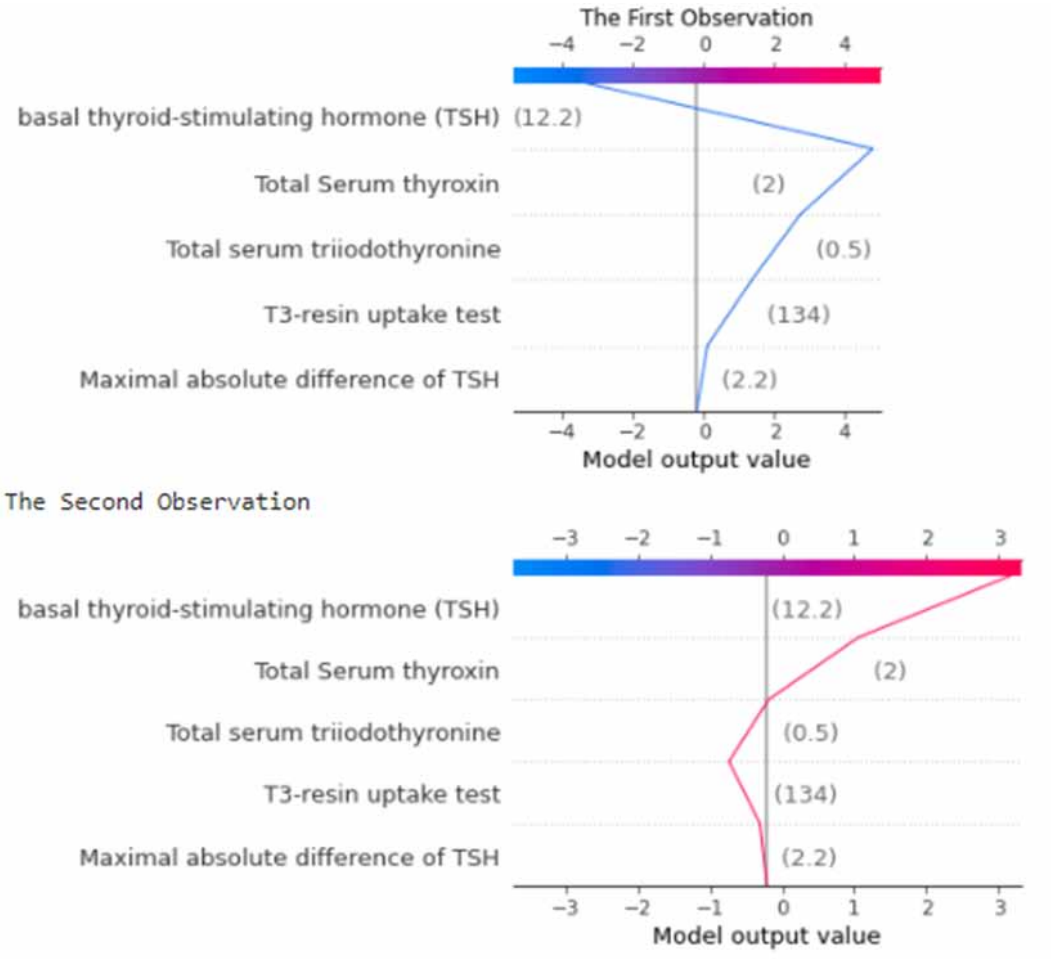


Figure 14. Force plot for understanding the decisions for different classes

post hoc SHAP method for interrelating the model locally and globally. The major contribution of the research is the interpretability of the results. The machine learning models are monolithic and thus capacious field is available for the research to understand the results. The results indicate the main cause of the disease and what is going wrong with the patients. In addition to this, the ordering of

features (causes) is presented in the paper, which suggests the health practitioner where to start the diagnosis. The basal thyroid-stimulating hormone (TSH) and total serum thyroxin are concluded as the most important features. The lower values of these tests indicate a higher risk of thyroid in the patient. In concluding remarks, machine learning models are now capable of achieving higher accuracy, but in real-world scenarios understandability of those results is more important for the stakeholders.

REFERENCES

Adadi, A., & M, B. (2020). Explainable AI for Healthcare: From Black Box to Interpretable Models. In *Embedded Systems and Artificial Intelligence* (pp. 327-337). Springer. 10.1007/978-981-15-0947-6_31

Chakraborty, D., Awolusi, I., & Gutierrez, L. (2021). An explainable machine learning model to predict and elucidate the compressive behavior of high-performance concrete. *Results in Engineering*, *11*, 100245. Advance online publication. doi:10.1016/j.rineng.2021.100245

Chandel, K., Kunwar, V., Sabitha, S., Choudhary, T., & Mukherjee, S. (2017). A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. *CSI Transactions on ICT*, *34*(2-4), 313–319. doi:10.1007/s40012-016-0100-5

Choubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2021). Thyroid Disease Prediction Using Machine Learning Approaches. *National Academy Science Letters*, *44*(3), 233–238. doi:10.1007/s40009-020-00979-z

Felzmann, H., Villaronga, E. F., Lutz, C., & Larrieux, A. T. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. doi:10.1007/s11948-020-00276-4 PMID:33196975

Fu, C., Chang, W., Liu, W., & Yang, S. (2019). Data-driven group decision making for diagnosis of thyroid nodule. *Science China. Information Sciences*, *62*(11), 212205. Advance online publication. doi:10.1007/s11432-019-9866-3

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artificial Intelligence and Law*, *28*(4), 415–439. doi:10.1007/s10506-020-09260-6

Islam, M. R., Barua, S., Begum, S., & Ahmed, M. U. (2019). Hypothyroid Disease Diagnosis with Causal Explanation using Case-based Reasoning and Domain-specific Ontology. In *International Conference on Case-based Reasoning (ICCBR-19)*. European Academy.

Kenny, E., & Keane, M. (2019). Twin-Systems to Explain Artificial Neural Networks using Case-Based Reasoning: Comparative Tests of Feature-Weighting Methods in ANN-CBR Twins for XAI. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 2708-2715). International Joint Conferences on Artificial Intelligence Organization. doi:10.24963/ijcai.2019/376

Liu, Z., Qiu, C., Song, Y., Liu, X., Wang, J., & Sheng, V. (2019). Texture Feature Extraction from Thyroid MR Imaging Using High-Order Derived Mean CLBP. *Journal of Computer Science and Technology*, *34*(1), 35–46. doi:10.1007/s11390-019-1897-9

Lundenberg, S., & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems. NIPS.

Messalas, A., Kanellopoulos, Y., & Makris, C. (2019). Model-Agnostic Interpretability with Shapley Values. *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1-7. doi:10.1109/IISA.2019.8900669

Moradi, M., & Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, *165*, 113941. Advance online publication. doi:10.1016/j.eswa.2020.113941

Pawar, U., O'Shea, D., Rea, S., & O'Reilly, R. (2020). Incorporating Explainable Artificial Intelligence (XAI) to aid Understanding of Machine Learning in the Healthcare Domain. *Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2020)*.

Pawar, U., O'Shea, D., Rea, S., & O'Reily, R. (2020). Explainable AI in Healthcare. In *International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*. IEEE. doi:10.1109/CyberSA49311.2020.9139655

Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, *7*(1), 9. Advance online publication. doi:10.1057/s41599-020-0501-9

Rong, G., Mendez, A., Assi, E. B., Zhao, B., & Sawan, M. (2020, March). Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering*, *6*(3), 291–301. doi:10.1016/j.eng.2019.08.015

Samek, W., & Muller, K. R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. Hansen, & K. R. Muller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (Vol. LNCS 11700, pp. 5-22). Springer Cham. doi:10.1007/978-3-030-28954-6_1

Schmidt, J., Marques, M. R., Botti, S., & Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *NPJ Computational Materials*, *5*(1), 83. Advance online publication. doi:10.1038/s41524-019-0221-0

Teixeira, P., Santos, P. B., & Pazos-Moura, C. C. (2020). The role of thyroid hormone in metabolism and metabolic syndrome. *Therapeutic Advances in Endocrinology and Metabolism*, *11*. Advance online publication. doi:10.1177/2042018820917869 PMID:32489580

Temurtas, F. (2009). A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications*, *36*(1), 944–949. doi:10.1016/j.eswa.2007.10.010

Thomas, J. M., & Haertling, T. (2020). AIBx, artificial intelligence model to risk stratify thyroid nodules. *Thyroid*, *36*(6), 878–884. doi:10.1089/thy.2019.0752 PMID:32013775

World Thyroid Day is Heralded by International Thyroid Societies. (2015, May 15). Retrieved from American Thyroid Association: https://www.thyroid.org/world-thyroid-day-is-heralded-by-international-thyroid-societies/

Yang, G., Ye, Q., & Xia, J. (2021). *Unbox the Black-box for the Medical Explainable AI via Multi-modal and Multi-centre Data Fusion: A Mini-Review, Two Showcases and Beyond.* arXiv:2102.01998 [cs.AI].

ENDNOTE

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/thyroiddisease/

Siddhartha Kumar Arjaria is B.E., Mtech, and Ph.D in computer science and Engineering. He has over 14 years of experience of teaching in institutes like BITS pilani Goa Campus, IIITM Gwalior. Currently, He is working as an Assistant Professor - IT Department in Rajkiya Engineering College, Banda, U.P, India. He has published over 20 papers/Book chapters in various national/international journals/Conference Proceedings. He is a reviewer of many international journals/conferences. He has organized many FDPs/Workshops as Co-Ordinator/Convenor. His research interests include Machine learning, Data Analytics and text mining.

*Dr. Rathore an established researcher in CSE, has over 12 years of experience as faculty, researcher and software engineer. He is currently working in the Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore. He is an alumni of IET, Indore and completed his Ph.D from NIT, Bhopal. He had taught dozens of workshops in Computer Science over number of universities in India. His significant research work contributions are in the areas of Machine Learning, Information Science, Agent Oriented Software Engineering. He is member of number of agent development groups, including i*wiki and JADE.*

Gyanendra Chaubey is working as Software Engineer in HCL Technologies Ltd. He has completed his B.Tech in Information Technology from Rajkiya Engineering College Banda. He has 4 published research papers in renowned journals and conferences. His area of interest is Machine Learning, Deep Learning, Data Mining, Data Science and Artificial Intelligence.