Explainable AI Diagnosis for Hypothyroidism

Rosana C. B. Rego
Department of Engineering and Technology
Federal University of Semi-Arid
rosana.rego@ufersa.edu.br

Caio M. V. Cavalcante
Computational Intelligence Laboratory
Federal University of Semi-Arid
caio.cavalcante@alunos.ufersa.edu.br

***Abstract*—The application of Artificial Intelligence (AI) in medical diagnosis has shown promising results. However, the lack of transparency in AI models poses challenges in comprehending and trusting their decisions, impeding their widespread clinical adoption. In this paper, we implemented the Random forest model for predicting hypothyroidism. Furthermore, we presented a transparent AI approach for diagnosing hypothyroidism. We applied Explainable AI (XAI) techniques, such as SHAP, LIME, DALEX, and SHAPASH, to explain the factors influencing hypothyroidism diagnosis by AI. The XAI methods emphasized the importance of features such as thyroid-stimulating hormone (TSH) and the level of Total Thyroxine (TT4) for the detection of hypothyroidism. The results highlight the importance of transparent AI approaches in enhancing the interpretability and accountability of AI systems in healthcare.**

***Index Terms*—Hypothyroidism, Explainable AI, Machine learning, Thyroid, Artificial Intelligence**

I. Introduction

The World Health Organization (WHO) reports that thyroid disorders are the second most common endocrine disorder globally, after diabetes. This highlights the substantial prevalence and impact of thyroid illness on a global scale [1]. Hypothyroidism are one of the most frequent thyroid gland illnesses [2]. The diagnosis of hypothyroidism generally encompasses clinical evaluation, laboratory tests, and the assessment of symptoms alongside medical history. Machine learning models show potential in the early detection of hypothyroidism. However, previous studies have not adequately designed models in a manner that facilitates human comprehension and interpretation of their decisions.

Explainable Artificial Intelligence (XAI) methods play a crucial role in enhancing the understanding of hypothyroidism diagnosis by machine learning models. We applied SHAP (SHapley Additive exPlanations) [3], SHAPASH (SHapley Additive exPlanations for Automated Science), LIME (Local Interpretable Model-agnostic Explanations) [4], and DALEX (Descriptive mAchine Learning EXplanations) [5] for enhancing the transparency and interpretability of model diagnosis.

II. Methods

A. Dataset

We used the Hypothyroid subset of the Thyroid Disease dataset sourced from the University of California Irvine (UCI) machine learning repository. The dataset has 29 attributes. We implemented a feature selection algorithm, and the features selected are thyroid-stimulating hormone (TSH), Total thyroxine (TT4), the level of Free Thyroid Index (FTI), Free thyroxine (T4U), triiodothyronine (T3), and the information if the patient is pregnant or not.

Thanks to UFERSA for financial support in granting a Scientific Initiation scholarship and UFERSA/PROPPG 65/2022 (PAPC) support for research groups.

B. Machine learning model

We applied the Random Forest (RF) classifier algorithm, configuring it with 200 decision trees. Each tree was constrained to a maximum depth of 20 levels. Additionally, we employed sampling criteria, requiring a minimum of 20 samples for node splitting and 1 sample for leaf node formation. The model was implemented with the scikit-learn library version $\geq 1.3.0$.

C. Explainable Artificial Intelligence methods

1) SHAP: is based on game theory and provides a suitable approach to explain the output of the machine learning model. For a set of M explanatory variables, the SHAP method approximates each prediction locally in terms of accuracy. It represents the prediction $f(x)$ by $g(\hat{x})$, along with a quantity ϕ_j, which is defined as follows:

$$f(x) = g(\hat{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j \hat{x}_j \tag{1}$$

For feature j in explaining the prediction $f(x)$ for instance x in a model F can be calculated using the formula

$$\phi_i(x) = \sum_{S \subseteq \{1,\ldots,M\} \setminus \{j\}} \frac{|S|!(M-|S|-1)!}{M!} [f(x_{S\cup\{j\}}) - f(x_S)] \tag{2}$$

where S Represents subsets of features excluding j, M is the total number of features, and $x_{S\cup\{j\}}$ is a instance x with feature j set to its value in x, while other features remain unchanged.

2) SHAPASH: is a Python library that extends SHAP's capabilities for interactive model interpretation.

3) LIME: explains the predictions of the machine learning model by approximating it with interpretable models (e.g., linear models) around specific data points. Each observation x of LIME can be calculated by

$$\gamma(x) = argmin_{q \in Q} L(f, q, \pi_x) + \Gamma(q) \tag{3}$$

where L denotes the locality aware loss, while Q represents potentially interpretable models and $\pi_x(s)$ signifies the distance between an instance s and x. Furthermore, $\Gamma(q)$ serves

as a metric assessing the complexity of the explanation, with q being a member of the set Q [6].

4) DALEX: is a framework for explaining machine learning models including feature importance plots, partial dependence plots, and accumulated local effects plots.

III. Results

Table I shows model metrics. The values indicates that the RF model generalizes well to unseen data, as evidenced by the similar performance metrics between the training and testing datasets.

TABLE I
MODEL METRICS

Model	***Accuracy***	***Precision***	***Recall***	***F1-score***
RF (Train)	0.9925	0.9992	0.9859	0.9925
RF (Test)	0.9863	0.9863	0.9863	0.9863

Fig. 1, achieved with SHAP, depicts the impact of features on the model output. Features such as TSH and TT4 influence on the model's predictions. High values of TSH and low levels of TT4 are closely associated with a heightened prediction of hypothyroidism. Moreover, elevated levels of T3 also contribute significantly to the model's output.

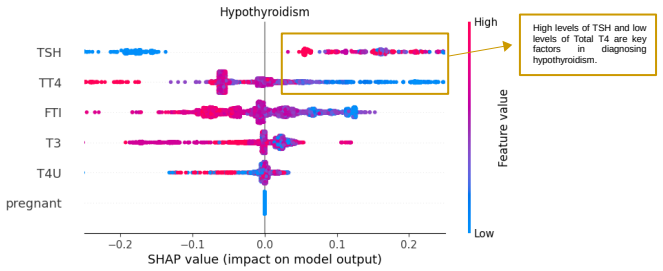
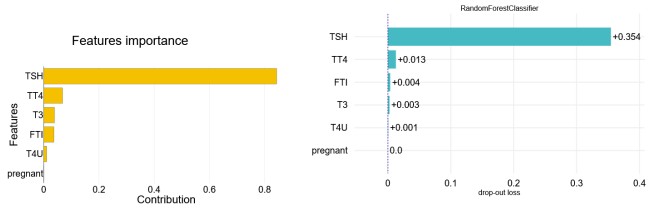


Fig. 1. Features importance and impact on model output for the hypothyroidism prediction.

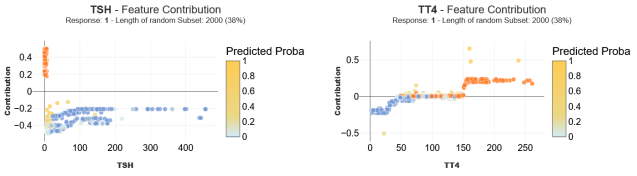
Fig. 2 (a) and (b) show the feature's importance reached with Shapash and DALEX. The results achieved with DALEX are equal to the ones reached with SHAP. Differently, Shapash's outputs suggest that T3 is more important than FTI.



(a) (b)

Fig. 2. Features importance: (a) Shapash and (b) DALEX.

Fig. 3 (a) and (b) depict the contribution of TSH and total T4 to the prediction of hypothyroidism. As shown in Figure (a), when the contribution of TSH is less than or equal to zero, it suggests a higher likelihood of hypothyroidism being present. Contrariwise, when the contribution of TSH is greater than zero, it indicates a lower likelihood of hypothyroidism. Specifically, for high values of TSH and lower values of total T4, the contribution is consistently less than or equal to zero, signifying a strong association with hypothyroidism. This observation implies that elevated levels of TSH are indicative of the presence of hypothyroidism according to the model's predictions.



(a) (b)

Fig. 3. Features contribution on model output for the hypothyroidism prediction: (a) TSH and (b) Total T4.

Fig. 4 shows a local explanation, providing insight into an individual prediction of hypothyroidism based on specific input values.

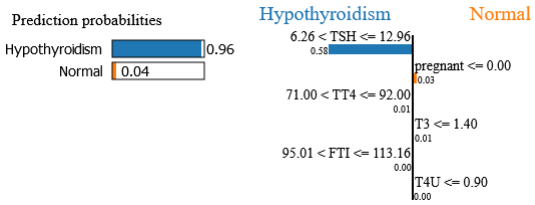


Fig. 4. Individual predictions explanation gives by LIME.

IV. Conclusions

The application of XAI methods in the diagnosis of hypothyroidism represents a significant advancement in medical decision-making. By generating explanations at both global and local levels, we have provided insights into the factors contributing to individual predictions, thus empowering them to make more informed decisions.

References

[1] "World Health Organization," https://www.who.int/, accessed on: 2024.

[2] L. Duan, H.-Y. Zhang, M. Lv, H. Zhang, Y. Chen, T. Wang, Y. Li, Y. Wu, J. Li, and K. Li, "Machine learning identifies baseline clinical features that predict early hypothyroidism in patients with graves' disease after radioiodine therapy," *Endocrine Connections*, vol. 11, no. 5, 2022.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[5] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, and P. Biecek, "dalex: Responsible machine learning with interactive explainability and fairness in python," *Journal of Machine Learning Research*, vol. 22, no. 214, pp. 1–7, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1473.html

[6] A. Gramegna and P. Giudici, "Shap and lime: an evaluation of discriminative power in credit risk," *Frontiers in Artificial Intelligence*, vol. 4, p. 752558, 2021.