IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 June 2024, accepted 25 June 2024, date of publication 28 June 2024, date of current version 8 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420415

RESEARCH ARTICLE

Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making

NADIA KHAN[1], MUHAMMAD NAUMAN[1], AHMAD S. ALMADHOR[2], NADEEM AKHTAR[1], ABDULLAH ALGHURIED[3], AND ADI ALHUDHAIF[4]

[1]Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Punjab 63100, Pakistan
[2]College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[3]Department of Industrial Engineering, Faculty of Engineering, University of Tabuk, Tabuk 47512, Saudi Arabia
[4]Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

Corresponding author: Muhammad Nauman (nauman@iub.edu.pk)

This work was supported by Prince Sattam Bin Abdulaziz University under Project PSAU/2023/R/1444.

ABSTRACT In recent years, Explainable Artificial Intelligence (XAI) has attracted considerable attention from the research community, primarily focusing on elucidating the opaque decision-making processes inherent in complex black-box machine learning systems such as deep neural networks. This spike in interest originates from the widespread adoption of black-box models, particularly in critical domains like healthcare and fraud detection, highlighting the pressing need to understand and validate their decision-making mechanisms rigorously. In addition, prominent XAI techniques, including LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive exPlanations), rely on heuristics and cannot guarantee the correctness of the explanations provided. This article systematically addresses this critical issue associated with machine learning and deep learning models, underscoring XAI's pivotal role in promoting model transparency to enhance decision-making quality. Furthermore, this study advocates integrating Formal Methods to provide correctness guarantees for black-box internal decision-making. The proposed methodology unfolds in three pivotal stages: firstly, training black-box models using neural networks to generate synthetic datasets; secondly, employing LIME and SHAP techniques to interpret the models and visualize their internal decision-making processes; and finally, training decision trees on the synthetic datasets to implement Formal Methods for ensuring the correctness of the black-box model's decision-making. To validate this proposed approach, experimentation was conducted on four widely recognized medical datasets, including the Wisconsin Breast Cancer and Thyroid Cancer (TC) datasets, which are available in the UCI Machine Learning Repository. Specifically, this research represents a significant contribution by pioneering a novel approach that seamlessly integrates XAI and Formal Methods, thereby furnishing correctness guarantees for internal decision-making processes within the healthcare domain.

INDEX TERMS Black-box machine learning, neural networks, interpretable machine learning, cancer prognosis, decision-making, formal methods, formal verification, colored petri nets.

I. INTRODUCTION

Explainable Artificial Intelligence (XAI) has attracted significant attention within the scientific community to enhance the comprehension of the internal decision-making processes in

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

black-box machine learning systems [1], [2], [3], [4], [5], [6], [7]. The lack of transparency of ML and deep learning-based models is a major issue in their implementation and is criticized due to their black-box nature [8], [9], [10]. XAI is the subdomain under AI to improve transparency by explaining the internal decision-making of such models [11]. In literature, few known efforts to explain the black-box models include SHAP [12] and LIME [13]. Numerous taxonomies have emerged in the research to categorize various methods for explaining AI models as model-specific and model-agnostic, either locally or globally, to explain models [14], [15]. Unfortunately, these methods lack direct access to the structural parameters or internal model weights. Most notably, model-agnostic XAI techniques lack assurances of thoroughness and can yield logically flawed explanations. The constraints embedded within these informal XAI methods present a significant hurdle to the reliability of explanations, especially in scenarios categorized as high-risk or crucial for safety [16]. Recent research in explainability focuses on revealing the primary features that have the most significant impact on a model's decision-making process [17]. As AI-based systems only make predictions without explaining their rationale, there is a need for mechanisms to explain and interpret their decisions. The lack of adequate tools for inspecting the functioning of the black-box model implementation and criticized due to its black-box nature, even with such tremendous results [9]. With their implementation in safety-critical domains, including healthcare, finance, and self-driving cars, it is also vital to provide correctness guarantees for the internal decision-making process. In such domains, trust from end-users majorly depends on explainability and reliability, making it crucial to address this limitation [18], [19], [20], [21].

According to recent studies, XAI methods rely on heuristics and cannot guarantee the correctness of the explanations. However, recent endeavors demonstrate the potential of Formal Methods to deliver verifiably correct explanations. While these formal approaches are theoretically sound, their scalability is hindered by the computational complexity of verification. Additionally, the explanations they provide may occasionally become excessively complex [22], [23]. In response to these challenges, this article introduces an approach based on SHAP, LIME, and Formal Methods. In addition, it also leverages formal correctness assurances of internal decision-making.

LIME and SHAP are highly effective for explainable AI because they are model-agnostic, meaning they can be applied to any machine learning model. They provide clear, human-interpretable explanations by highlighting the contribution of each feature to individual predictions (LIME) or consistently across the entire model (SHAP). Their ability to align with human intuition and visualize feature importance makes them invaluable tools for understanding and trusting AI decisions.

Formal Methods are tools and techniques for specifying, verifying, and mathematically validating complex systems. Accordingly, they are used for specifying, modeling, and formal verification in fields ranging from medicine to finance [24], [25]. Formal Methods have two main techniques: model checking and theorem proving. Model checking constructs the system's traceable model to explore the entire system state space [26]. Theorem proving utilizes complex mathematical equations to prove the system specifications are correct mathematically. The Formal Methods can assist in validating and explaining the internal decision-making of complex systems [27], [28], [29]. Coloured Petri nets (CP-Nets) [30] are formal tools based on Petri nets [31]. They are useful for modeling non-deterministic and stochastic processes. They enable a systematic and exhaustive exploration of the mathematical model to prove, refute, or analyze the correctness of complex systems [32]. CP-Nets come with a graphic interface that makes it easy to operate and understand the model [33].

This article introduces an approach to explain the black-box ML Model of internal decision-making with formal correctness guarantees. The major advantage of this work is the proposal of a method that explains diagnostic decisions made by black-box machine learning models and provides formal correctness guarantees. This capability is critical for ensuring the reliability and trustworthiness of AI-driven diagnostics in sensitive fields such as healthcare. By integrating explainability with formal verification, the proposed method addresses both the interpretability and accuracy requirements, thereby enhancing the robustness and acceptance of AI-based diagnostic tools. This research will assist medical practitioners in making decision-making during prognostic decisions. This work presents visual explanations that align with established prior beliefs and gain best practices in providing general explanations. The explainability of the model significantly improves its suitability for practical use in clinical decision-making.

This research is different from the existing literature on the explanation of black-box ML Models in that it proposes the use of Formal Methods to explain and provide correctness guarantees of internal decision-making. This article focuses on interpreting prognostic decisions made by black-box ML models to improve early cancer detection and help healthcare professionals. The correctness of these prognostic rules is verified using CP-Nets state space analysis. The specificity, sensitivity, and accuracy are the principal performance measures used to demonstrate the efficacy of the proposed approach. The research results will help physicians identify life-threatening diseases in the early stages and facilitate a healthier society.

In summary, we introduced an approach for improved understanding of decision-making through the integration of XAI and Formal Methods. The main contributions of this research work are:

1) A novel approach to ensure the correctness of internal decisions made by black-box ML algorithms using formal methods.
2) It involves using decision tree induction over a synthetic dataset generated by a black-box algorithm to facilitate the explanation of its internal decision-making processes.
3) It uses CP-Nets to validate the accuracy and reliability of black-box model predictions, ensuring they align with formally specified preferences.
4) The effectiveness of the proposed approach is demonstrated by the empirical results of four real-world cancer datasets.

The paper is organized as follows. Section II presents the background. Then, the literature review is discussed in Section III. Afterward, Section IV discusses the materials and methods, and finally, the conclusion is presented in Section V.

II. BACKGROUND

A. EXPLAINABLE AI FRAMEWORKS

The domain of XAI aims to study and develop AI systems that provide clear and intelligible explanations for predictions and decision-making [6]. This interdisciplinary domain seeks to enhance the human understandability of black-box machine learning models by employing techniques that generate explanations [34]. Some common methods for explaining machine learning models include SHAP [12],LIME [13], Grad-CAM [35] and Grad-CAM++ [36], [37].

LIME employs a post-hoc method applied after the model has been trained. It is model-agnostic, meaning it is not tied to any specific ML algorithm. LIME operates by using input features and model outputs without interacting with the internal weights and layers of the model. This broad applicability allows LIME to provide interpretable explanations for various types of machine learning models by approximating the behavior of the model locally around the prediction of interest [38]. Its results have interpretability and are also beneficiary for alerts generated by a classifier compared to model confidence scores [39].

SHAP is a popular XAI technique known for aligning well with practitioners' intuitions, especially with decision tree models. It enhances interpretability and aids alert processing more effectively than relying solely on model confidence scores by decomposing predictions into understandable contributions from individual features. This makes SHAP particularly valuable in critical fields like medical diagnosis, banking, and fraud detection, where it improves task efficiency and highlights key decision-making features through domain knowledge. By using SHAP, practitioners can determine the most important features in predictions and their impacts, aiding in informed decision-making without compromising predictive accuracy. This increased interpretability helps identify potential immunotherapies to improve patient survivability rates and provides transparent reasoning [9], [40].

In general, SHAP is slower but suitable for deep learning applications, particularly when using optimized variations like deep and gradient explainers. Conversely, results indicate that LIME demonstrates superior reproducibility and execution time compared to the SHAP gradient explainer [41]. The success of LIME has spurred numerous enhancements to its capabilities as a prominent image explainer. Notable examples include Anchor LIME [42], which employs anchors for high precision and coverage in generating interpretable explanations; KL-LIME [43], which focuses on local interpretability; NormLIME [44], which incorporates normalization techniques for improved performance; and the most recent BMB-LIME [45], which aims to provide more robust explanations. Each of these efforts contributes additional value to the existing LIME methodology, underscoring its versatility and effectiveness in the field of explainable AI [42].

The field of XAI focuses on developing AI systems that offer transparent explanations for predictions and decision-making, employing methods like SHAP, LIME, Grad-CAM, and Grad-CAM++. SHAP, a model-agnostic technique, improves interpretability by revealing key features influencing predictions and aiding decision-making in various domains like medical diagnosis and fraud detection. While SHAP enhances understanding and reasoning, LIME stands out for its reproducibility and execution time, with ongoing efforts to enhance its capabilities through methodologies like Anchor LIME.

B. ARTIFICIAL NEURAL NETWORKS

ANNs are a type of supervised ML algorithm inspired by the functioning of the human brain and formed by an interconnection of neurons. A neuron is a single processing unit of the human brain. Billions of neurons are interconnected in a complex way in the human brain. Hence, ANNs are formed by complex interconnections of the artificial neurons [46]. ANNs have high accuracy with the ability to provide fault tolerance for big datasets. Their self-organized and adaptive learning behaviors make them suitable for a large set of applications, including industrial process control, sensory data recognition, medical diagnosis, weather forecasting, image and text classifications, and pattern recognition [47]. ANNs leverage the rapid information processing, mapping capabilities, fault tolerance, generalization, and robustness that make ANNs smart and powerful modeling and forecasting tools [48], [49]. However, despite the aforementioned characteristics, the internal decision-making of these models is not easily understood. With their use in security-sensitive areas, the focus is on reliable validation to avoid disaster situations.

In general, ANNs consist of one input layer, several hidden layers, and one output layer [50]. The neurons of each layer are connected to the neurons of the next layer, but the neurons of the same layer are not interconnected. The input layer neurons receive the data and then pass it into subsequent layers until they reach the output layer.

C. COLOURED PETRI NETS

Coloured Petri nets (CP-Nets) are a type of Petri nets that are used in modeling systems that contain discrete, concurrent, and scattered events. CP-Nets are directed graphs containing two kinds of nodes: places represented by ellipses or circles and transitions represented by rectangular boxes [51]. In CP-Nets, the edges connect nodes of different types and are represented as arcs. CP-Nets have strong mathematical logic associated with Standard ML [52] programming language. CP-Nets support a strong mathematical foundation for the state space analysis and reachability analysis to verify numerous properties, including boundedness, liveness, fairness, terminating, and cycles in the model [53]. CP-Nets suffer from the problem of state space explosion when the number of states greatly increases. To overcome the state explosion problem, state space reduction methods are implemented through Linear-time Temporal Logic (LTL) [54] or Computation Tree Logic (CTL) [55].

III. LITERATURE REVIEW

A. XAI AND BLACK-BOX MODELS

The black-box issue in ML algorithms has been present in literature for decades [56], [57], [58]. The use of rule extraction to overcome this issue for neural networks is being discussed by research community [3], [59], [60], [61], [62], [63], [64], [65], [66]. However, there is very little work done on guaranteeing the correctness of these extracted rules. This research emphasizes visualizing the internal decision-making processes of black-box models using LIME and SHAP techniques. Additionally, it focuses on providing correctness guarantees for these internal decisions through the application of formal methods. This dual approach ensures not only the interpretability of the models' outputs but also the reliability and accuracy of their decision-making processes, thereby enhancing the trustworthiness and robustness of AI systems.

Wang et al. [67] proposed a rule extraction method to derive interpretable classification rules using the ensemble decision tree technique for the diagnosis of breast cancer. The Random Forest algorithm with a multi-objective evolutionary algorithm was used for the optimal classification rule to find the best trade-off between accuracy and interpretability. In [65], a shallow Artificial Neural Network model was proposed for breast cancer diagnosis. In [66], a hybrid approach called CWV-BANNSVM was proposed, combining boosting ANNs and two SVMs.

Jia et al. [68] proposed an approach to transform black-box conventional neural network models to decision trees using rule extraction. The authors have decomposed conventional neural networks into a feature extractor and a classifier. A human-readable decision tree was extracted from the classifier. They have built a visual tool to enable users to explore surrogate decision trees. Our work provides the formal guarantee of model correctness using Formal Methods. In [69], the design space of explainable artificial intelligence is explored to enhance the best design practices and future opportunities in the domain. A question bank is discussed on creating user-centered explainable artificial intelligence.

Bhatt et al. [70] discussed explainable ML in deployments and argued that ML algorithms are explained only to ML engineers but not to the end users of these algorithms. Roscher et al. [71] explored the recent scientific works and the contributions towards explainable ML in several application domains. In [72] a Dynamic Cell Structure neural network is proposed to build predictive models for forest fire detection and analyze environmental factors leading to forest fires. The rule extraction algorithms were applied to extract fire prediction rules.

Sultana et al. [73] used deep learning techniques to discover the knowledge from Twitter data for analyzing public sentiment towards education. The neural network model predictions were used to construct decision trees to extract rules. Ueno and Zhao [74] interpreted neural networks-based decision layers using decision trees. The research work focused on extracting interpretable knowledge from the hidden layers. Empirical results have demonstrated that accurate decision trees may be extracted from hidden layers.

Augasta and Kathirvalavakumar [75] presented a rule extraction algorithm RxREN to explain the predictive rules for trained neural networks. The presented algorithm relies on reverse engineering techniques for an explanation of neural network rules. In [76], the author proposed a rule extraction algorithm to extract rules from a neural network that contains continuous and discrete literals. The neural network decomposed by creating decision trees to obtain the production rules. These rules were merged to interpret the neural network. In [77], a rule extraction algorithm neural-network decision tree algorithm (ANN-DT) was proposed to extract binary decision trees from a trained neural network. It extracts rules from feed-forward neural networks with continuous output. The IF-THEN rules extraction algorithm from sample training data is presented in [78]. The dataset contained different attributes, values, and classes to cover a wide range of problems. Arslan et al. [79] proposed a web-based approach based on classification associations rule extraction using the R programming language.

B. XAI AND FORMAL METHODS

The integration of XAI and Formal Methods for proving the correctness has been attracting attention from the research community recently [23], [80], [81], [82], [83], [84].

Marques-Silva [80] emphasized that formal XAI, despite scalability challenges, offers more accurate and dependable alternatives to current XAI practices. Antonio et al. [81] highlighted the importance of integrating nonmonotonic reasoning and typicality-based logic to create explainable AI systems that cater to diverse user needs, aiming to overcome the limitations of traditional recommendation systems, which often reinforce user biases.

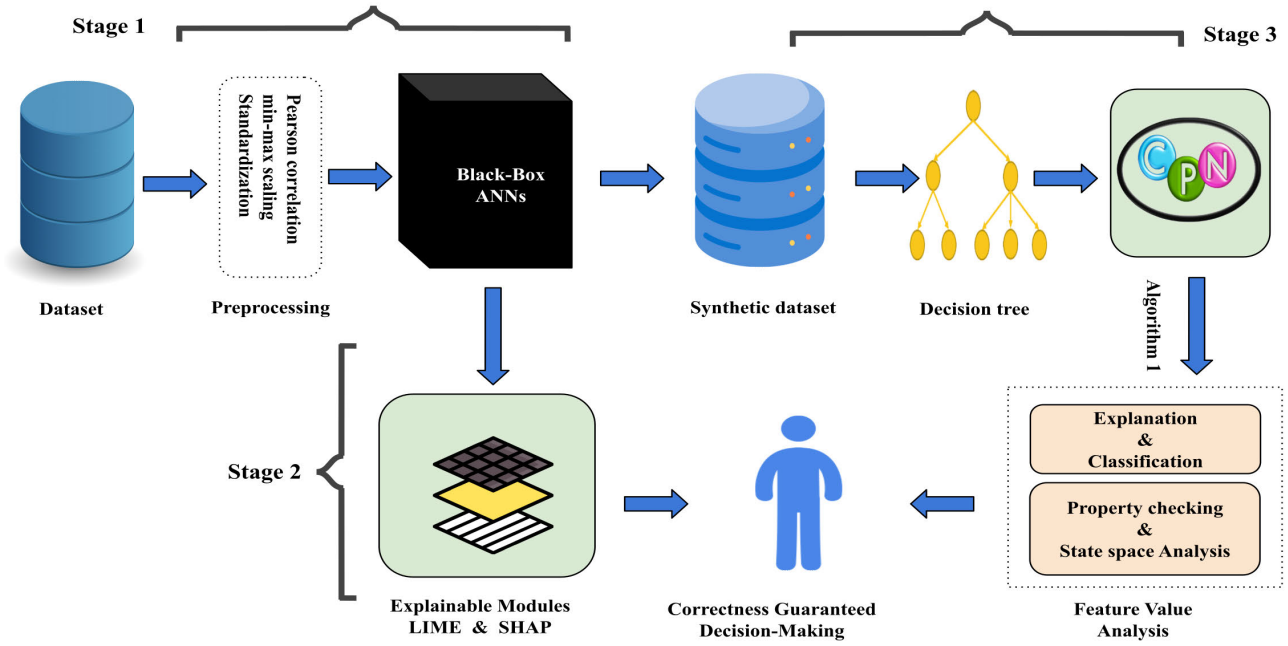

FIGURE 1. Proposed research methodology for guaranteeing correctness in AI-based medical decisions.

In [82], the authors formulate the creation of these mimic programs as a SyGuS problem, using if-then-else grammars to construct decision tree-like structures that replicate the model's decisions on given data points. They demonstrate their method's effectiveness through experiments on neural networks trained on the MNIST and Pima Indians diabetes datasets, showing that the synthesized mimic programs are both accurate and interpretable. They also highlight the potential of Formal Methods in providing reliable and comprehensible explanations for AI decision-making, particularly in critical applications like healthcare and autonomous systems.

Bassan and Katz [22] proposed a new method for efficient, verification-based minimal explanations that approximate the global minimum explanation. They also introduced heuristics to improve scalability and suggested using "bundles" for more interpretable explanations. In [84], the authors discussed the need for trustworthy AI in high-risk settings. They proposed formal XAI approaches that provide sound and irredundant explanations with formal guarantees of rigor.

Bassan et al. [85] discussed a novel approach for providing formal explanations of Deep Neural Networks within reactive systems, addressing the challenge of opacity and the limitations of existing heuristic-based Explainable AI techniques. The authors proposed a verification-based XAI technique that leveraged the system's transition constraints to efficiently compute minimal and minimum explanations, demonstrating significant improvements over state-of-the-art methods in terms of efficiency and reliability of explanations. The approach was evaluated using automated navigation benchmarks, highlighting its potential for enhancing trustworthiness and reliability in critical systems where DNNs are controllers.

In [86], the authors discuss the limitations of Shapley values, commonly known as SHAP scores, in XAI. They highlight that SHAP scores can provide misleading information about the importance of features in ML model predictions. The paper presents theoretical and empirical evidence showing that SHAP scores may incorrectly rank the significance of features, leading to potential misinterpretations by decision-makers. The authors analyze simple classifiers to demonstrate these issues and suggest that exact definitions of SHAP scores do not accurately reflect feature relevancy.

In conclusion, the body of work underscores the critical role of Formal Methods in advancing the field of explainable AI. By addressing the inherent limitations of current XAI practices, such as the opacity of deep neural networks and the potential misinterpretations of feature importance through methods like SHAP, these studies collectively advocate for more rigorous, scalable, and reliable approaches. Integrating formal guarantees, nonmonotonic reasoning, and novel techniques like syntax-guided synthesis and verification-based explanations significantly contribute to developing trustworthy AI systems, particularly in high-stakes environments such as healthcare and autonomous systems.

C. XAI IN CANCER DIAGNOSTIC

The recent research community has started to focus on the use of XAI in explaining medical diagnostics decisions [14], [15], [58], [87], [88], [89]. Khater et al. [58] devised machine-learning models to classify breast cancer and elucidate the model outcomes. Their research advanced the comprehension of breast cancer diagnosis and treatment by pinpointing crucial tumor features through the utilization of the SHAP algorithm. The results underscore the capacity

of machine learning to augment breast cancer diagnosis and therapy planning, underscoring the significance of interpretability and transparency in healthcare systems driven by artificial intelligence. Silva-Aravena et al. [15] presented a decision-support methodology for healthcare teams leveraging ML and explainability algorithms. Their investigation assessed diverse ML algorithms for patient classification into cancer and non-cancer groups, with XGBoost demonstrating the highest accuracy. Furthermore, employing the SHAP algorithm enabled the identification of significant variables, facilitating personalized early alerts for patients and augmenting clinical decision-making processes.

Hurtado et al. [89] conducted a study using a melanoma detection dataset to evaluate the effectiveness of the existing explainer against LIME. Their analysis conclusively demonstrated that LIME surpasses SHAP in diagnosing melanoma. This research underscores the advantages of employing XAI methods for interpreting model outcomes in melanoma image classification. Specifically, LIME exhibits superior performance compared to SHAP gradient explainer in terms of reproducibility and execution time. In [88], the authors introduced a two-stage XAI-MethylMarker framework, which is an explainable AI-based approach for biomarker discovery applied to DNA methylation data to identify a concise set of biomarkers for breast cancer classification. In the initial stage, they developed a deep-learning network incorporating an autoencoder for dimensionality reduction and a feed-forward neural network for breast cancer subtype classification. In the subsequent stage, they proposed a biomarker discovery algorithm utilizing various explainable techniques to analyze the model and identify a compact set of 52 biomarkers. Through 5-fold cross-validation, they attained a classification accuracy of 0.8145 Â± 0.07 with a 95% confidence interval. To validate the clinical significance of the discovered biomarkers, they conducted a gene set analysis, revealing 14 druggable genes, nine genes associated with prognostic outcomes, and several enriched pathways known to be significantly correlated with distinct breast cancer subtypes.

IV. MATERIALS AND METHODS

The proposed approach to interpret and provide formal correctness guarantees of the internal decision-making processes of black-box models is illustrated in Fig. 1. In this work, the internal decision-making is visualized using LIME and SHAP. The presented approach has three major steps: first, training black-box models using neural networks; second, employing LIME and SHAP techniques to interpret the models and visualize their internal decision-making processes; and finally, training decision trees on the synthetic datasets to employ Formal Methods for providing the correctness guarantees of internal decision-making process.

A. DATA COLLECTION

To illustrate our research methodology, experiments were performed with Wisconsin cancer and Thyroid cancer datasets available in the UCI ML repository [90]. The detail of these datasets is shown in Table 1.

The Wisconsin diagnosis cancer (WDBC) dataset includes 569 cases, including 32 attributes, with 212 malignant (M) and 357 benign (B) cases, respectively. Features are computed from a digitized image of a fine needle aspirate (FNA) of a mass and describe characteristics of the cell nuclei present in the image. The Wisconsin prognosis cancer (WPBC) dataset includes 198 instances with 34 features and contains 47 malignant and 151 benign cases, respectively. The WDBC and WPBC have the same fetuses, except WPBC has two extra features: tumor size and lymph node status. Tumor size is the diameter of the excised tumor in centimeters, and lymph node status is the number of positive axillary lymph nodes observed at the time of surgery. The feature values for the WDBC and WPBC datasets are shown in Table 2. The Wisconsin cancer (DSBC) dataset has 286 records, each with nine attributes excluding the ID number and classification label. It has 85 malignant and 201 benign cases, respectively. The feature values for the DSBC datasets are shown in Table 3.

The thyroid cancer (TC) dataset comprised 12 clinical and demographic patient attributes, except for the ATA risk score, which included 383 cases. It contains 108 malignant and 275 benign cases. The feature values for the TC dataset are shown in Table 4.

B. PRE-PROCESSING

Pre-processing steps were applied to a dataset before applying ML algorithms to enhance the learning process of the training model [91]. The pre-processed data was divided into training and test data using a 10-fold cross-validation methodology.

In the case of the WDBC, the patient's ID feature does not have a major impact on the diagnosis of the disease and thus has been removed. No additional pre-processing was conducted for WDBC, as no missing values were found.

In the case of the DSBC dataset, 9 instances with missing values were found and removed. In ML models, it is often necessary to convert categorical text features into their numeric representation. Consequently, the categorical values were converted into numerical values. In the WPBC dataset, 4 instances with missing feature ''Lymph node status'' values were found and removed. No additional pre-processing was conducted for the WPBC dataset. In the case of the TC dataset, the categorical values were converted into numerical values.

Additionally, the data is normalized and scaled to ensure that the features are on a consistent scale. This normalization process helps prevent any particular feature from dominating the model's training process due to its larger magnitude.

C. BLACK-BOX MODELS

Black-box neural network models were trained on WDBC, DSBC, WPBC, and TC datasets. The architecture of these black-box ANN models is shown in Table 5, and the neural

TABLE 1. Wisconsin breast cancer datasets.

Dataset	No. of attribute	No. of instances	No. of class	Instances with missing values
Wisconsin diagnosis breast cancer (WDBC)	32	569	2	-
Wisconsin breast cancer (DSBC)	11	277	2	9
Wisconsin prognosis breast cancer (WPBC)	34	198	2	4
Thyroid cancer (TC)	12	383	2	-

TABLE 2. Description of the features in WDBC and WPBC.

Feature	Description	Mean	Standard error	Worst/largest value
Radius	Mean of the distances between the center and the points of the perimeter	6.98–28.11	0.11–2.87	7.93–36.04
Texture	Standard deviation of gray level values	9.71–39.28	0.36–4.89	12.02–49.54
Perimeter	The total distance between the snake points constitute the nuclear perimeter	43.79–188.50	0.76–21.98	50.41–251.20
Area	Measured simply by counting the number of pixels on the interior of the snake	143.5–2501.00	6.80–542.20	185.20–4254.00
Smoothness	Local variation in radius lengths	0.05–0.16	0.00–0.03	0.07–0.22
Compactness	Perimeter^2/area - 1	0.02–0.35	0.00–0.14	0.03–1.06
Concavity	Severity of concave portions of the contour	0.00–0.43	0.00–0.40	0.00–1.25
Concave points	Number of concave portions of the contour	0.00–0.20	0.00–0.05	0.00-0.29
Symmetry	The major axis or longest chord through the center	0.11–0.30	0.01–0.08	0.16–0.66
Fractal dimension	Coastline approximation - 1	0.05–0.10	0.00–0.03	0.06–0.21

TABLE 3. Description of the features in DSBC.

Feature Name	Feature Values	Type
Class	No-recurrence-events(Benign), Recurrence-events(Malignant)	Dichotomous
Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99	Linear
Menopause	Lt40, Ge40, Premeno	Nominal
Tumor-Size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59	Linear
Inv-Nodes(IN)	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39	Linear
Node-Caps	Yes, No	Dichotomous
Deg-Malig	1, 2, 3	Linear
Breast	Left, Right	Dichotomous
Breast-Quad(BQ)	Left-up, Left-low, Right-up, Right-low, Central	Nominal
Irradiat	Yes, No	Dichotomous

TABLE 4. Description of the features in TC.

Feature Name	Feature Values	Type
Age	15-82	Linear
Gender	F, M	Dichotomous
Smoking	Yes, No	Dichotomous
Hx Smoking	Yes, No	Dichotomous
Hx Radiothreapy	Yes, No	Dichotomous
Thyroid Function	Clinical Hyperthyroidism, Clinical Hypothyroidism, Euthyroid, Subclinical Hyperthyroidism, Subclinical Hyperthyroidism	Nominal
Physical Examination	Diffuse goiter, Multinodular goiter, Normal, Single nodular goiter-left, Single nodular goiter-right	Nominal
Adenopathy	Bilateral, Extensive, Left, No, Posterior, Right	Nominal
Pathology	Follicular, Hurthel cell, Micropapillary, Papillary	Nominal
Focality	Multi-Focal, Uni-Focal	Dichotomous
Risk	High, Intermediate, Low	Nominal
T	T1a, T1b, T2, T3a, T3b, T4a, T4b	Nominal
N	N0, N1a, N1b	Nominal
M	M0, M1	Dichotomous
Stage	I, II, III, IVA, IVB	Nominal
Response	Biochemical Incomplete, Excellent, Indeterminate, Structural Incomplete	Nominal
Recurred	No(Benign), Yes(Malignant)	Dichotomous

network trained on the DSBC dataset is demonstrated in Fig. 2. To reduce the features, the final input features were selected using the Pearson correlation coefficient to enhance performance. This method involves calculating the linear correlation between pairs of features and eliminating those that exhibit high redundancy, thereby retaining only the most informative and uncorrelated features for model training. By applying this technique, we aim to improve model accuracy and generalization by minimizing multicollinearity and ensuring that each feature contributes unique information to the predictive process.

The number of features selected using the Pearson correlation coefficient algorithm for the DSBC dataset was 4. The neural network's architecture trained on the DSBC

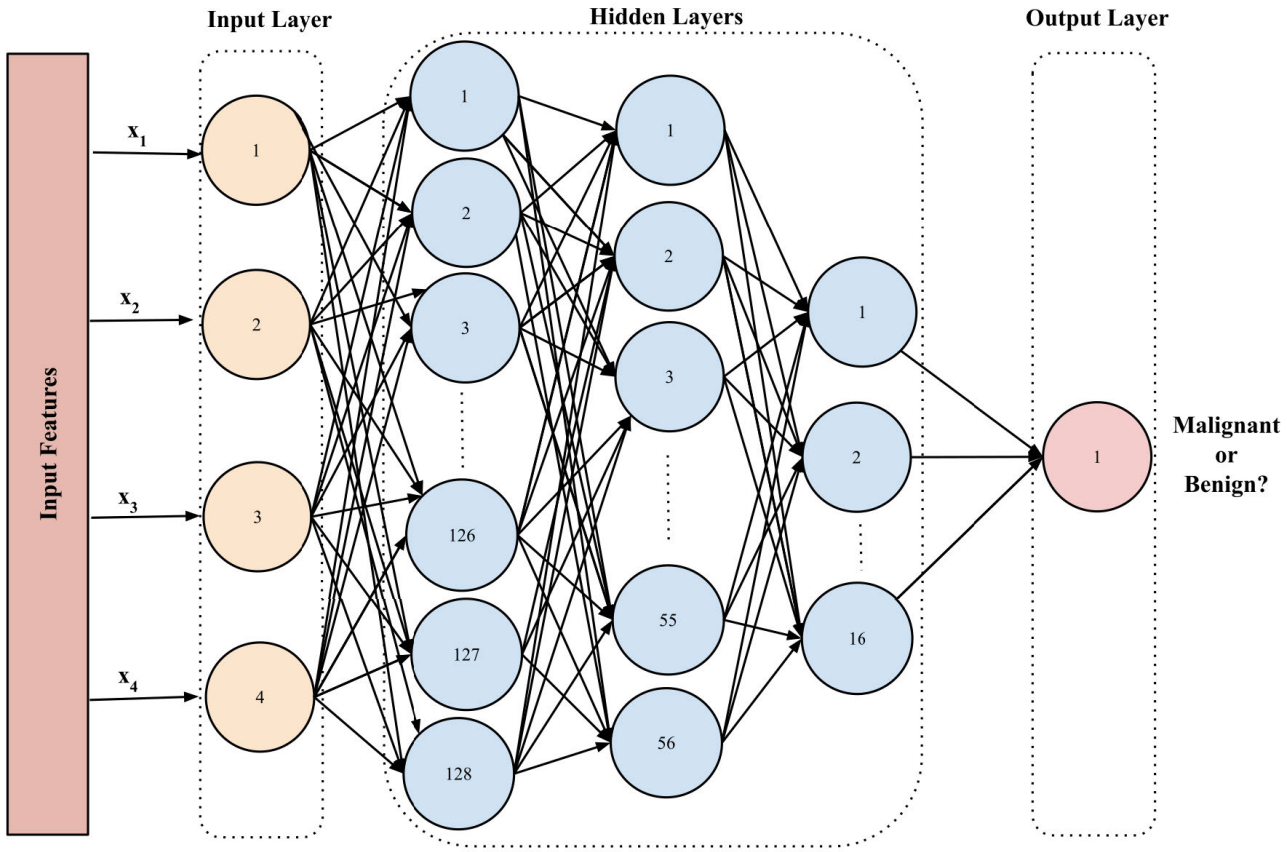

FIGURE 2. Breast cancer classification using neural networks on DSBC data.

TABLE 5. Architectural details of Black-Box neural networks for cancer classification.

Layer Type	WDBC			DSBC			WPBC			TC		
	Units	Activation Function	Output Shape	Units	Activation Function	Output Shape	Units	Activation Function	Output Shape	Units	Activation Function	Output Shape
Input	10	-	(1,15)	4	-	(1,4)	12	-	(1,12)	5	-	(1,5)
Hidden 1	256	ReLU	(1,256)	128	ReLU	(1,128)	56	ReLU	(1,56)	56	ReLU	(1,56)
Hidden 2	128	ReLU	(1,128)	56	ReLU	(1,156)	32	ReLU	(1,32)	32	ReLU	(1,32)
Hidden 3	56	ReLU	(1,56)	16	ReLU	(1,16)	16	ReLU	(1,16)	16	ReLU	(1,16)
Hidden 4	32	ReLU	(1,32)	-	-	-	-	-	-	-	-	-
Output	1	-	(1,1)	1	-	(1,1)	1	-	(1,1)	1	-	(1,1)

dataset has 4 neuron units in the input layer and hidden layers with 128, 56, and 32 neurons. Similarly, the number of features selected for the WDBC dataset was 15. Accordingly, the neural network trained on WDBC has 15 neurons on the input layer and hidden layers with 256, 128, 56, and 32 neurons.

On the other hand, the architecture of the neural network trained on the WPBC dataset has 12 neuron units in the input layer, and the hidden layers have 56, 32, and 16 neurons. Moreover, all the neural networks have a single neuron at the output layer with the rectified linear unit (ReLU) activation function to classify the input as malignant or benign.

D. EXPLAINING BLACK-BOX DECISION-MAKING

1) LIME

At this phase, we utilize the LIME framework, a method crafted to clarify specific predictions through the creation of a local, understandable model that approximates the behavior of any complex machine learning model. This procedure entails modifying the original data points, feeding them into an inscrutable ML model, and examining the outcomes of these predictions. Fig. 3 (a), (b), (c), and (d) illustrate the discovery of features for the WDBC, WPBC, DSBC, and TC datasets using LIME. LIME has effectively highlighted the most crucial characteristics contributing to the expected diagnostic outcomes.

2) SHAP

Upon choosing the ANN algorithm, we decided to enhance the tool provided to doctors and administrative health teams by integrating an interpretability algorithm. This addition is designed to clarify the model's patient classification process. Fig. 4 (a) and (c) showcase the identification of the most impactful variables within the ANN model based

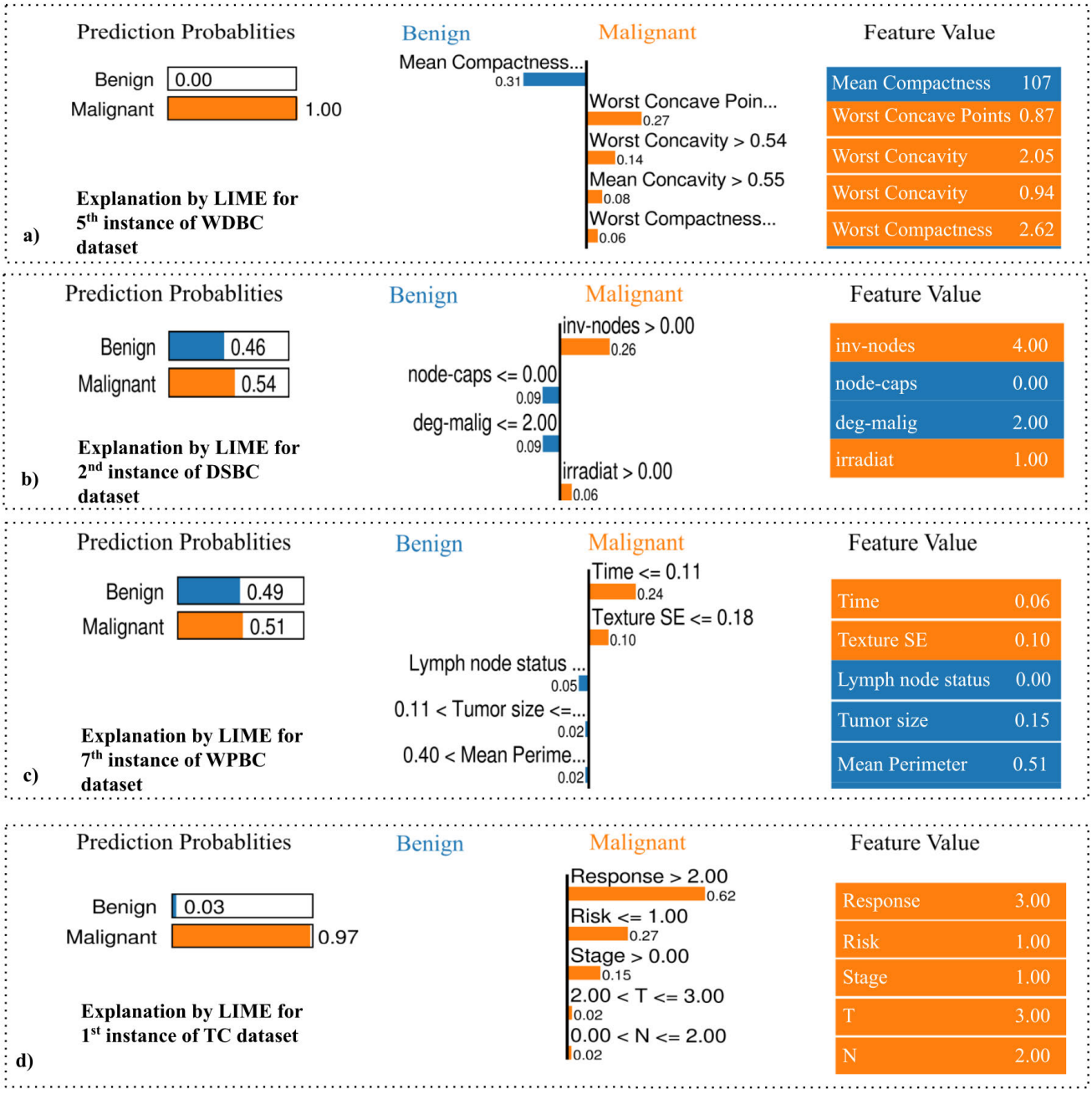

FIGURE 3. LIME local explanations for cancer datasets.

on the patient data at hand. Notably, ''Worst Area'' and ''Time'' emerge as crucial variables, enabling the model to distinguish between healthy individuals and those diagnosed with cancer. Figure 4 (b) highlights the significance of the ''inv-nodes'' feature, which plays a key role in enhancing the model's interpretability. Similarly, fig. 4 (d) highlights the significance of the ''Response'' and ''RISK'' features, which play a key role in enhancing the model's interpretability. Thus, the combination of the ANN model with the SHAP interpretability algorithm significantly enriches the information available for healthcare teams' decision-making processes.

E. CORRECTNESS GUARANTEES

Most existing XAI techniques rely on heuristics and fail to ensure the correctness of their explanations. In contrast, recent advancements have shown that formal methods can generate explanations with provable correctness, significantly improving the reliability and trustworthiness of AI systems [22]. The main advantage of the approach presented is the application of CP-Nets to demonstrate the correctness of prognostic decisions. The decision trees were induced from the synthetic datasets and produced using original input features with the classification predictions of the black-box models to mimic the transparency of equivalent

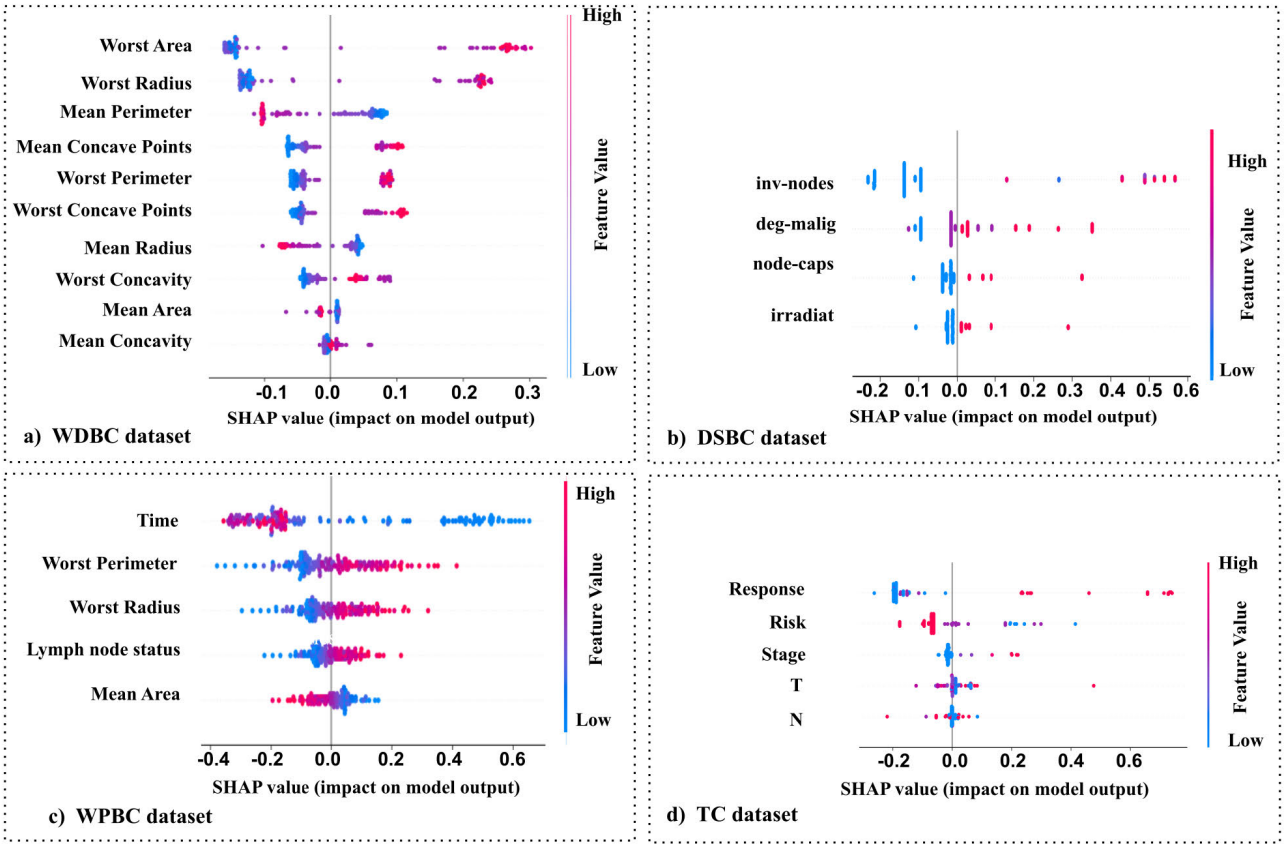

FIGURE 4. SHAP summary plots for cancer datasets.

internal decisions. The prognostic decisions were derived by traversing the path from the root node to the leaf nodes as described in [92]. The Hierarchical CP-Nets model used for formal verification is shown in Fig. 6. To confirm the correctness of CP-Nets colour tokens were constructed using cancer datasets.

More precisely, the following steps were performed to extract decision choices and provide formal correctness guarantees of decision-making for black-box models.

1) DECISION CHOICE AND CP-NETS TRANSFORMATIONS

In literature, the decision tree has emerged as a useful tool to transform black-box models into transparent models using the synthetic datasets [68]. The synthetic datasets consisted of original input features with corresponding output classification labels produced by the black-box models. To induce decision trees, an open-source implementation of C4.5 [93] in WEKA [94] was used with a ten-fold cross-validation. The decision tree in Fig. 5 exhibits decision points for the cancer datasets in the tree structure.

Decision choices in the models needed to be identified to transform black-box models into the CP-Nets model. The transparent decision trees were induced using synthetic datasets to identify decision choices. From these transparent decision trees, predictive decision decisions were extracted using a path traversing through the root to the leaf nodes

Algorithm 1 Decision Tree to CP-Nets Model Transformations

Input: Decision Tree
Output: CP-Net Sub-module
for *Decision Paths in Decision Tree* **do**
  Create a Decision Rule for the Decision Path Add CP-Net Transition for Decision Rule
  for *Attributes in Decision Rule* **do**
    Add Guard to transition for Attribute
  end
end

using transformation algorithm 1. To be more specific, to guarantee the correctness of prognostic decisions for a particular patient, which needs to be either classified as malignant or benign, CP-Nets models were constructed for all paths in decision trees.

2) CORRECTNESS GUARANTEES FOR INTERNAL DECISION-MAKING

CPN tools [95] provide mechanisms to simulate models and construct state space reports. The presence of errors in the models can be analyzed via simulations, but it cannot guarantee its absence. On the other hand, the state space analysis can be exploited to check the absence of an error in

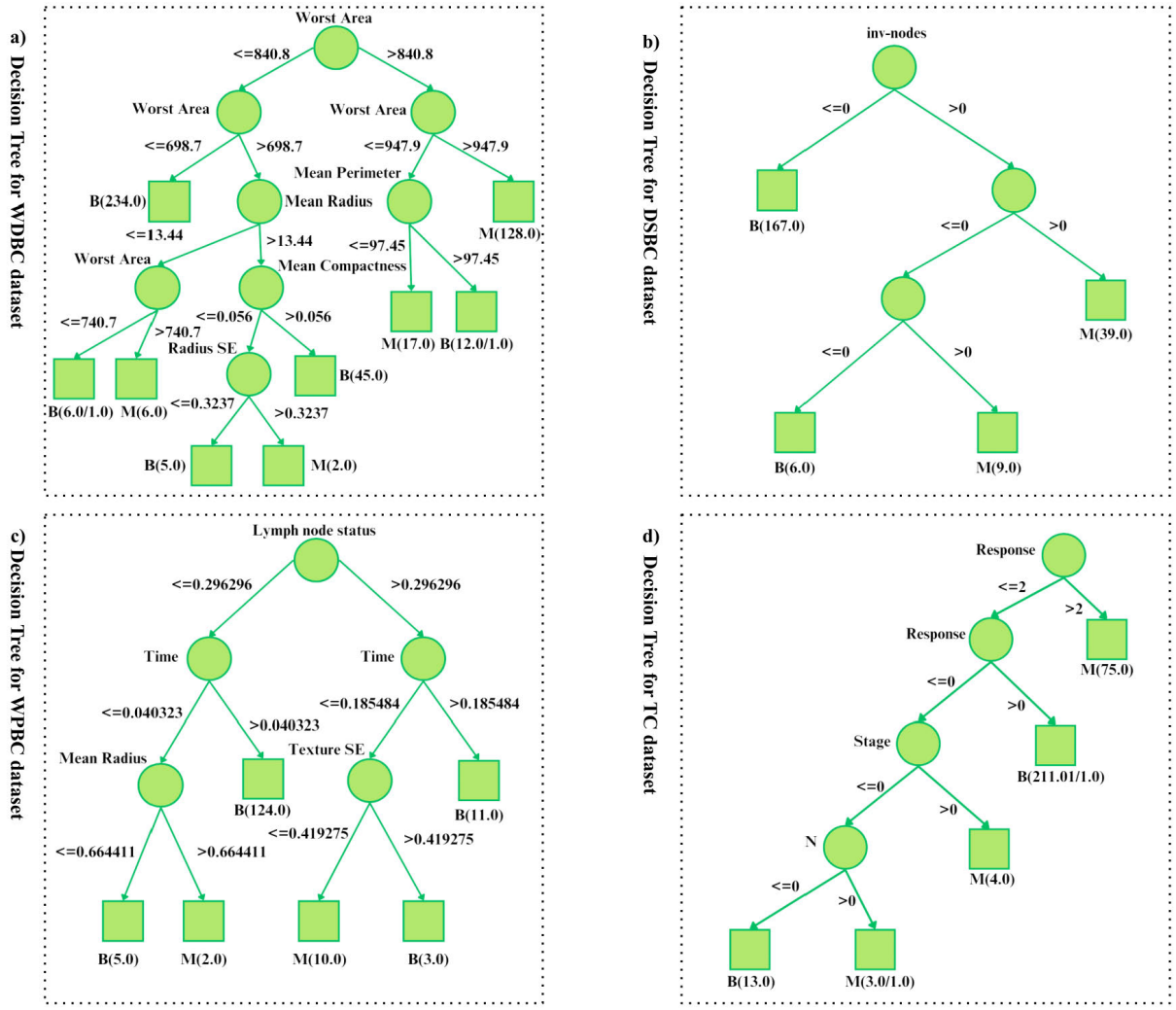

FIGURE 5. Decision trees trained on synthetic cancer datasets.

the model. CP-Nets, as a representation of the equivalent of prognostic decisions, was selected because of the expressive support it provides and the strong simulation capabilities. Furthermore, it also leverages the hierarchy concept to allow the composition in a modular way. In addition, the availability of probability distribution support allows for modeling performance aspects. Finally, data-related aspects are supported by the introduction of colour tokens.

The model's formal verification begins at the start of the simulation from the first transition state, which is called initial marking. Every prognosis decision was modeled as a transition in CP-Nets. Prognostic decision-making choices were verified in a three-stage process. First, all the color tokens were loaded on the initial marking. Second, the simulation tool was used to simulate the color tokens. In the third and final stage, the colour tokens' reachability was analyzed for the transition *Correctness_Verification*. The sub-module *Predicted_Diagnosis_Verifier* was used to analyze and compare colour tokens with actual and predicted labels. This comparison on the transition place *Correctness_Verification* provided the correctness profs for the prognostics decision-making choices. All CP-Nets models are accessible on the link.[1]

Fig. 6 demonstrates the hierarchical CP-Nets model proposed in this research work. The CP-Nets model architecture consists of three sub-models, namely *Diagnosis_M_Model*, *Diagnosis_B_Model*, and *Predicted_Diagnosis_Verifier*. The *Diagnosis_M_Model* and *Diagnosis_B_Model* sub-modules were responsible for modeling malignant and benign prognostic choices, respectively. The third sub-module, namely *Predicted_Diagnosis_Verifier*, is responsible for verifying the colour tokens with predicted and actual labels to confirm the correctness of diagnostic choices.

In addition, the reachability graph computed by the CPN tool is analyzed for formal verification of specific properties and discussed in the section below.

[1]https://drive.google.com/file/d/11ivjh4CsBFoz5UGM_rg5N9hiOxW328VJ/view?usp=sharing

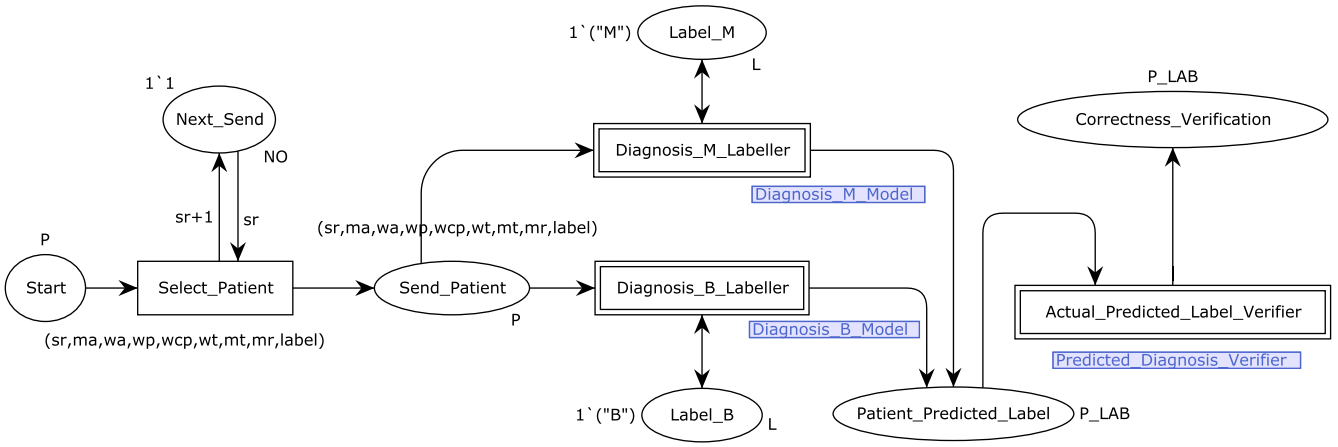

FIGURE 6. Hierarchical colored petri nets model for formal correctness proofs.

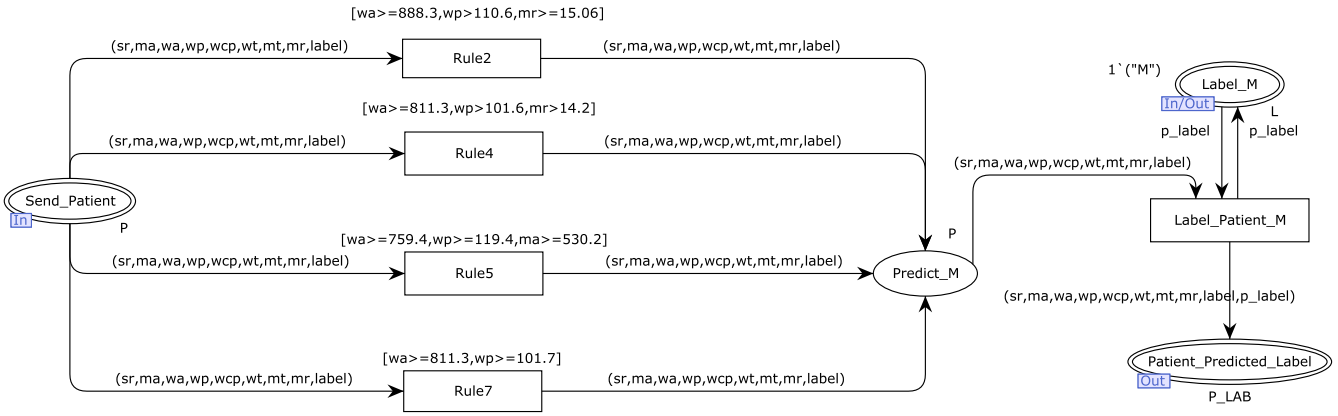

FIGURE 7. CP-Net Sub-Model: ensuring correctness of prognostic decisions for malignancy.

3) MALIGNANT CASES VERIFICATION

Property 1: Malignant sub-module always predicts the correct label. The decision specifies that when a token satisfies the criteria for the **"M"** classification, then the model is expected to assign a **"M"** prediction label and direct it to the *Patient_Predicted_Label* place. The labeling within the *Patient_Predicted_Label* ensures that all tokens that pass through transitions guarded with decision-making logic within the *Classifier_M_Labels* sub-modules, are classified as **"M"**.

Property 2: Classifiers put "M" label on colour tokens for labeling them according to decision choice. Marking the place *Label_M* on transition *Label_Patient_M* confirms that colour tokens are always labeled as **"M"** at the *Patient_Predicted_Label* place.

4) BENIGN CASES VERIFICATION

Property 1: Benign sub-module always predicts correct labels. The property indicates that if a token fulfills the criteria for the **"B"** classification, the model is expected to apply a **"B"** predictive label and pass to the *Patient_Predicted_Label* place. The labeling within the *Patient_Predicted_Label* ensures that all tokens that pass through transitions guarded with decision-making logic within the Patient_B_Labels sub-modules are classified as **"B"**.

Property 2: Classifiers put a "B" label on colour tokens to label them according to classification decision choice.

The marking of the place *Label_B* during the transition *Label_Patient_B* confirms that colour tokens are always labeled as **"B"** at the *Patient_Predicted_Label* place.

5) ACTUAL AND PREDICTED LABELS VERIFICATION SUB-MODULE

Property 1: The colour tokens with actual and predicted labels must reach Compare_Actual_Predicted_Labels place.

The property asserts that colour tokens, influenced by classification decision-making, ought to arrive at *Compare_Actual_Predicted_Labels* place for the comparison of actual and predicted labels. It's important to highlight that, at this point, a comparison of labels is employed to confirm the correctness of the decisions made by the classifiers.

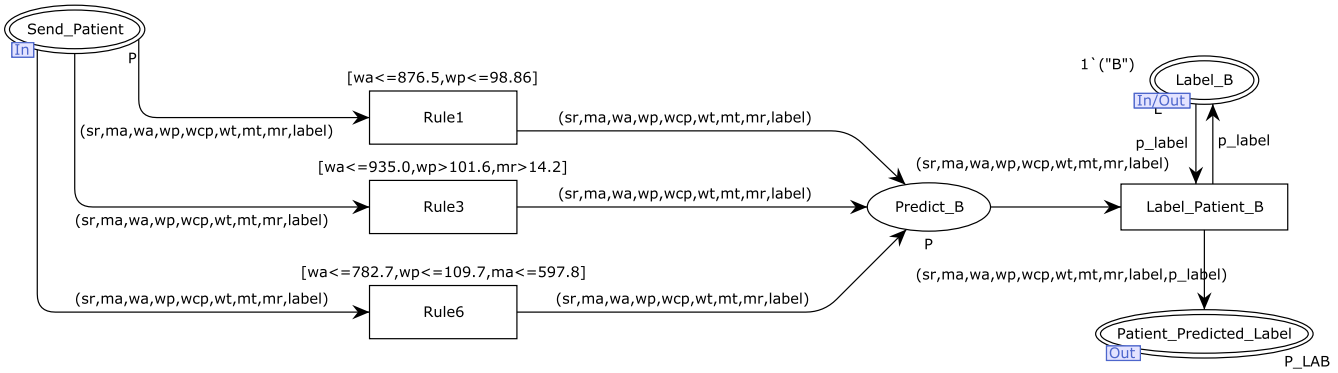

FIGURE 8. CP-Net Sub-Model: verification of prognostic decisions for benign cases.

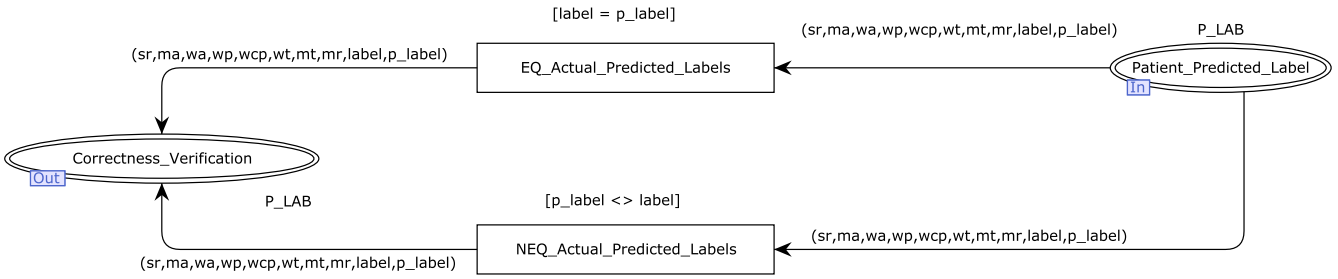

FIGURE 9. CP-Net Sub-Model: verifying prognostic decisions.

F. PERFORMANCE METRICS

The ML models were trained to predict the patient's cancer status, either as malignant or benign. The Confusion Matrix was calculated to classify cases as shown in Table 6. The performance of the proposed approach is evaluated according to Accuracy (ACC), Sensitivity(TPR), and Specificity(TNR). Here, the total number of instances is represented as N and computed such as:

$$N = TP + TN + FP + FN. \tag{1}$$

Accuracy (ACC): ACC specifies the accuracy of the classification model such that:

$$ACC = (TP + TN)/N. \tag{2}$$

Sensitivity (TPR): TPR specifies the correct classification rate of positive instances such as:

$$TPR = TP/(TP + FN). \tag{3}$$

Specificity (TNR): TNR specifies the correct classification rate of negative instances such as:

$$TNR = TN/(TN + FP). \tag{4}$$

G. RESULTS

Once the performance of the ML model has been evaluated, it becomes imperative to elucidate and scrutinize the findings to gain insights into the model's performance. This involves discerning the crucial features influencing the model's predictions, comprehending the relationships between these features and the target variable, and identifying any pertinent patterns or trends within the dataset. This study employed two model-agnostic techniques, namely LIME and SHAP, for this purpose. Additionally, the proposed approach implemented CP-Nets formalism to provide correctness guarantees for diagnostic decision-making.

TABLE 6. Confusion matrix.

		Predicted Labels	
		M	B
Actual Labels	M	TP	TN
	B	FP	FN

The effectiveness of the proposed approach is illustrated using cancer datasets from the University of Wisconsin. The black-box ANN models were trained on the cancer dataset. LIME and SHAP were employed to provide interpretations of internal decision-making. The results of Table 7 and Table 8 empirically demonstrate that our approach has improved performance as compared with the baseline classifiers in terms of accuracy, sensitivity, and specificity. The analysis of Table 7 and Table 8 reveals that the true negative rate (TNR) attains a value of 100% for the WPBC dataset. This observation substantiates the assertion that the proposed approach effectively safeguards benign instances with high accuracy.

TABLE 7. Comparisons of the proposed approach with other supervised ML methods.

Method	WDBC			DSBC			WPBC			TC		
	TPR (%)	TNR (%)	ACC (%)	TPR (%)	TNR (%)	ACC (%)	TPR (%)	TNR (%)	ACC (%)	TPR (%)	TNR (%)	ACC (%)
ANN	97.25	97.31	97.14	60.35	81.07	74.16	24.61	90.86	76.27	97.46	80.55	92.17
SVM	95.98	94.44	96.82	25.00	82.14	77.66	15.38	100.00	77.14	93.67	77.77	90.63
K-NN	95.21	95.86	94.02	26.66	91.93	70.65	5.55	97.87	72.30	96.66	72.97	89.76
RF	95.73	97.84	92.64	38.66	87.64	73.92	12.73	98.80	76.61	100.00	89.47	96.52
NB	92.39	93.51	90.47	53.57	69.64	64.28	23.07	78.26	66.10	97.46	72.22	89.56
XGBoost	93.65	98.14	96.49	35.71	82.14	66.66	23.07	89.13	74.57	97.46	83.33	93.04
Bagging	90.47	95.37	93.56	39.28	83.92	69.04	15.38	93.47	76.27	96.20	83.33	92.17
Boosting	95.23	99.07	97.66	46.42	76.78	66.66	15.38	93.47	76.27	97.46	83.33	93.04
Proposed approach	97.67	97.18	97.00	42.10	86.48	71.00	57.14	100.00	92.00	98.27	84.21	95.00

TABLE 8. Comparisons of the proposed approach with other decision extraction methods.

Method	WDBC			DSBC			WPBC			TC		
	TPR (%)	TNR (%)	ACC (%)	TPR (%)	TNR (%)	ACC (%)	TPR (%)	TNR (%)	ACC (%)	TPR (%)	TNR (%)	ACC (%)
Decision Tables	93.75	89.59	96.91	97.82	64.28	92.16	38.46	92.22	80.17	95.80	9.00	95.82
OneR	88.5	87.28	89.42	92.02	89.28	91.56	53.84	95.55	86.20	92.70	13.60	92.68
PART	91.75	89.59	93.39	95.65	85.71	93.97	80.76	95.55	91.75	94.80	8.00	94.77
RIPPER	92.5	91.32	93.39	90.57	71.42	87.34	69.23	93.33	87.93	94.00	10.20	93.99
Proposed approach	97.67	97.18	97.00	42.10	86.48	71.00	57.14	100.00	92.00	98.27	84.21	95.00

1) LIME

In our investigation, we utilized LIME as a potent tool for interpretability to delve into the predictions generated by our ML models. The model-agnostic characteristic of LIME enabled us to deploy it across diverse ML models, regardless of their architectures or training algorithms [96]. Leveraging LIME, we elucidated the predictions of binary classifications concerning cancer datasets. Through a focus on the key relevant features identified by LIME, we presented visual confirmation of our ML model's efficacy and provided insights into the underlying patterns associated with different classes of cancer. To summarize, our research harnessed LIME's capabilities to uncover the decision-making mechanisms of our ML models in the realm of cancer diagnostics.

Fig. 3 (a), (b), (c), and (d) illustrate the discovery of features for WDBC, DSBC, WPBC, and TC datasets using LIME. LIME has effectively highlighted the most crucial characteristics contributing to the expected diagnostic outcomes. Among the top relevant features in the WDBC dataset, the ''Worst Concave Point'' feature exhibited the highest impact on the model's estimation capability. Fig 3 (c) indicates that ''TIME'' and ''Texture SE'' features have the most significant influence. Similarly, Figure 3 (b) demonstrates that ''inv-nodes'' is the top impacting feature for DSBC dataset. Through LIME, it becomes evident that these attributes play essential roles in cancer diagnosis.

2) SHAP

SHAP method [97] relies on Shapley values, which offer explanations for individual instances rather than overarching ones. By leveraging Shapley values, we can ascertain the significance of each feature in a specific prediction. When our primary concern is understanding the importance of features for a particular prediction rather than gaining insight into the model's general behavior, SHAP proves valuable. Utilizing SHAP [98], the prediction for a given instance 'x' is elucidated by computing the contribution of each feature. In this study, the SHAP summary plot was generated to evaluate the impact of various features on cancer classification. As depicted in Fig. 4, the ''Worst Area'' feature emerges as the most influential, corroborating findings from permutation analysis for WDBC and WPBC datasets. The feature ''inv-nodes'' is most influential for the DSBC dataset.

Shapley values were computed for selected top relevant features identified through Pearson correlation to develop an interpretable ML model. This approach aimed to discern the impact of these features on predicting cancer classes. Concerning interpretability, the feature ''Worst Area'' represents the total area occupied by the nucleus and stands out as the most influential factor in the classification of cancer. The SHAP plot depicted in Fig. 4 demonstrates that larger values of the area feature positively impact the classification task. In other words, as the area feature increases, the model tends to predict a higher likelihood of cancer.

In previous literature, researchers typically utilize a single XAI method to identify influential features within a dataset. Contrary to this, our extensive research demonstrates the efficacy of incorporating a diverse array of XAI techniques, enabling a more thorough comprehension of the underlying factors shaping the dataset and its predictive outcomes. Furthermore, our study endeavors to establish formal correctness guarantees for XAI results. This synergistic approach not only enhances our understanding of cancer but also

potentially unveils novel findings that might otherwise remain undiscovered when studying each dataset in isolation.

H. DISCUSSION

The effective use of ML in safety-sensitive applications like the medical field has increased over the past several years. It has raised a new challenge for the research community to explain internal ML decisions for achieving a particular outcome. The proposed methodology sought to explain black-box ML models using SHAP and LIME. It also proposed the use of CP-Nets formalism to guarantee the correctness of internal decision choices of black-box ML models. To be more precise, this research concludes that the black-box ML model internal decision-making can be verified using CP-Nets formalism.

Khater et al. [58] advocated for the adoption of the SHAP algorithm to interpret ML models designed for cancer classification. This method promises to deep understanding of cancer diagnosis and treatment by identifying pivotal tumor features crucial for accurate classification. The leading ML model achieved a 97.7% accuracy employing k-nearest neighbors, with a precision of 98.2%, utilizing the Wisconsin cancer dataset. Additionally, an accuracy of 98.6% was reached using an ANN, with a precision of 94.4%, based on the Wisconsin diagnostic cancer dataset.

Confalonieri et al. [3] introduced an ontology-based approach to improve human understanding of black-box models using surrogate decision trees. The Trepan Reloaded algorithm is introduced, which extracts the surrogate decision trees from black-box models. In this work, CP-Nets with a surrogate decision tree classifier were used to interpret the internal functioning of the ANN black-box models and formally prove the correctness of prognostic decisions. In [68], a methodology is proposed to explain black-box deep neural network models. It decomposed the convolutional neural network into a feature extractor classifier and later extracted the decision trees to explain the internal decision-making of the black-box model. This work uses SHAP and LIME to interpret the internal functioning of the ANN black-box models and CP-Nets formalism is implemented to provide the correctness proof of internal decision choices.

Gorzałczany and Rudziński [59] proposed a rule extraction method to derive accurate and interpretable classification rules using multi-objective evolutionary optimization algorithms. Wang et al. [67] proposed a rule extraction method to derive interpretable classification rules using the ensemble decision tree technique for the diagnosis of cancer. The Random Forest algorithm with a multi-objective evolutionary algorithm was used for the optimal classification rule to find the best trade-off between accuracy and interpretability. The approach showed an accuracy of 97%. In [65], a shallow Artificial Neural Network (ANN) model was proposed to diagnose and predict cancer using the Wisconsin breast cancer datasets without employing feature optimization or selection algorithms. The approach showed promising performance with an average accuracy of 99.85%, specificity of 99.72%, sensitivity of 100%, precision of 99.69%. In [66], a hybrid approach called CWV-BANNSVM, combining boosting ANNs and two SVMs, was proposed with an accuracy of 99.7%. Our methodology has 98% accuracy with an improved interpretation of decision choices and also demonstrates the correctness of these prognostic decisions with CP-Nets formalism.

To the best of our knowledge, this is the first attempt to apply CP-Nets formalism to provide correctness guarantees for ML-based prognostic decisions. The empirical results demonstrate that the application of state space analysis not only enhances the interpretability of these decisions but also ensures the correctness of the medical prognosis process. Consequently, our proposed approach significantly improves the quality of the medical decision-making process, offering a more reliable and accurate framework for healthcare applications.

It should be noted that this study's results are not generalizable on a broader scale. Therefore, further work on comparing the results with expanded datasets from other medical datasets would be beneficial. Despite this limitation, the findings remain important in medical data analysis.

V. CONCLUSION

The ML models pose black-box behaviors that reduce the fidelity of internal decision-making rules. Therefore, there is a need for a mechanism to explain, interpret, and build correctness proofs for these black-box models. In this research, a formal approach was developed to provide correctness guarantees for the internal decision-making of black-box models using Formal Methods. The black-box models were interpreted by SHAP and LIME algorithms. The primary aim of this research is to enhance the understanding of black-box models and to provide guarantees of their accuracy. By employing formal methods, this work ensures that the internal decision-making processes of these models are both interpretable and verifiably correct, thereby improving the reliability and trustworthiness of machine learning applications in critical domains. The equivalent decision trees were surrogated using the synthetic datasets to prove the correctness of the black-box model's internal decision-making. In the proposed approach, black-box models were first trained using neural networks to generate synthetic datasets. Subsequently, SHAP and LIME techniques were employed to interpret these models and visualize their internal decision-making processes. Finally, decision trees were trained on the synthetic datasets to implement Formal Methods, ensuring the correctness of the black-box models' decision logic. This multi-step process aims to enhance the interpretability and reliability of machine learning models by combining model-agnostic explainability techniques with Formal Methods.

In this work, we explored only ANNs-based black-box models using CP-Nets formalism. The proposed approach has limitations and requires further investigation. In future studies, we plan to examine other black-box machine learning

techniques for interpretive decision-making, including SVM, Random Forest, and Convolutional Neural Networks. Additionally, it is worth mentioning the potential for integrating fuzzy techniques with CP-Nets formalism to enhance decision rule interpretations. While this work focused solely on CP-Nets formalism, other formal verification techniques and tools, such as UPPAAL, SAT solvers, and VDM, still need to be investigated to validate the correct interpretations of black-box models.

REFERENCES

[1] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[2] M. Lang, A. Bernier, and B. M. Knoppers, "AI in cardiovascular imaging: 'Unexplainable' legal and ethical challenges?" *Can. J. Cardiol.*, vol. 38, no. 2, pp. 225–233, 2021.

[3] R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artif. Intell.*, vol. 296, Jul. 2021, Art. no. 103471.

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. D. 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.

[5] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—A brief history, state-of-the-art and challenges," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2020, pp. 417–431.

[6] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[7] E. Pastor and E. Baralis, "Explaining black box models by means of local rules," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 510–517.

[8] H. J. P. Weerts, W. van Ipenburg, and M. Pechenizkiy, "A human-grounded evaluation of SHAP for alert processing," 2019, *arXiv:1907.03324*.

[9] D. Chakraborty, C. Ivan, P. Amero, M. Khan, C. Rodriguez-Aguayo, H. Başağaoğlu, and G. Lopez-Berestein, "Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer," *Cancers*, vol. 13, no. 14, p. 3450, Jul. 2021.

[10] S. Kawakura, M. Hirafuji, S. Ninomiya, and R. Shibasaki, "Analyses of diverse agricultural worker data with explainable artificial intelligence: XAI based on SHAP, LIME, and LightGBM," *Eur. J. Agricult. Food Sci.*, vol. 4, no. 6, pp. 11–19, Nov. 2022.

[11] D. Chen, H. Zhao, J. He, Q. Pan, and W. Zhao, "An causal XAI diagnostic model for breast cancer based on mammography reports," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 3341–3349.

[12] S. C. Chelgani, H. Nasiri, A. Tohry, and H. Heidari, "Modeling industrial hydrocyclone operational variables by SHAP-CatBoost—A 'conscious lab' approach," *Powder Technol.*, vol. 420, Apr. 2023, Art. no. 118416.

[13] H. Hakkoum, A. Idri, and I. Abnane, "Artificial neural networks interpretation using LIME for breast cancer diagnosis," in *Trends and Innovations in Information Systems and Technologies*. Springer, 2020, pp. 15–24.

[14] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artif. Intell. Med.*, vol. 94, pp. 42–53, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718304846

[15] F. Silva-Aravena, H. Núñez Delafuente, J. H. Gutiérrez-Bahamondes, and J. Morales, "A hybrid algorithm of ML and XAI to prevent breast cancer: A strategy to support decision making," *Cancers*, vol. 15, no. 9, p. 2443, Apr. 2023. [Online]. Available: https://www.mdpi.com/2072-6694/15/9/2443

[16] J. Marques-Silva, "Logic-based explainability in machine learning," in *Proc. 18th Int. Summer School Reasoning Web. Causality, Explanations Declarative Knowl.*, Berlin, Germany. Springer, Sep. 2022, pp. 24–104.

[17] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, Jan. 2022.

[18] F.-L. Fan, J. Xiong, M. Li, and G. Wang, "On interpretability of artificial neural networks: A survey," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 6, pp. 741–760, Nov. 2021.

[19] L. Fan, C. Liu, Y. Zhou, T. Zhang, and Q. Yang, "Interpreting and evaluating black box models in a customizable way," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 5435–5440.

[20] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.

[21] M. Nauman, N. Akhtar, O. H. Alhazmi, M. Hameed, H. Ullah, and N. Khan, "Improving the correctness of medical diagnostics (Don't short) based on machine learning with coloured Petri nets," *IEEE Access*, vol. 9, pp. 143434–143447, 2021.

[22] S. Bassan and G. Katz, "Towards formal XAI: Formally approximate minimal explanations of neural networks," in *Proc. Int. Conf. Tools Algorithms Construct. Anal. Syst.*, S. Sankaranarayanan and N. Sharygina, Eds. Cham, Switzerland: Springer, 1007, pp. 187–207.

[23] X. Huang, "Recent advances in formal explainability," Ph.D. dissertation, Université Paul Sabatier-Toulouse III, Toulouse, Toulouse, 2023.

[24] G. J. Holzmann, "The theory and practice of a formal method: NewCoRe," in *Proc. IFIP World Comput. Congr.*, 1994.

[25] T. Lecomte, D. Deharbe, E. Prun, and E. Mottin, "Applying a formal method in industry: A 25-year trajectory," in *Proc. Brazilian Symp. Formal Methods*. Springer, 2017, pp. 70–87.

[26] C. Pecheur, "Verification and validation of autonomy software at NASA," NASA, Washington, DC, USA, Rep., 2000, p. 20.

[27] F. Mayr, S. Yovine, and R. Visca, "Property checking with interpretable error characterization for recurrent neural networks," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 1, pp. 205–227, Feb. 2021.

[28] A. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *Proc. 23rd Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 108, 2020, pp. 895–905.

[29] S. Tripakis, "Data-driven and model-based design," in *Proc. IEEE Ind. Cyber-Physical Syst. (ICPS)*, May 2018, pp. 103–108.

[30] K. Jensen, *Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use*, vol. 1. Springer, 2013.

[31] C. A. Petri, "Kommunikation mit automaten," Ph.D. thesis, Univ. Bonn, Bonn, Germany, 1962.

[32] A. V. Ratzer, L. Wells, H. M. Lassen, M. Laursen, J. F. Qvortrup, M. S. Stissing, M. Westergaard, S. Christensen, and K. Jensen, "CPN tools for editing, simulating, and analysing coloured Petri nets," in *Proc. Int. Conf. Appl. Theory Petri Nets*. Springer, 2003, pp. 450–462.

[33] M. Westergaard and L. M. Kristensen, "The access/CPN framework: A tool for interacting with the CPN tools simulator," in *Proc. Int. Conf. Appl. Theory Petri Nets*. Springer, 2009, pp. 313–322.

[34] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.

[35] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2131–2145, Sep. 2019.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[37] K. Mridha, Md. M. Uddin, J. Shin, S. Khadka, and M. F. Mridha, "An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system," *IEEE Access*, vol. 11, pp. 41003–41018, 2023.

[38] M. Pocevičiūtė, G. Eilertsen, and C. Lundström, "Survey of XAI in digital pathology," in *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*, 2020, pp. 56–88.

[39] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, K. Balananthanan, M. Arunkumar, and C. Rajkumar, "A prediction and recommendation system for diabetes mellitus using XAI-based lime explainer," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Apr. 2022, pp. 1472–1478.

[40] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102470.

[41] R. J. Haynes and R. Naidu, "Influence of lime, fertilizer and manure applications on soil organic matter content and soil physical conditions: A review," *Nutrient Cycling Agroecosystems*, vol. 51, pp. 123–137, Jun. 1998.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1.

[43] T. Peltola, "Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback–Leibler projections," 2018, *arXiv:1810.02678*.

[44] I. Ahern, A. Noack, L. Guzman-Nateras, D. Dou, B. Li, and J. Huan, "NormLime: A new feature importance metric for explaining deep neural networks," 2019, *arXiv:1909.04200*.

[45] Y.-H. Hung and C.-Y. Lee, "BMB-LIME: LIME with modeling local nonlinearity and uncertainty in explainability," *Knowl.-Based Syst.*, vol. 294, Jun. 2024, Art. no. 111732. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124003678

[46] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, 2015.

[47] P. J. Lisboa, *Industrial Use of Safety-Related Artificial Neural Networks*, document HSE CR 327/2001, 2001.

[48] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, Art. no. e00938.

[49] A. H. Elsheikh, S. W. Sharshir, M. Abd Elaziz, A. E. Kabeel, W. Guilan, and Z. Haiou, "Modeling of solar energy systems using artificial neural network: A comprehensive review," *Sol. Energy*, vol. 180, pp. 622–639, Mar. 2019.

[50] P. M. Atkinson and A. R. L. Tatnall, "Introduction neural networks in remote sensing," *Int. J. Remote Sens.*, vol. 18, no. 4, pp. 699–709, Mar. 1997.

[51] L. Wells, "Performance analysis using CPN tools," in *Proc. 1st Int. Conf. Perform. Eval. Methodolgies Tools*, 2006, p. 59.

[52] R. Milner, M. Tofte, R. Harper, and D. MacQueen, *The Definition of Standard ML: Revised*. Cambridge, MA, USA: MIT Press, 1997.

[53] T. Murata, "Petri nets: Properties, analysis and applications," *Proc. IEEE*, vol. 77, no. 4, pp. 541–580, Apr. 1989.

[54] E. M. Clarke and E. A. Emerson, "Design and synthesis of synchronization skeletons using branching time temporal logic," in *Proc. Workshop Log. Programs*. Springer, 1981, pp. 52–71.

[55] C. Baier and J.-P. Katoen, *Principles of Model Checking*. Cambridge, MA, USA: MIT Press, 2008.

[56] O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[57] F. D. Pereira, S. C. Fonseca, E. H. T. Oliveira, A. I. Cristea, H. Bellhäuser, L. Rodrigues, D. B. F. Oliveira, S. Isotani, and L. S. G. Carvalho, "Explaining individual and collective programming students' behavior by interpreting a black-box predictive model," *IEEE Access*, vol. 9, pp. 117097–117119, 2021.

[58] T. Khater, A. Hussain, R. Bendardaf, I. M. Talaat, H. Tawfik, S. Ansari, and S. Mahmoud, "An explainable artificial intelligence model for the classification of breast cancer," *IEEE Access*, 2024.

[59] M. B. Gorzałczany and F. Rudziński, "Interpretable and accurate medical data classification—A multi-objective genetic-fuzzy optimization approach," *Expert Syst. Appl.*, vol. 71, pp. 26–39, Apr. 2017.

[60] T. A. Etchells and P. J. G. Lisboa, "Orthogonal search-based rule extraction (OSRE) for trained neural networks: A practical and efficient approach," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 374–384, Mar. 2006.

[61] A. Gupta, S. Park, and S. M. Lam, "Generalized analytic rule extraction for feedforward neural networks," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 6, pp. 985–991, Dec. 1999.

[62] R. Setiono and H. Liu, "Symbolic representation of neural networks," *Computer*, vol. 29, no. 3, pp. 71–77, Mar. 1996.

[63] R. Setiono, "Extracting rules from neural networks by pruning and hidden-unit splitting," *Neural Comput.*, vol. 9, no. 1, pp. 205–225, Jan. 1997.

[64] S. S. Roy, A. Mallik, R. Gulati, M. S. Obaidat, and P. V. Krishna, "A deep learning based artificial neural network approach for intrusion detection," in *Proc. Int. Conf. Math. Comput.* Springer, 2017, pp. 44–53.

[65] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103141.

[66] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," *Measurement*, vol. 146, pp. 557–570, Nov. 2019.

[67] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105941.

[68] S. Jia, P. Lin, Z. Li, J. Zhang, and S. Liu, "Visualizing surrogate decision trees of convolutional neural networks," *J. Visualizat.*, vol. 23, no. 1, pp. 141–156, Feb. 2020.

[69] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing design practices for explainable AI user experiences," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2020, pp. 1–15.

[70] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 648–657.

[71] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.

[72] O. Elsarrar, M. Darrah, and R. Devine, "Analysis of forest fire data using neural network rule extraction with human understandable rules," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1917–19176.

[73] J. Sultana, M. U. Rani, and M. A. H. Farquad, "Knowledge discovery from recommender systems using deep learning," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 1074–1078.

[74] T. Ueno and Q. Zhao, "Interpretation of deep neural networks based on decision trees," in *Proc. IEEE 16th Int. Conf. Dependable, Autonomic Secure Comput., 16th Int. Conf. Pervasive Intell. Comput., 4th Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 256–261.

[75] M. G. Augasta and T. Kathirvalavakumar, "Reverse engineering the neural networks for rule extraction in classification problems," *Neural Process. Lett.*, vol. 35, no. 2, pp. 131–150, Apr. 2012.

[76] M. Sato and H. Tsukimoto, "Rule extraction from neural networks via decision tree induction," in *Proc. Int. Joint Conf. Neural Networks*, Jul. 2001, pp. 1870–1875.

[77] G. P. J. Schmitz, C. Aldrich, and F. S. Gouws, "ANN-DT: An algorithm for extraction of decision trees from artificial neural networks," *IEEE Trans. Neural Netw.*, vol. 10, no. 6, pp. 1392–1401, Nov. 1999.

[78] D. T. Pham and M. S. Aksoy, "RULES: A simple rule extraction system," *Expert Syst. Appl.*, vol. 8, no. 1, pp. 59–65, Jan. 1995.

[79] A. K. Arslan, T. Zeynep, İ. B. Çiçek, and C. Çolak, "A novel interpretable web-based tool on the associative classification methods: An application on breast cancer dataset," *J. Cogn. Syst.*, vol. 5, no. 1, pp. 33–40, 2020.

[80] J. Marques-Silva, "Disproving XAI myths with formal methods—Initial results," in *Proc. 27th Int. Conf. Eng. Complex Comput. Syst. (ICECCS)*, Jun. 2023, pp. 12–21.

[81] L. Antonio, P. G. Luca, M. STRIANI, Z. Stefano, and D. Rossana, "Formal methods meet XAI: The tool DEGARI 2.0 for social inclusion," in *Proc. 4th Workshop Artif. Intell. Formal Verification, Logic, Automata, Synthesis*, 2022.

[82] K. Bjørner, S. Judson, F. Cano, D. Goldman, N. Shoemaker, R. Piskac, and B. Könighofer, "Formal XAI via syntax-guided synthesis," in *Proc. Int. Conf. Bridging Gap Between AI Reality*. Springer, 2023, pp. 119–137.

[83] K. Larsen, A. Legay, G. Nolte, M. Schlüter, M. Stoelinga, and B. Steffen, "Formal methods meet machine learning (F3ML)," in *Proc. Int. Symp. Leveraging Appl. Formal Methods*. Springer, 2022, pp. 393–405.

[84] J. Marques-Silva and A. Ignatiev, "Delivering trustworthy AI through formal XAI," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 11, pp. 12342–12350.

[85] S. Bassan, G. Amir, D. Corsi, I. Refaeli, and G. Katz, "Formally explaining neural networks within reactive systems," in *Proc. Formal Methods Comput.-Aided Design (FMCAD)*, Oct. 2023, pp. 1–13.

[86] X. Huang and J. Marques-Silva, "On the failings of Shapley values for explainability," *Int. J. Approx. Reasoning*, vol. 171, Aug. 2024, Art. no. 109112.
[87] B. Zhang, A. Vakanski, and M. Xian, "BI-RADS-NET-v2: A composite multi-task neural network for computer-aided diagnosis of breast cancer in ultrasound images with semantic and quantitative explanations," *IEEE Access*, vol. 11, pp. 79480–79494, 2023.
[88] S. Rajpal, A. Rajpal, A. Saggar, A. K. Vaid, V. Kumar, M. Agarwal, and N. Kumar, "XAI-MethylMarker: Explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data," *Expert Syst. Appl.*, vol. 225, Sep. 2023, Art. no. 120130. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423006322
[89] S. Hurtado, H. Nematzadeh, J. García-Nieto, M.-Á. Berciano-Guerrero, and I. Navas-Delgado, "On the use of explainable artificial intelligence for the differential diagnosis of pigmented skin lesions," in *Proc. Int. Work-Conf. Bioinf. Biomed. Eng.* Springer, 2022, pp. 319–329.
[90] D. Dua and C. Graff. (2022). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml
[91] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, Oct. 2012.
[92] M. Nauman, N. Akhtar, A. Alhudhaif, and A. Alothaim, "Guaranteeing correctness of machine learning based decision making at higher educational institutions," *IEEE Access*, vol. 9, pp. 92864–92880, 2021.
[93] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.
[94] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, "Weka—A machine learning workbench for data mining," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2009, pp. 1269–1277.
[95] K. Jensen, C. Søren, L. M. Kristensen, and M. Westergaard. (2022). *Cpntools Home*. [Online]. Available: https://www.cs.au.dk/CPNTools
[96] M. K. Islam, M. M. Rahman, M. S. Ali, S. M. Mahim, and M. S. Miah, "Enhancing lung abnormalities detection and classification using a deep convolutional neural network and GRU with explainable AI: A promising approach for accurate diagnosis," *Mach. Learn. Appl.*, vol. 14, Dec. 2023, Art. no. 100492.
[97] L. Gianfagna and A. Di Cecco, *Explainable AI With Python*. Springer, 2021.
[98] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

NADIA KHAN is currently pursuing the Ph.D. degree in computer science with The Islamia University of Bahawalpur (IUB), Pakistan. Her research interests include machine learning, artificial neural networks, explainable artificial intelligence, software engineering, data mining, medical diagnostics, and decision trees.

MUHAMMAD NAUMAN received the Ph.D. degree in computer science from The Islamia University of Bahawalpur (IUB), Pakistan. He is currently a Lecturer with IUB. His research interests include formal approaches, formal methods, big data, big data analytics, machine learning, data mining, explainable artificial intelligence, artificial neural networks, decision trees, and predictive models.

AHMAD S. ALMADHOR received the B.S.E. degree in computer science from Aljouf University (formerly Aljouf College), Aljouf, Saudi Arabia, in 2005, the M.E. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2010, and the Ph.D. degree in electrical and computer engineering from the University of Denver, Denver, CO, USA, in 2019. From 2006 to 2008, he was a Teaching Assistant and the College of Sciences Manager with Aljouf University, where he was a Lecturer, from 2011 to 2012. He was a Senior Graduate Assistant and a Tutor Advisor with the University of Denver, from 2013 to 2019. He is currently an Assistant Professor of CEN and VD with the Computer and Information Sciences College, Jouf University, Saudi Arabia. His research interests include AI, blockchain, networks, smart and microgrid cyber security and integration, image processing, video surveillance systems, PV, EV, and machine and deep learning. His awards and honors include the Aljouf University Scholarship (Royal Embassy of Saudi Arabia in D.C.), Aljouf's Governor Award for Excellency, and several others.

NADEEM AKHTAR received the M.S. degree in information system architecture from the Institut Universitaire Professionnalis (IUP), University of South Brittany, Bretagne, France, in 2006, and the Ph.D. degree (magna cum laude) from IRISA, University of South Brittany (UBS), Bretagne, in September 2010. He is currently the Chairperson and an Associate Professor with the Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan. He has 20 years of experience in teaching and research at universities. His Ph.D. thesis was titled "Contribution to the Formal Specification and Verification of a Multi-Agent Robotic System." He is supervising the Ph.D. (CS) and the M.S. (CS) research students. His contributions to scientific research are reflected in the form of 55 research articles published in reputed international journals. His research interests include formal verification and validation, formal modeling, safety-critical systems, data analytics, and machine learning. He was a recipient of several awards, scholarships, and research grants, such as the 2004 French Embassy Scholarship for M.S. (computer science) studies in France, the 2006 Higher Education Commission (HEC) Overseas Scholarship for Ph.D. studies in France, the Teaching Assistant for ENSIBS—UBS France, the HEC Start-Up Research Grant of 0.5 million, in 2012, and the Student Research Project Grant from ICT, in 2014.

ABDULLAH ALGHURIED received the Ph.D. degree in industrial engineering from the University of Miami, in 2020. He is currently an Assistant Professor of industrial engineering with the University of Tabuk. His research interests include stochastic modeling and optimization, data mining, big data, data analytics, decision-making under uncertainty, lean six sigma, and sustainability.

ADI ALHUDHAIF received the bachelor's degree in computer science from King Saud University, the master's degree in computer science (information security and big data) from George Washington University, Washington D.C., the master's degree in law LLM in internet law from the University of Strathclyde, Glasgow, U.K., and the Ph.D. degree in computer science (information security and big data) from George Washington University. He is currently an Associate Professor with Prince Sattam Bin Abdulaziz University. He received several IT professional certificates in IT governance, risk management, project management, and more.