IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 June 2024, accepted 16 July 2024, date of publication 22 July 2024, date of current version 31 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3431691

RESEARCH ARTICLE

A Methodology for Reliability Analysis of Explainable Machine Learning: Application to Endocrinology Diseases

FIRAS KETATA[1,2], ZEINA AL MASRY[1], SLIM YACOUB[2], AND NOUREDDINE ZERHOUNI[1]
[1]SUPMICROTECH, CNRS, Institute FEMTO-ST, 25000 Besançon, France
[2]Remote Sensing Laboratory and Information Systems Spatial Reference Laboratory, INSAT, Tunis 1080, Tunisia

Corresponding author: Firas Ketata (firasketata@yahoo.fr)

ABSTRACT Machine learning (ML) has transformed various sectors, including healthcare, by enabling the extraction of complex knowledge and predictions from vast datasets. However, the opacity of ML models, often referred to as ''black boxes,'' hinders their integration into medical practice. Explainable AI (XAI) has emerged as a crucial area for enhancing the transparency and understandability of ML model decisions, particularly in healthcare where reliability and accuracy are paramount. However, the reliability of the explanations provided by ML models remains a major challenge. This mainly concerns the difficulty of maintaining the validity and relevance of the new training and test data explanations. In this study, we propose a structured approach to enhance and evaluate the reliability of explanations provided by ML models in healthcare. We aim to improve the reliability of explainability by combining the XAI approaches with the k-fold technique. We then developed several metrics to assess the generalizability, concordance, and stability of the combined XAI and k-fold approach, which we applied to case studies on hypothyroidism and diabetes risk prediction using SHAP and LIME frameworks. Our findings reveal that the SHAP approach combined with k-fold exhibits superior generalizability, stability, and concordance compared to the combination of LIME with k-fold. SHAP and k-fold integration provide reliable explanations for hypothyroidism and diabetes predictions, providing strong concordance with the internal explainability of the random forest model, the best generalizability, and good stability. This structured approach can bolster practitioner's confidence in ML models and facilitate their adoption in healthcare settings.

INDEX TERMS Explainable machine learning, reliability analysis, concordance, stability, generalizability, medical decision support.

I. INTRODUCTION

Machine learning (ML) has revolutionized many fields owing to its ability to extract complex knowledge and make predictions from large amounts of data. In the medical field, ML offers important prospects in the diagnosis, treatment, and prediction of diseases. The main goal is to support medical decision-making by analyzing clinical, genetic, and image data sets [1]. However, despite these advances, the integration of ML models into medical practice remains limited. A crucial reason for this is that these models are often opaque and known as ''black boxes'' which cannot explain their prediction and decision-making processes [2], [3]. This lack of explainability raises concerns, particularly regarding trust and acceptance by healthcare professionals.

ML explainability, or (XAI), is emerging as a critical area aimed at making ML model decisions transparent, understandable and justifiable to human users [4], [5]. In the healthcare sector, the importance of XAI is particularly

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Kavak.

pronounced, as medical decisions require unrivaled reliability and accuracy [6]. The ability to explain and justify model predictions is an important factor in improving the trust of practitioners and patients, facilitating the adoption of ML in healthcare, and ensuring that ML-based medical decisions are made ethically and responsibly [7].

However, the reliability of explanations provided by XAI approaches remains a critical challenge [8], [9], [10], [11]. These challenges include the difficulty of explanations that remain valid and relevant on new training and testing data. The explanation changes each time the selection of the test and training data is changed, especially for small datasets [14]. Hence, there is ambiguity regarding generalization and the robustness of explainability. In addition, XAI approaches may admit a lack of consistency, where explanations can vary considerably for similar instances, raising concerns regarding their stability. Furthermore, there may be a difference between the importance of features evaluated by the model and the importance reflected in the explanations, calling into question the concordance of the explanations.

Overall, an ML-based prediction provided to practitioners is generally considered opaque information. When this prediction is accompanied by explainability, it becomes clearer. However, explainability can vary when the data selected for testing and training changes. This variability leads to a lack of confidence in the reliability of explainability validation, highlighting the need for improvement and thorough reliability studies. Therefore, we propose to provide practitioners with predictions that include more reliable explainability, validated by the combination with a data sampling technique and assessed using metrics that evaluate the reliability of the provided explainability.

This study aims to develop a structured approach to improve and assess the reliability of explanations provided by XAI approaches in healthcare. The main contributions of this study are as follows:

- Combining the explainable ML approach with a data sampling method to improve the reliability of the explainability.
- We define and develop metrics to assess the reliability of explainability after combining it with a data sampling technique.
- Develop a global metric for reliability assessment of XAI.

Our ultimate goal was to increase practitioners' confidence in ML by exploiting tabular datasets (Biological and Clinical) to predict the risk of abnormalities for each subject to assist physicians in personalizing the screening or treatment of diseases. These include thyroid disorders and diabetes which are the most common endocrine diseases worldwide according to the World Health Organization [15]. It represents a significant global burden in terms of morbidity and mortality, and the ability to accurately predict risk and understand underlying factors through reliable explanations could transform the management and treatment of these conditions [12], [13]. Consequently, several case studies are proposed to test and validate our approach on various datasets and ensure its applicability and effectiveness in real-world scenarios. Through this research, we aim not only to improve the understanding of the decisions made by ML models in healthcare but also to facilitate wider adoption of these technologies by providing reliable, actionable explanations to healthcare professionals.

The remainder of this paper is structured as follows: Section II offers a review of the literature and analysis of related works. Section III describes the methodology of the proposed approach, and Section IV presents the results and discussion. Finally, Section V concludes the paper.

II. RELATED WORK

To evaluate XAI performance, both objective and subjective metrics have been developed [16]. Subjective metrics are based on user feedback such as clarity, transparency, and satisfaction. However, an objective metric is based on mathematical and statistical tools used to assess the reliability of data-driven explainability. Hence, the diversity of metrics depends on the data type and the field of application.

In collaboration with an expert, XAI can achieve optimum accuracy. The author in [17] studied advanced imaging techniques in radiology to detect disease, and showed that the XAI agent achieves 99.5% accuracy. In a field where decisions are critical, it is essential to provide doctors with clear explanations of classification errors. The study proposes four criteria for evaluating the explanations provided by the agent: difference in performance (D) between the agent's model and the explanation's logic, number of rules (R), number of features (F) used to construct the explanation, and stability (S). These criteria highlight the limitations of current studies, which oversimplify the logic of initial models without considering legal, ethical or safety implications. They offer the advantage of not depending on the task or XAI algorithm used.

In the field of recommendation systems, studies often interchangeably use specific terms, as highlighted by Tintarev and Masthoff [18]. They emphasize that aspects such as user transparency, persuasiveness, scrutability, effectiveness, satisfaction, efficiency, and trust are as crucial as traditional accuracy measures such as precision and recall. To gauge the effectiveness of explanations within these systems, they introduced metrics such as transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and user satisfaction. These metrics focus on ensuring that users understand how and why recommendations are made, providing them with the ability to correct the system when necessary, and maintaining trust through clear and effective explanations. This approach highlights the importance of accurate and user-friendly recommendation systems.

The authors in [19] showed that a system's design can influence perceived trustworthiness. Trust is assessed using

methods such as surveys or by evaluating user engagement, which can be observed through metrics such as login frequency or sales. The concept of persuasiveness involves encouraging users to purchase or try something as assessed by their reactions to explanations. Effectiveness enables users to eliminate unsuitable choices through quality explanations, whereas efficiency, particularly relevant to chat systems, gauges the speed of task completion, often measured by the required number of explanations. User satisfaction, which reflects the system's usefulness and user-friendliness, is determined through user feedback metrics.

Doshi-Velez and Kim in [20] introduced various concepts related to the quality of system explanations. They discussed how explanations can empower users to make corrections (actionability and correctability), the link between causes and effects (causality [21]), and the thoroughness of system descriptions (completeness). They also cover the ease of understanding explanations (comprehensibility [22]), the selection of pertinent features (faithfulness), and whether explanations align with expert knowledge (justifiability). The consistency of explanations given similar inputs (robustness [23]), the ability to examine unsuccessful training (scrutability), and the focus on essential explanatory features (simplicity) are also explored. Other aspects such as sensitivity to input changes, explanation stability, and truthfulness (soundness) are considered, with a note on the debate over the priority of completeness versus soundness in [24]. They also mentioned that while many concepts such as transparency, interactivity, and security are discussed within XAI, formal definitions, and practical applications are still lacking.

Authors in [25] introduced seven cognitive metrics in their research: explanation quality, user satisfaction, engagement of user curiosity/attention, trust/reliance on the system, user comprehension, performance/productivity, and system usability/interaction. They suggested that these evaluation techniques are derived from cognitive science principles, involving three main processes: assessing system behavior under various scenarios, creating models to predict behavior, and utilizing factor analysis for behavior explanation based on feature significance. These metrics are predominantly subjective and require user feedback for accurate measurements. The categorization of evaluation methods targets specific user groups, including AI beginners, domain experts, and AI professionals. Key interpretability metrics focus on the user's cognitive model, the utility and impact of explanations, trust in the system, and overall performance in human-AI collaborative tasks studied in [26].

Computational evaluation encompasses measures such as explanation accuracy, which closely ties model consistency, reliability of explanations, and trustworthiness of models, independent of human-based studies. For time series classification, aspects such as explanation stability, robustness of the classification model, and computational demands of explanation methods are crucial. The authors in [27] analyzed the effectiveness of saliency maps in providing explanations that identify critical components for predictions within time series data. A truly informative explanation highlights parts that are crucial for prediction. The stability and robustness of these explanations are tested through repeated trials, assessing their resilience to changes and the computational resources required for generating such explanations.

The authors of [28] discussed explainability in AI as a blend of interpretability (how understandable the explanations are to humans) and fidelity (how accurately these explanations reflect the model's behavior). They argue against the feasibility of universal computation metrics for evaluating XAI methods owing to factors such as the subjective nature of explanations, varying contexts, dependencies on users and models, and specific types of explanations needed. They categorized objective evaluation metrics into three groups: model-based, attribution-based, and example-based explanations. Model based explanations involve the use or creation of models to elucidate ML algorithms, with metrics such as model size, interaction strength, or complexity. Attribution-based metrics focus on the significance or ranking of features by employing metrics such as monotonicity or sensitivity [29].

Arrieta et al. [30] proposed the development of specific evaluation metrics for future improvement, focusing on the quality, utility, and satisfaction derived from explanations, enhancing the mental model of the audience, and the impact of explanations on model performance and user trust. They mentioned tools such as goodness checklists, satisfaction scales, and computational measures to assess explainer fidelity and reliability, as described in [31] and [32].

The study also discusses the use of Bayesian Networks in various applications, highlighting BayLime as an enhancement of the LIME technique to address instability and improve consistency and robustness through Bayesian reasoning [33]. The necessity of explaining Bayesian networks, especially in legal contexts, is acknowledged [34].

In [35], a review of explanation methods for Bayesian networks was presented, outlining essential properties such as content, communication, and adaptation to user needs. They emphasize the importance of explaining the knowledge base, reasoning process, and evidence-justifying conclusions. The communication aspect involves the presentation of explanations, including the format and expression of probabilities. Adaptation refers to tailoring explanations to a user's level of knowledge and the details required. Although various tools exist for explaining Bayesian Networks, there remains a lack of metrics for assessing their accuracy and effectiveness.

Authors in [36] introduce four principles for XAI systems: providing evidence or reasons for outputs, explanations understandable to users, accuracy of explanations reflecting the system's processes, and operating only under designed conditions. These principles take into account the varied

needs of developers, decision-makers, and end users, influencing the necessary explanations based on specific contexts and requirements.

Finally, in 2024, the authors in [37] proposed an end-to-end framework to evaluate XAI methods for network intrusion detection. It evaluates global and local scopes and analyzes metrics such as descriptive accuracy, sparsity, stability, efficiency, robustness, and completeness. The framework uses three network intrusion datasets and seven AI methods and is released as a baseline for the network security community. This reveals the limitations and strengths of the current XAI methods.

Table 1 summarizes several metrics developed in the literature, with the objectives of each study and the type of application of data and models.

In our research, we aim to consider the influence of train and test data changes on the feature ranking provided by XAI. Hence, our goal was to combine the XAI approach with a data sampling method to validate the XAI more reliably. Then, we developed a global metric based on generalizability, concordance, and stability to evaluate the reliability of the XAI and k-fold combination.

III. METHODOLOGY

Our study aims to define and develop an approach to improve and assess the reliability of XAI. Our methodology begins by combining XAI approaches with the k-fold cross-validation method as shown in Fig.1 to enhance the reliability of explainability, particularly for small datasets. This integration considered multiple samples within a dataset. To evaluate the reliability of this combination we propose to study the generalizability, concordance, and stability of the explanation. Next, we define and develop a global metric for assessing the reliability of XAI that considers generalizability, concordance, and stability. Our approach will be employed across two explainability frameworks: SHAP (SHapley Additive exPlanations) [40] and LIME (Local Interpretable Model-agnostic Explanations) [41] as these two XAI approaches are the most used approaches found in the literature in the context of endocrine disease prediction. Moreover, these approaches are relatively generic, can be applied to several ML models, and are not limited by a specific model.

A. IMPROVING AND ASSESSING XAI RELIABILITY

The challenge with explainable ML is that whenever the selection of test and training data for prediction is changed, the explainability of the prediction outcomes leads to a novel order of feature contributions [14]. To overcome this problem and improve the reliability of the explainability, we propose to combine the extraction of importance coefficients by XAI approaches with the k-fold technique. Subsequently, metrics were developed to study the generalization, concordance, and stability of the combination of XAI and k-fold.

1) IMPROVING RELIABILITY: COMBINING XAI WITH K-FOLD

For this purpose, data were divided into k samples or files using the k-fold technique. For the first iteration, one of the k samples was chosen as the validation set, with the remaining k-1 samples serving as the training set for model learning. Then, for each iteration, the data file selected as validation data is used for training, and one of the files selected previously for training is used for validation. For each iteration, we concatenate the feature importance coefficients in the XAI list. Eventually, we obtain a list of feature importance coefficients divided into multiple sub lists, each of which results from a prediction made using distinct training and test sets of data, as shown in Fig.2. As a result, visualizing the explainability of each variable's significance to prediction is more generalizable and reliable for validating the feature importance ranking.

2) GENERALIZATION

The objective of this part is to identify the value of k in k-fold, where the ranking of the best feature contributors remains unchanged. The idea is to calculate the augmented XAI coefficients for several values of k. Subsequently, we computed the average absolute of the coefficients for each feature. Thus far, we have a separate list for each value of k, that contains the average importance coefficient for each feature. Each list provides a distinct rating of features for each k. Subsequently, Spearman correlation similarity analysis was conducted to examine the generalizability or variability of the feature ranks across various k values. Ultimately, a matrix was generated to display the correlation among all k values and construct a curve to assess the variation between each pair of consecutive k values.

We propose studying the generalizability metric to first choose the optimum k value in the k-fold approach and to study whether the feature importance ranking provided by XAI changes when the data selected for testing or training changes. In other words, we aim to check whether the final feature ranking is generalizable and stable within the dataset.

This metric is based on the calculation of the similarity (correlation). Therefore, it was between 0 and 1. A value of 1 indicates maximum generalizability. However, a value of zero indicates a zero similarity.

The algorithm 1 describes the computing process of similarity between features importance ranking.

3) CONCORDANCE

This metric evaluates the correlation between the feature importance provided by XAI, named XAI_coefficients and those provided by the predictive model explanation (impurity-based importance for the predictive model) [38]. A high correlation indicates that the explanations provided by the XAI agree with the intrinsic importance of the features according to the model, which is a reliable and significant indicator of explainability.

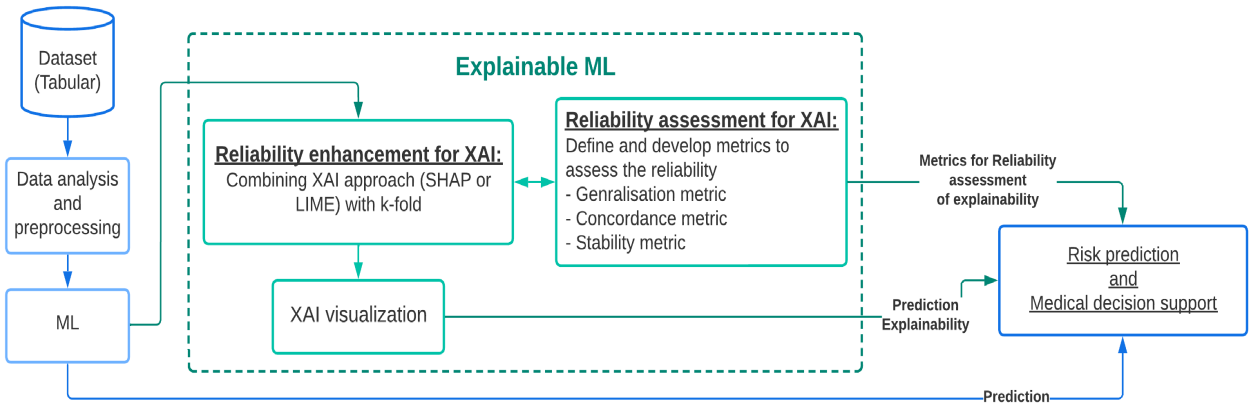

FIGURE 1. Process of the proposed methodology.

Algorithm 1 Generalizability Analysis of Feature Sorting Based on k Values

Step 1. Increase the XAI coefficient extraction for each k-value

Step 2. Compute the average absolute values of the XAI coefficients for each feature represented by the following formula.

$$M = \frac{1}{m}\sum_{i=0}^{m-1} |XAI\_coefficients_i| \tag{1}$$

where m is the number of subjects

Step 3. Generate a list of values, including the mean absolute values of the feature importance for each k-value:

$$L_j = [M_{0j}, M_{1j}, \ldots, M_{(n-1)j}] \tag{2}$$

for j in [1,...,n], where n is the max value of k

Step 4. The similarity between the ranked lists for all k values was calculated using the Spearman correlation presented by the following formula:

$$Generalisation = \rho(\mathbf{L}_j, \mathbf{L}_{j+1}) \tag{3}$$

where ρ refers to Spearman correlation.

Step 5. Display the Generalisation_metric between feature ranks for each k value.

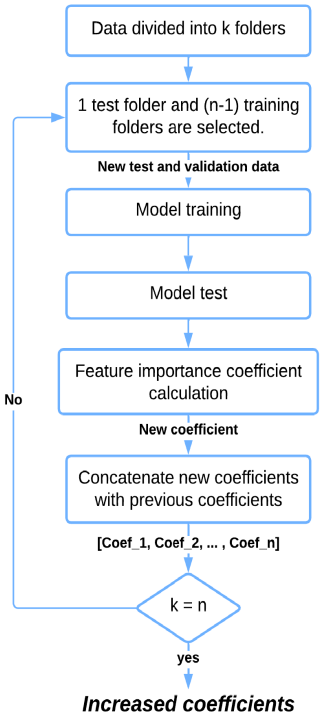

FIGURE 2. XAI with k-fold.

We propose to study concordance to assess the degree to which the final ranking of the most important features provided by XAI is directly correlated with the importance of these features in the kernel of the model in the ML process.

The concordance equation is presented below (4).

$$\text{Concordance} = \Phi(\mathbf{I}_{\text{Model}}, \mathbf{I}_{\text{XAI}}) \tag{4}$$

where Φ represents the Pearson correlation function, $\mathbf{I}_{\text{model}}$ is the vector of feature importance as assessed by the model, and $\mathbf{I}_{\text{XAI}}$ is the vector of mean feature importance derived from XAI.

The concordance was also between 0 and 1. A higher concordance indicates good reliability.

4) STABILITY

This measure assesses the extent to which the explanations provided by XAI are consistent for similar instances [39].

TABLE 1. XAI evaluation metrics summary.

Reference	Metric	Metric Definition	Study Objective	Metric type	Data	Model
[17]	Performance (D), Number of rules (R), Number of features (F), Stability (S)	Difference in performance between the agent's model and the explanation's logic.	Demonstrate the effectiveness of XAI in enhancing diagnostic accuracy in radiology.	Objective	Images	Neural Networks
[18]	transparency, persuasiveness, scrutability, effectiveness, satisfaction, efficiency, trust	Metrics evaluate how well a system communicates its decision-making process, convinces users, and allows for inspection.	Assess the role of clarity and user control in recommendation systems acceptance.	Subjective	-	-
[19]	Login Frequency, Sales	Indicators of how often users log in and the volume of sales, reflecting engagement and trust.	Investigate how system design influences user trust and engagement metrics.	Obj./Subj.	-	-
[20]	Actionability, Correctability, Causality, faithfulness, justifiability, robustness, soundness.	Focus on the user's ability to act upon, correct explanations, and understand cause-effect relationships.	Evaluate how well XAI systems enable user interaction and understanding through explanations.	Obj./Subj.	-	-
[25]	Explanation Quality, User Satisfaction, performance/productivity, usability/interaction.	Cognitive metrics that gauge the clarity, helpfulness, and satisfaction levels of explanations from the user's perspective.	Identify and measure cognitive metrics reflecting user interaction with XAI systems.	Subjective	-	-
[27]	Explanation Stability, Robustness	Evaluate how consistent and resilient the explanations are to changes and noise.	Explore the consistency and resilience of explanations in time series classification.	Objective	Time Series	-
[28]	Model Size, Complexity, Monotonicity, Complexity.	Assessments of the explanatory model's size, intricacy, and the predictability of feature importance.	Assess the interpretability and accuracy of explanations across different models and data types.	Objective	Images, Tabular	-
[30]	Goodness Checklist, Satisfaction Scale	Tools to evaluate the effectiveness, adequacy, and satisfaction with the explanations provided.	Develop and refine metrics for evaluating explanation effectiveness and user satisfaction.	Subjective	-	-
[35]	Explanation Focus, Explanation Level.	Criteria for determining the scope, depth, and approachability of explanations in Bayesian networks.	Review and define the necessary properties of explanations in Bayesian networks for user understanding.	Obj./Subj.	-	Bayesian Nets
[37]	Accuracy, sparsity, stability, efficiency, robustness, and completeness	End-to-end framework.	evaluate both global and local scopes of XAI for network intrusion detection.	Objective.	-	-

A small distance close to zero means that similar instances receive similar explanations, indicating good stability in how the model assigns importance to features.

This metric is dedicated to assessing the stability of the explainability provided by XAI for two similar instances. In other words, it is a useful metric for evaluating the certainty of its explainability. The stability metric is an important assessment in conjunction with the generalizability metric because unstable and uncertain explainability can affect the similarity between feature rankings as a function of k values, thus causing disorder in the generalizability metric.

The equation for stability is presented below (5).

$$\text{Stability} = \frac{1}{N}\sum_{k=1}^{N} d(\mathbf{S}_{k1}, \mathbf{S}_{k2}) \tag{5}$$

where N is the number of pairs of similar instances examined, d represents euclidean distance function, and $\mathbf{S}_{k1}$ and $\mathbf{S}_{k2}$ are the XAI value vectors for the k pair of similar instances.

The euclidean distance is calculated by the following formula (6):

$$d(\mathbf{S1}, \mathbf{S2}) = \sqrt{\sum_{i=1}^{n} (S1_i - S2_i)^2} \tag{6}$$

where **S1** and **S2** are two vectors of the XAI values for compared instances, and $S1_i$, $S2_i$ are the corresponding components in these vectors.

5) GLOBAL RELIABILITY

Our final metric considered generalization, stability, and concordance. A higher metric (equal to 1) for generalizability or concordance showed good reliability. In contrast, a lower stability metric close to zero indicates good reliability. Hence, the final reliability metric is the product of generalizability and concordance and 1 minus stability as shown in (7). Therefore, even reliability_metric will be between 1 and 0. A reliability close to 1 indicates the best XAI reliability.

$$\text{Reliability} = \textit{Concordance} \cdot \textit{Generalizability} \cdot (1 - \textit{Stability}) \tag{7}$$

This metric is very helpful for comparing and choosing between various explainability approaches, considering generalizability, concordance, and stability.

B. APPLICATION TO EXPLAINABILITY APPROACHES

Our approach was applied to two explainability frameworks: SHAP and LIME. Hence, the XAI coefficients presented here

are equivalent to the SHAP or LIME coefficients. We begin by defining these two approaches and calculating the feature importance coefficients.

1) SHAP

SHAP is an advanced method for the explainability of ML models, based on cooperative game theory, to evaluate the impact of each feature on model prediction. In particular, it uses Shapley values, a concept derived from game theory, to assign each feature a fair share of the contribution to the final prediction, regardless of the order in which the features are evaluated. This approach provides both local and global views of the feature importance [40].

The strength of the SHAP lies in its ability to decompose the predictions of a complex model into individual contributions attributable to each input feature. This decomposition was achieved by examining the marginal effect of adding a feature to the model prediction, considering all possible combinations of features. This provided an accurate and fair measure of the importance of each feature, reflecting its true impact on the prediction.

The contribution of feature "l" to the prediction is calculated by the difference between the model predictions when feature "l" is included $f(x^k_{+l})$ and when it is excluded $f(x^k_{-l})$, summed over all k examples in the dataset (K) as shown in the formula (8):

$$G(l) = \sum_{k=1}^{K} (f(x^k_{+l}) - f(x^k_{-l}). \tag{8}$$

where K is the number of iterations, l is the feature index, x is the data matrix, and f(x) is the prediction for x, but with a random number of feature values replaced by feature values from a random data point, except for the respective values of features with index l.

2) LIME

LIME is a local explainability technique designed to aid in the understanding of the predictions of complex ML models. Unlike SHAP, which focuses on distributing the contributions of each feature across all possible combinations of features, LIME simplifies this process by creating an interpretable model (such as a linear regression or decision tree) that approximates the predictions of the complex model in the local neighborhood of the instance to be explained [41].

The LIME approach is based on the idea that although the overall model may be complex and non-linear, it is possible to approximate its behavior around a specific prediction with a simple, linear model that is much easier to interpret. To achieve this, LIME perturbs the instance of interest by generating a set of new instances in its neighborhood (by slightly modifying the feature values) and observing the model's predictions for these new instances. It then weighs these perturbed instances according to their proximity to the original instance and uses this weighted data to train a simple model. The coefficients of the simple model are interpreted as the importance of the features, for instance predictions.

The general formula for the simplified model is based on a local linear approximation of the model's predictive function f(.), explaining how each feature contributes to the prediction $\hat{f}(x)$, as shown in equation (9):

$$\hat{f}(x) = \alpha + \sum_{i=1}^{P} \beta_i x_i \tag{9}$$

where $\hat{f}(x)$ is the prediction of the simplified model for instance x, P is the number of features, β_i is the weight attributed to feature i indicating its importance in the local prediction, and x_i is the value of feature i for the instance.

C. CASE STUDIES

According to the World Health Organization, thyroid disorders and diabetes are the most prevalent endocrine diseases globally [15]. These conditions present significant challenges for healthcare professionals. ML models have demonstrated promising capabilities in forecasting the risks associated with these diseases, thereby aiding physicians in making informed decisions regarding screening, diagnosis, and treatment strategies. However, these models have often been criticized for their opaque nature, leading to their limited adoption in clinical practice. As previously mentioned, even when XAI methods are employed to enhance transparency, the reliability of these explanations remains a matter of concern.

The idea is to apply our approach to improve and assess the reliability of the XAI, specifically SHAP and LIME, in two real-world scenarios: thyroid disorders and diabetes. Our objective was to provide a more reliable framework for interpreting thyroid and diabetes risk predictions. By comparing the efficacy of SHAP and LIME across these case studies, we aimed to determine which approach offers a more trustworthy explanation of the predictions, enhancing the decision-making process in clinical settings.

IV. RESULTS AND DISCUSSION

As previously highlighted, the idea is to evaluate and test the proposed approach for improving and assessing the reliability of XAI on two XAI frameworks (SHAP and LIME) as well as two case studies. The first case study focused on predicting the risk of hypothyroidism, whereas the second case study focused on diabetes risk prediction. For both case studies, we used the same ML predictive model to ensure a fair comparison of the application to two different datasets. This model should already have internal explainability to calculate the concordance between XAI and the explainability of the model architecture. To achieve this, logistic regression, random forest, and support vector machine were tested in these two case studies. The random forest model performed the best. Hence, random forest was chosen for risk prediction in both cases.

In this section, Python and its libraries were used to develop and apply the proposed approach. Sklearn for model

development [42], Numpy for list manipulation [43], Pandas for dataset management and manipulation [44], Seaborn for visualizations [45], and SHAP and LIME for XAI [46].

A. APPLICATION OF OUR APPROACH TO PREDICT THE RISK OF HYPOTHYROIDISM

1) DATASET DESCRIPTION

Data were obtained from the UCI Machine Learning Repository [47]. The dataset contains 3772 subjects and 29 features including a binary output column reporting the presence and absence of hypothyroidism. The age range of the population in the dataset was 1–95 years, of which 67.9% were female and 32.1% were male.

Table 2 lists the existing features in the dataset associated with their corresponding definitions and types.

TABLE 2. Data description and type for hypothyroid dataset.

Attribute	Signification	Type
age	Patient's age	Int
sex	Gender of the patient	Int
On thyroxine	Whether patient currently taking thyroxine	Bool
Query on thyroxine	Patient is questioned about use of thyroxine	Bool
On antithyroid meds	Patient currently taking antithyroid medication	Bool
pregnant	If the patient now is pregnant	Bool
Thyroid surgery	whether the individual has had thyroid surgery	Bool
I131 treatment	If the patient has receiving I131 treatment	Bool
Query hypothyroid	patient thinks they have developed hypothyroidism	Bool
lithium	If patient currently taking lithium	Bool
goitre	Patient have goitre	Bool
tumor	If patient diagnosed with a tumor	Bool
hypopituitary	Patient hypopituitarism	Float
TSH	Blood test results for TSH level	Float
T3	Blood test results for T3 level	Float
TT4	Blood test results for TT4 level	Float
T4U	Blood test results for T4U level	Float
FTI	Blood test results for FTI level	Float
TBG	Blood test results for TBG level	Float
Target-hypothyroid	Diagnosis of hypothyroidism	Int

The positive class of the population significantly dominated the negative class, with 3387 cases identified as positive for hypothyroidism and only 291 identified as negative, as depicted in Fig.3.

Therefore, there is imbalanced data between the two classes of the target column. Imbalanced data results in an inappropriate ML of the model. This impacts the capacity to predict positive or negative values. Additionally, during the testing and validation steps of the prediction, it is likely that the data chosen for this step may be inconsistent. Generally, the test phase is considered unreliable.

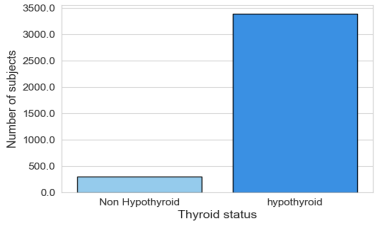

FIGURE 3. Imbalanced hypothyroidism output classes.

There are two solutions to this unbalanced data problem. The first technique employed oversampling. This concept involves randomly duplicating the minority class to achieve a balanced distribution between the two classes. Undersampling was the second proposed solution. This includes the elimination of instances from the dominant class to achieve a balanced distribution between the two classes.

Most studies in the literature have used oversampling. However, this methodology induces overfitting and frequently results in unjust testing and validation. This study employed an undersampling technique. This strategy reduces dataset volume. However, this strategy ensures reliable and equitable testing and validation. Fig.4 shows the data obtained after a random undersampling. Data were transformed from 3772 to 582 subjects. However, we guaranteed better learning for the ML model and a reliable testing and validation process.

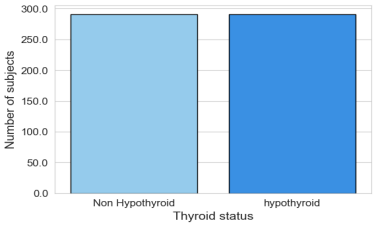

FIGURE 4. Balanced hypothyroidism output classes.

2) RISK PREDICTION OF HYPOTHYROIDISM USING RANDOM FOREST

Several tests were carried out to select the best hyperparameters of the random forest model. The ideal hyperparameters were a 300 of tree numbers in the forest, the function to measure the quality of a split is "GINI" and a max depth of 10.

To provide a reliable risk prediction evaluation, we applied a k-fold cross-validation (k = 10) to show the metric raters [48].

Random forest showed performances of 99.1%, 99.5%, 98.8%, and 99.1% for the accuracy, precision, recall, and f1-score.

3) SHAP RELIABILITY IMPROVEMENT AND ASSESSMENT

To improve the reliability of explainable ML, in this section, we analyze the variability in the sorting feature importance for every k value. The correlation matrix presented in Fig.5 illustrates the degree of similarity in the feature ranks across the various k values. When the value of k is less than 5, there is a limited correlation between k-values. The rankings of the features showed major changes. This implies a lack

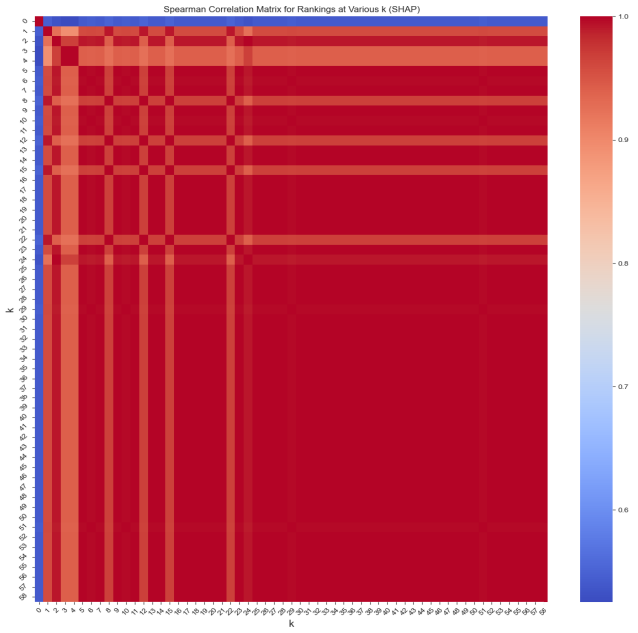

FIGURE 5. Correlation matrix between feature rankings: SHAP for hypothyroid.

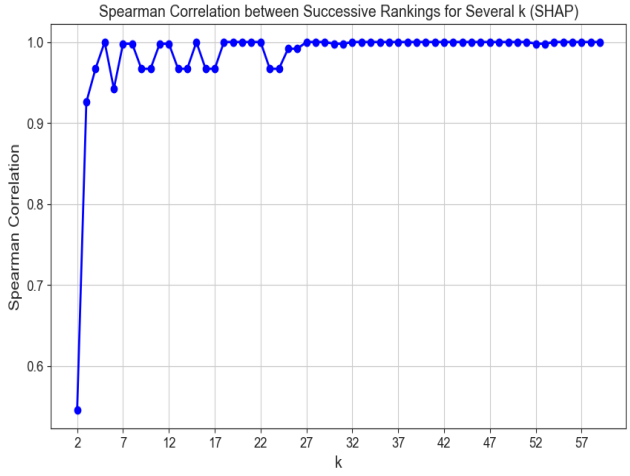

FIGURE 6. Correlation between characteristic rankings of successive k values: SHAP for hypothyroid.

of generalization by only applying SHAP without K-fold or even with small k-values. Starting from k = 27, we observed that the ranks have nearly identical similarity, which is equal to one.

To effectively visualize the similarity, Fig.6 shows the correlation between consecutive k values to assess the consistency of the rankings. Fig.6 demonstrates that starting from k = 27, there is a strong correlation between rankings, with an approximate value of 1. Therefore, the ranking remained mostly unchanged.

Based on this analysis, we can confidently state that utilizing a combination of Shapley and k-fold with a value of k = 27 is the most reliable method for studying the most

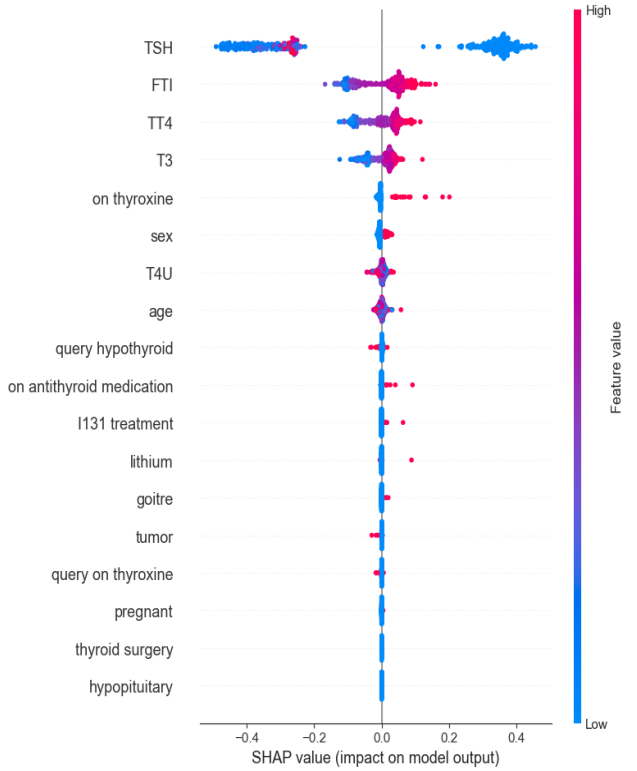

FIGURE 7. SHAP with k-fold after study k-value (k = 27): SHAP for hypothyroid.

important features in the dataset, with a generalization metric equal to 1. The ultimate ranking of feature importance is displayed in Fig.7 in descending order from top to bottom.

We begin by explaining how to analyze the graph. Features are shown on the y-axis, and Shapley values on the x-axis. The color blue and red indicates whether the characteristic values for each subject are at minimum (blue) or maximum (red). Each point on the graph represents the Shapley value for each specific characteristic associated with a patient. Therefore, the number of points for each entry was equal to the number of patients. The characteristics were ranked from the most important (top) to the least important (bottom) in predicting hypothyroidism. If the biggest values (in red) of a variable admit positive Shapley values, it means that the bigger the variable, the greater the risk of having the anomaly (Output = 1). Alternatively, if the smallest values of a variable admit positive shapley values, the smaller that variable is, the greater the risk of having the anomaly. As shown in Fig.7, TSH was the most important feature for the risk prediction of hypothyroidism. We can see the difference between the TSH SHAP values and the SHAP values of other features. The final ranking of the top contributors is highly reasonable from a medical perspective [49]. This means that our forecast and risk assessment outcomes are dependable and are not influenced by random chance or skewed data.

We computed the developed metrics to assess the reliability of SHAP after combining it with the k-fold. The preceding

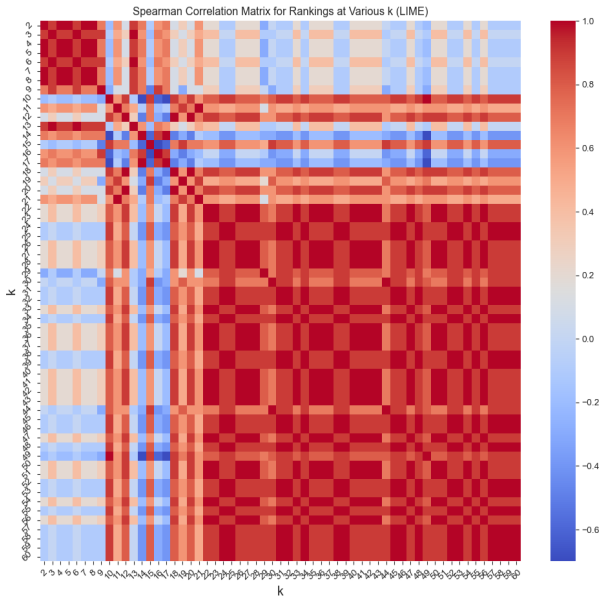

FIGURE 8. Correlation matrix between feature rankings: LIME for hypothyroid.

curve shows that generalizability is perfect and equal to one. The stability of the similarity between the several feature classifications was perfectly correlated. In addition, SHAP with k-fold has a very good concordance of 0.994 and a good stability of 0.087. Therefore, the overall reliability metric is 0.91, which is good reliability.

4) LIME RELIABILITY IMPROVEMENT AND ASSESSMENT

In this section, we display the same reliability assessment methodology with the same graphs presented in the previous section with SHAP and k-fold to study the combination of LIME and k-fold.

Fig. 8 shows that by combining LIME and k-fold for k values below 17, the correlation is very weak. This demonstrates the lack of generalizability of LIME without the application of k-fold, even with low k values. However, for k values above 17, the correlation is much higher and very close to 1.

By analyzing the variation of similarities in Fig. 9, we can see that, unlike SHAP, the generalization metric is well perturbed and not stable. This means that the feature ranking was not similar for several k-values. However, we can see a trend towards stability with a correlation between 0.75 and 1 from k = 23. Hence, the generalization metric has an average value of 0.875. The concordance was relatively good at 0.81, and the stability was very good at 0.017. This means the LIME and k-fold approach were stable and certain of its explainability. On the other hand, it is non-generalizable and not sufficiently correlated with the internal interpretability of random forest. This provided an overall reliability metric of 0.69, affirming a lack of LIME and k-fold reliability for the hypothyroid case study.

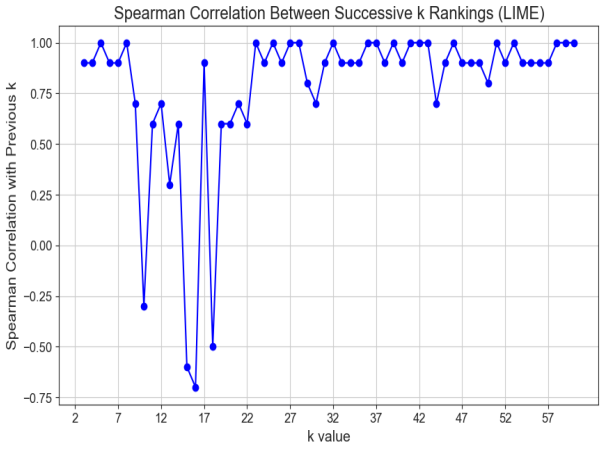

FIGURE 9. Correlation between characteristic rankings of successive k values: LIME for hypothyroid.

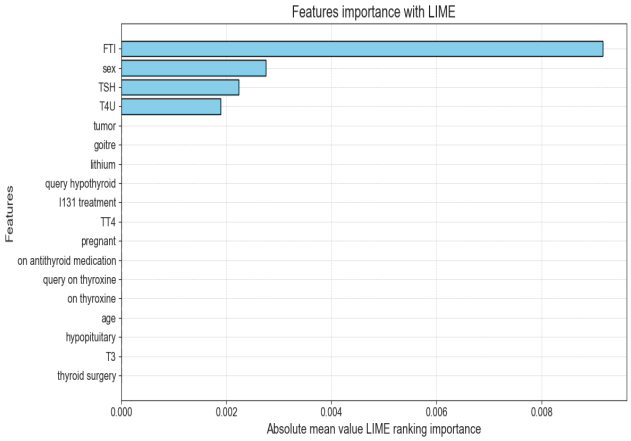

FIGURE 10. LIME with k-fold after study k-value (k = 27): LIME for hypothyroid.

Fig. 10 shows the ranking of the feature importance after combining LIME and k-fold. The TSH variable was ranked as the most important variable, as shown by LIME and SHAP. However, there is a change in the importance of the other variables, which leads us to compare the reliability of the two explainability approaches.

B. APPLICATION OF OUR APPROACH TO DIABETES RISK PREDICTION

For the second case study, we applied the same approaches as in the first case study, with identical graphical presentations, to analyze the differences between the two case studies. The same predictive model was used in this case study to predict diabetes.

1) DATASET DESCRIPTION

The Pima diabetes dataset was obtained from UCI Machine Learning. It originated from the National Institute of Diabetes and Digestive and Kidney Diseases. The goal of the

TABLE 3. Data description and type for the diabetes dataset.

Features	Definition	Type
Pregnancies	Number of times pregnant	int
Glucose	Plasma glucose concentration	int
BloodPressure	Diastolic blood pressure (mm Hg)	int
SkinThickness	Triceps skin fold thickness (mm)	int
Insulin	2-Hour serum insulin (mu U/ml)	int
BMI	Body mass index (weight in kg/(height in m)^2)	float
DiabetesPedigreeFunction	Diabetes pedigree function	float
age	age (years)	int
Outcome	0: No diabete, 1: diabete	binary

dataset was to use the diagnostic features presented in Table 3 to predict whether a patient has diabetes from a diagnostic perspective. The dataset contained 9 features and 768 subjects, 268 of whom are diagnosed as positive for diabetes.

2) RISK PREDICTION OF DIABETES USING RANDOM FOREST

To provide a reliable risk prediction evaluation, we applied a k-fold cross-validation (k = 10) to show the metric raters.

Random forest showed a moderate performance in predicting diabetes risk. Accuracy 76.8%, precision 69.0%, recall 59.5% and F1-score 62.8%.

3) SHAP RELIABILITY IMPROVEMENT AND ASSESSMENT

The correlation matrix presented in Fig.11 shows that for k less than 8, the correlation is lower and perturbed than that for k greater than 9, which is a strong correlation. Hence, the ranking of features after the combination of SHAP and k-fold was stable quickly and perfectly from k = 10.

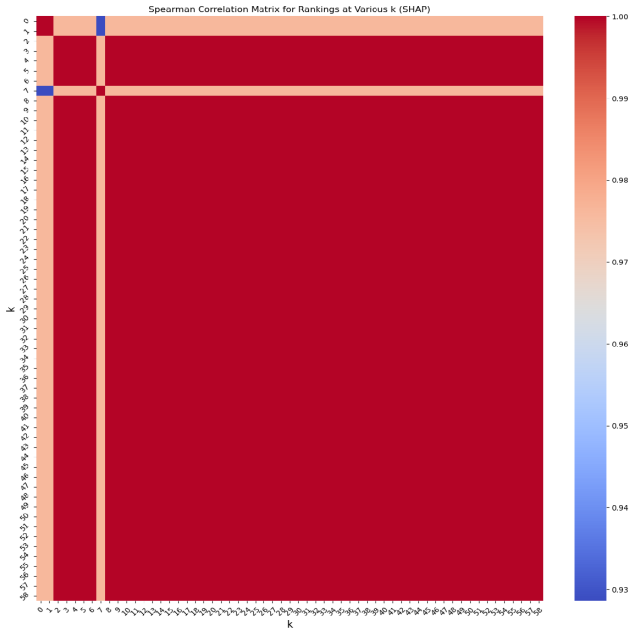

FIGURE 11. Correlation matrix between feature rankings: SHAP for diabetes.

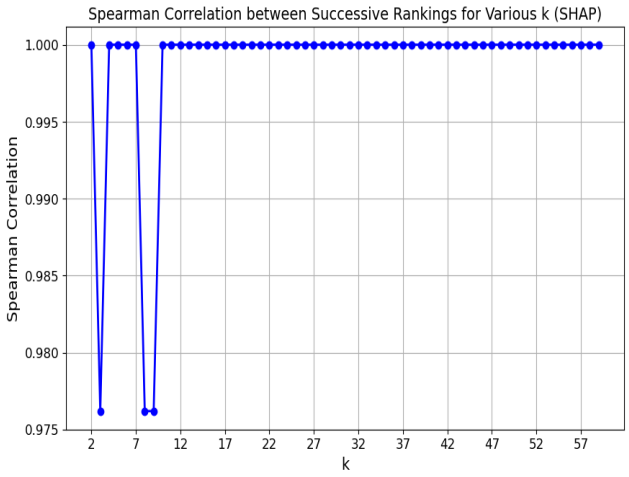

FIGURE 12. Correlation between characteristic rankings of successive k values: SHAP for diabetes.

Moreover, the curve in Fig. 12 shows that from k = 10, the correlation is perfectly equal to 1. Even for k values below 9, the correlation was strong between 0.975 and 1.

After combining SHAP and k-fold, Fig. 13 shows that glucose was the most important variable for predicting diabetes risk. In addition, all other variables were important for prediction, particularly BMI and age.

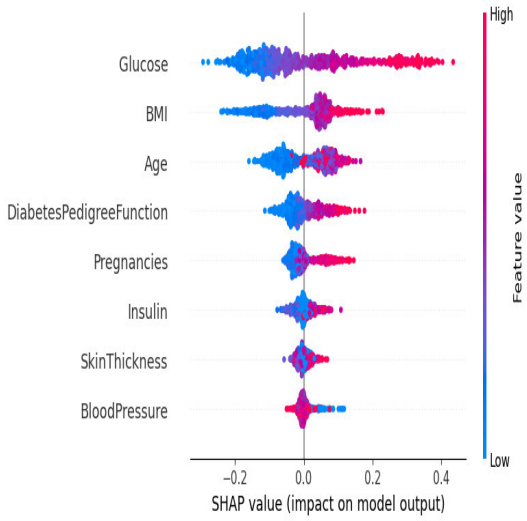

FIGURE 13. SHAP with k-fold after study k-value (k = 10): SHAP for diabetes.

Finally, the generalizability of SHAP and k-fold was perfectly equal to 1, proving an identical feature importance ranking. The concordance of this combination and the random forest explanation was also significant at 0.98, with a good stability of 0.01. Hence, a very good global reliability score of 0.97 for the combination of SHAP and k-fold for the diabetes case study.

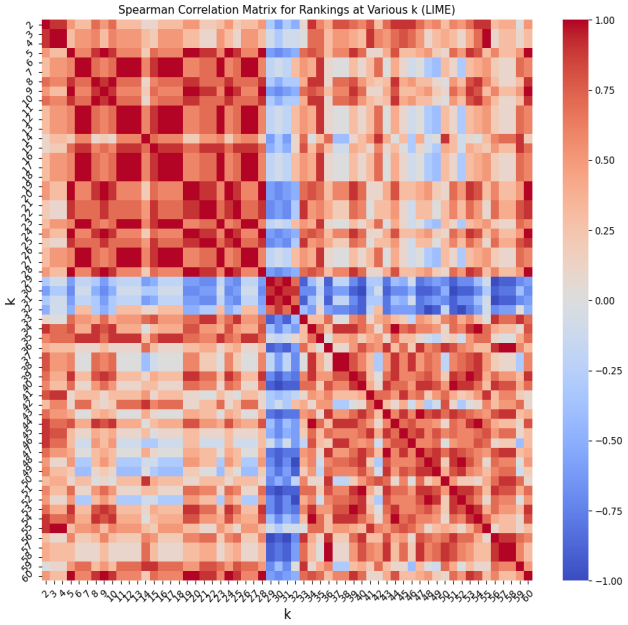

FIGURE 14. Correlation matrix between feature rankings: LIME for diabetes.

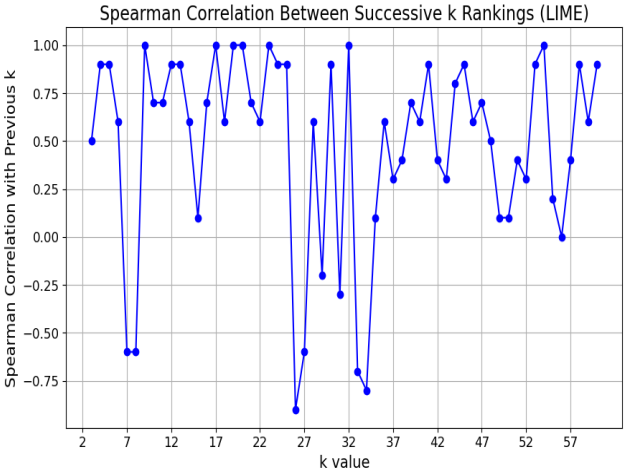

FIGURE 15. Correlation between characteristic rankings of successive k values: LIME for diabetes.

4) LIME RELIABILITY IMPROVEMENT AND ASSESSMENT

The correlation matrix presented in Fig.14 shows a weak and skewed correlation for the combination of LIME and k-fold, demonstrating the lack of generalizability of LIME in this case study.

The same is true for the correlation curve shown in Fig.15, which confirms that the LIME and k-fold combination are not stabilized. Hence, there is a lack of generalizability of the LIME and k-fold combination. Therefore, the generalizability metric is the mean value of all correlations. This mean value was equal to 0.7.

Fig.16 shows the feature rankings for LIME, which differ from those for SHAP. This confirms the need to study the reliability of both approaches.

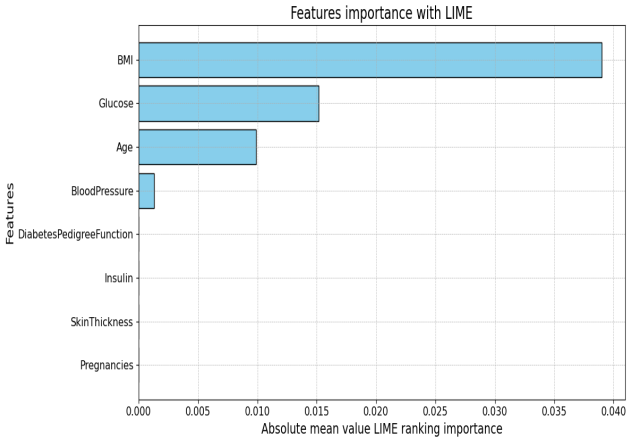

FIGURE 16. LIME with k-fold after study k-value (k = 32): LIME for diabetes.

The concordance between LIME plus k-fold and Random Forest explainability was very low, equal to 0.27 with a stability of 0.01. This means a high degree of surety in the explainability provided by LIME and k-fold, but not sufficiently correlated with the internal interpretability of random forest. In addition, LIME and k-fold showed poor and unstable generalizability, which shows the lack of similarity between the different rankings of the most important features, leading to poor reliability. Hence, the global reliability score is 0.187.

C. DISCUSSION

In this study, we propose to combine the k-fold technique with the SHAP and LIME approaches. Subsequently, we developed metrics to evaluate the concordance, generalization, stability, and overall reliability of this combination. We then tested and applied these to two different datasets to predict hypothyroidism and diabetes. Table 4 summarizes the results obtained for both case studies. First, we noticed a difference in the ranking of the most important features between SHAP, kfold and LIME, kfold. This adds to the obligation to evaluate and compare the reliability of these two combinations. this can be observed from Table 4, that the SHAP approach with k-fold was more generalizable and less influenced by changes in the test and training data for both case studies. This combination achieved feature ranking stability more quickly in the second case study, which may be explained by either the lower number of features or the higher amount of data in the second case study compared with the first. However, for the combination of LIME and k-fold, feature ranking did not achieve stability for either case study, indicating a lack of generalizability and a significant influence on changing the selected test and training data. Additionally, the SHAP and k-fold combinations showed strong concordance with the internal explainability of Random Forest at 0.994 and 0.98 respectively in both case studies, ensuring the convergence of SHAP explainability

TABLE 4. Summary of XAI reliability assessment.

Metrics	Thyroid prediction		Diabetes prediction	
	SHAP	LIME	SHAP	LIME
Generalization	1	0.875	1	0.7
Concordance	0.994	0.81	0.98	0.27
Stability	0.087	0.017	0.01	0.01
Overall reliability	0.91	0.69	0.97	0.18

even when combined with k-fold. In contrast, the LIME plus k-fold approach showed poor concordance, especially for the second case study at 0.27, indicating that combining k-fold with LIME was not effective or reliable. Both approaches provide stable explanations for the two case studies. The good stability of the LIME with k-fold indicates the relative suitability of this approach. The good stability of LIME with k-fold, poor concordance, and poor generalizability raise doubts about the effectiveness of the LIME with k-fold. In contrast, the SHAP and k-fold combination demonstrated excellent stability, concordance, and generalizability. Finally, the overall reliability score effectively demonstrates that the explanations provided by the SHAP and k-fold combination were reliable at 0.91 and 0.97 for both hypothyroidism and diabetes predictions. Moreover, without this combination, the explainability of SHAP remains influenced by the change in training and test data presented when the k variable is weak in the correlation curves. This supports the usefulness of our idea of combining SHAP and k-fold. In contrast, the LIME and k-fold combination showed mediocre scores, particularly for predicting the risk of diabetes, with a reliability score of 0.18.

Finally, in this study, we proposed not only to share an identification of at-risk individuals that is ambiguous for practitioners but also to provide a reliable explanation of this identification combined with k-fold and metrics to assess this reliability and its degree of certainty. A reliable XAI such as SHAP combined with k-fold and metrics that address its reliability, such as generalizability, concordance, and overall reliability, can increase physicians' confidence in the prediction and its explanatory power. This may lead to greater integration of ML models for risk prediction in hospitals, particularly in the context of endocrine diseases.

D. LIMITS AND PERSPECTIVES

In this work, we proposed an approach to improve and evaluate the reliability of XAI validation in response to a specific limitation related to explainability change provided by XAI when training or test data is changed. This specific and limited scope is embedded within several other XAI limitations in the literature.

Our proposed approach to enhance and evaluate the reliability of XAI is limited to basic ML models that exploit tabular data. It may not be applicable to deep learning models and to other types of data, such as images. Our study tested this methodology in only two case studies using the Random Forest model. Our findings indicated that the combination of SHAP and k-fold validation is reliable, based on the evaluation metrics developed in this study.

However, the combination of LIME and k-fold validation showed poor performance. This lack of reliability could be attributed to the fact that LIME is typically used as a local explainer. It may also be caused by the uncertainty of the AI model. Since it is a frequent problem, many models in the literature are likely over-optimistic due to leakage and over-fitting. Our focus in this article was more on the reliability of XAI, and we hope in future projects to study the impact of model reliability and optimization on XAI reliability. Another assumption is presented on the limited performance of LIME with k-fold about the data quantity. Based on our analyses, a high number of features may affect the stability of the XAI feature ranking, especially when the number of subjects is relatively small. However, a high number of subjects with a low number of features may increase the stability of feature ranking as a function of k values. This prompted us to investigate this combination further using several ML models and case studies in the future. In addition, in our research, we have proposed a combination of the k-fold technique with SHAP and LIME approaches, and we envisage testing other data sampling techniques in place of k-fold to discuss the reliability of the validation produced by this combination.

Combining the XAI approach with external validation may also lead to reliable XAI validation [50]. Therefore, it looks interesting to test this combination for reliable validation of the XAI.

Ultimately, the existence of XAI aims to ensure and strengthen confidence in predictions by identifying the basis on which these predictions are made. This explainability is crucial in the medical field, as it is a sensitive area where a prediction can recommend specific treatments or tests. However, a thorough study of the reliability of XAI is also essential to reinforce this confidence. Evaluating and improving the reliability of XAI is, therefore, a priority, motivating us to delve deeper into this field in our future research.

V. CONCLUSION

In this study, we propose a structured approach to improve and assess the reliability of explanations provided by ML models in the healthcare sector. We aimed to enhance the reliability of explainability by combining XAI approaches with the k-fold method. We developed metrics to evaluate the generalizability, concordance, and stability of the combined XAI and k-fold approach, and applied them to case studies of hypothyroidism and diabetes prediction using SHAP and LIME frameworks.

Our results indicated that the SHAP approach combined with k-fold demonstrated higher generalizability, stability, and concordance than the LIME approach. The SHAP and k-fold combination provided reliable explanations for both hypothyroidism and diabetes predictions, with strong concordance with the internal explainability of the Random

Forest model. In contrast, the LIME and k-fold combination showed poor concordance and reliability, particularly in predicting the risk of diabetes.

Overall, our study highlights the importance of combining XAI approaches with robust validation methods such as k-fold to improve the reliability of model explanations in healthcare applications, particularly with the combination of SHAP and k-fold. The structured approach and metrics developed in this study can enhance practitioners' confidence in ML models and facilitate their adoption of these technologies in real-world scenarios. Further research and validation on diverse datasets are needed to validate the effectiveness and applicability of LIME with k-fold in other case studies.

REFERENCES

[1] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100370.

[2] A. S. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, and M. Deveci, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, Aug. 2023.

[3] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using Shapley additive explanation and application for real data in hospital," *Comput. Methods Programs Biomed.*, vol. 214, Feb. 2022, Art. no. 106584.

[4] P. P. Angelov, "Explainable artificial intelligence: An analytical review," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 11, no. 5, 2021, Art. no. e142.

[5] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023.

[6] H. Javed, H. A. Muqeet, T. Javed, A. U. Rehman, and R. Sadiq, "Ethical frameworks for machine learning in sensitive healthcare applications," *IEEE Access*, vol. 12, pp. 16233–16254, 2024.

[7] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[8] C.-K. Yeh, "On the (In)fidelity and sensitivity of explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[9] I. E. Kumar, "Problems with Shapley-value-based explanations as feature importance measures," in *Proc. Int. Conf. Mach. Learn.*, 2020.

[10] C. Schwartzenberg, T. van Engers, and Y. Li, "The fidelity of global surrogates in interpretable machine learning," in *Proc. BNAIC/BeneLearn*, 2020, p. 269.

[11] M. Charles, "But are you sure? An uncertainty-aware perspective on explainable AI," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2023, pp. 7375–7391.

[12] A. Sutradhar, M. Al Rafi, P. Ghosh, F. M. J. M. Shamrat, M. Moniruzzaman, K. Ahmed, A. Azad, F. M. Bui, L. Chen, and M. A. Moni, "An intelligent thyroid diagnosis system utilizing multiple ensemble and explainable algorithms with medical supported attributes," *IEEE Trans. Artif. Intell.*, vol. 5, no. 6, pp. 2840–2855, Jun. 2024.

[13] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 103737–103757, 2021.

[14] F. Ketata, Z. A. Masry, N. Zerhouni, and S. Yacoub, "Explainable machine learning approach with augmentation for mortality prediction," in *Proc. IEEE Int. Conf. Adv. Syst. Emergent Technol. (IC_ASET)*, Apr. 2023, pp. 01–06.

[15] B. Biondi, G. J. Kahaly, and R. P. Robertson, "Thyroid dysfunction and diabetes mellitus: Two closely associated disorders," *Endocrine Rev.*, vol. 40, no. 3, pp. 789–824, Jun. 2019.

[16] L. Coroama and A. Groza, "Evaluation metrics for explainable artificial intelligence techniques: State of the art review and challenges," in *Proc. Nom de la Revue ou de la Conf.*, pp. 1–8.

[17] A. Rosenfeld, "Better metrics for evaluating explainable artificial intelligence," in *Proc. 20th Int. Conf. Auton. Agents Multiagent Syst.*, May 2021, pp. 45–50.

[18] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," presented at the IEEE 23rd Int. Conf. Data Eng. Workshop, Apr. 2007.

[19] B. J. Fogg, "Web credibility research: A method for online experiments and early study results," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Mar. 2001, pp. 295-296.

[20] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[21] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 4, Jul. 2019, Art. no. e1312.

[22] I. Askira-Gelman, "Knowledge discovery: Comprehensibility of the results," in *Proc. 31st Hawaii Int. Conf. Syst. Sci.*, Jan. 1998, pp. 247–255.

[23] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018, *arXiv:1806.08049*.

[24] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *Proc. IEEE Symp. Vis. Lang. Human Centric Comput.*, Sep. 2013, pp. 3–10.

[25] J. H.-W. Hsiao, H. H. T. Ngai, L. Qiu, Y. Yang, and C. C. Cao, "Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI)," 2021, *arXiv:2108.01737*.

[26] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," 2018, *arXiv:1811.11839*.

[27] T. T. Nguyen, T. Le Nguyen, and G. Ifrim, "A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification," in *Proc. Int. Workshop Adv. Anal. Learn. Temporal Data*, vol. 6, Ghent, Belgium. Cham, Switzerland: Springer, Sep. 2020, pp. 77–94.

[28] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, Mar. 2021.

[29] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," 2020, *arXiv:2007.07584*.

[30] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[31] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv:1812.04608*.

[32] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.

[33] X. Zhao, "Baylime: Bayesian local interpretable model-agnostic explanations," in *Proc. Uncertainty Artif. Intell.*, 2021.

[34] C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij, "A method for explaining Bayesian networks for legal evidence with scenarios," *Artif. Intell. Law*, vol. 24, no. 3, pp. 285–324, Sep. 2016.

[35] C. Lacave and F. J. Díez, "A review of explanation methods for Bayesian networks," *Knowl. Eng. Rev.*, vol. 17, no. 2, pp. 107–127, Jun. 2002.

[36] P. J. Phillips, "Four principles of explainable artificial intelligence," U.S. Dept. Commerce, Nat. Inst. Standards Technol., 2021.

[37] O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," *IEEE Access*, vol. 12, pp. 23954–23988, 2024.

[38] S. Stassin, A. Englebert, G. Nanfack, J. Albert, N. Versbraegen, G. Peiffer, M. Doh, N. Riche, B. Frenay, and C. De Vleeschouwer, "An experimental investigation into the evaluation of explainability methods," 2023, *arXiv:2305.16361*.

[39] C. Munoz, K. da Costa, B. Modenesi, and A. Koshiyama, "Evaluating explainability for machine learning predictions using model-agnostic metrics," 2023, *arXiv:2302.12094*.

[40] C. Molnar, "Interpreting machine learning models with SAP: A guide with Python examples and theory on Shapley values," in *Chistoph Molnar C/O MUCBOOK, Heidi Seibold*, 2023.

[41] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo, "Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models," *J. Oper. Res. Soc.*, vol. 73, no. 1, pp. 91–101, Jan. 2022.

[42] P. Fabian, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.
[43] C. R. Harris, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.
[44] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.
[45] W. McKinney, "Pandas: A foundational python library for data analysis and statistics," *Python High Perform. Sci. Comput.*, vol. 14, no. 9, pp. 1–9, 2011.
[46] Datacamp. (May 2023). *Explainable AI Understanding and Trusting Machine Learning Models*. [Online]. Available: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models
[47] D. Dua and C. Graff, "UCI machine learning repository," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 1, p. 129, Sep. 2019.
[48] S. Borra and A. Di Ciaccio, "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods," *Comput. Statist. Data Anal.*, vol. 54, no. 12, pp. 2976–2989, Dec. 2010.
[49] B. O. Svold, "Serum TSH within the reference range as a predictor of future hypothyroidism and hyperthyroidism: 11-year follow-up of the HUNT Study in Norway," *J. Clin. Endocrinology Metabolism*, vol. 97, no. 1, pp. 93–99, 2012.
[50] R. D. Riley, L. Archer, K. I. E. Snell, J. Ensor, P. Dhiman, G. P. Martin, L. J. Bonnett, and G. S. Collins, "Evaluation of clinical prediction models (part 2): How to undertake an external validation study," *Brit. Med. J.*, vol. 384, Jan. 2024, Art. no. e074820.

FIRAS KETATA is currently pursuing the Ph.D. degree in data science and artificial intelligence with the University of Bourgogne Franche-Comté (UBFC), France. He was a Trainer in artificial intelligence, sharing his knowledge and expertise with other professionals. He holds a position as a Temporary Teaching and a Research Associate with École Nationale Supérieure de Mécanique et des Microtechniques (SUPMICROTECH-ENSMM), Besançon, France. He is affiliated with several renowned institutions, including SUPMICROTECH-ENSMM, UBFC, and the FEMTO-ST Institute. He is based in Besançon. His research focuses on developing data science and artificial intelligence based approaches for medical decision support systems.

ZEINA AL MASRY received the Ph.D. degree in applied mathematics from the University of Pau and Pays de l'Adour, France, in 2016. She joined École Nationale Supérieure de Mécanique et des Microtechniques (SUPMICROTECH-ENSMM), Besançon, France, as an Associate Professor, in 2017. She is doing her research activities with FEMTO-ST Institute, Prognostics and Health Management (PHM) Research Group. Her research works concern data quality management, applied statistics, and stochastic processes in healthcare and industry. She is interested in interdisciplinary research and mainly in diagnosis and prognosis.

SLIM YACOUB is currently a Professor with INSAT, Tunisia, and a Researcher with the LTSIRS Laboratory. His research interests include signal processing for biomedical signals, particularly ECG, EMG, EEG, and AMG signals.

NOUREDDINE ZERHOUNI received the Ph.D. degree in automatic-productivity from the National Polytechnic Institute of Grenoble (INPG), France, in 1991. He was a Lecturer with the National School of Engineers (ENI, UTBM), Belfort. Since 1999, he has been a Professor of universities with École Nationale Supérieure de Mécanique et des Microtechniques (SUPMICROTECH-ENSMM), Besançon. He is doing his research in the Automatic Department, FEMTO-ST Institute, Besançon. He is also an Expert in adult education in the areas of process improvement and project management. His areas of research are related to *Prognostics and Health Management*.