Knowledge-Based Systems 271 (2023) 110552

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

An optimized Belief-Rule-Based (BRB) approach to ensure the trustworthiness of interpreted time-series decisions

Sonia Farhana Nimmy [a,*], Omar K. Hussain [a], Ripon K. Chakrabortty [b], Farookh Khadeer Hussain [c], Morteza Saberi [c]

[a] School of Business, University of New South Wales, Canberra, Australia
[b] School of Engineering and Information Technology, University of New South Wales, Canberra, Australia
[c] School of Computer Science, University of Technology Sydney, Sydney, Australia

ARTICLE INFO

Article history:
Received 17 August 2022
Received in revised form 26 October 2022
Accepted 4 April 2023
Available online 17 April 2023

Keywords:
Explainable AI
Credibility
Glass-box
Post-hoc explainers
Trustworthiness

ABSTRACT

The accuracy and reliability of XAI methods are important to establish their credibility and use in complex decision-making tasks. Existing XAI methods provide little information about the correctness and reliability of their outputs. Furthermore, post-hoc explanation approaches explain the outcomes after producing them, not in a step-by-step glass-box manner to explain how an output is reached. Our proposed approach addresses these drawbacks by designing a Belief-Rule-Based (BRB) framework that interprets in a glass-box manner why a particular decision has been reached. It does that by determining the chance of different output classes occurring for a specific time period by considering the different possible permutations of the inputs along with their influence. This also assists the user to determine if the given input dataset is incomplete, vague, imprecise or inconsistent before trusting the analysis emanating from it. We compare the performance of the proposed BRB approach against the different eXplainable artificial intelligence (XAI) methods, such as SHAP, LIME and LINDA-BN to ensure the users of the trustworthiness of its analysis. This also enables users to determine the extent to which each of the XAI techniques meets the requirements of XAI and the gaps that need to be addressed.

1. Motivation of the article

The explainability of ML models is becoming an increasingly important research question. This is especially important in domains such as risk management, health diagnosis, high-stake businesses, financial analysis etc., where the accuracy, reliability, and stability of the recommended decision outputs is needed to instil confidence [1,2]. As a result, the number of eXplainable artificial intelligence (XAI) methods that interpret why a black-box model has reached a certain conclusion has significantly increased [3]. LIME [4], SHAP [5], MAPLE [6], Anchors [7], LINDA-BN [8] are examples of such techniques that work by interpreting local approximations to the predicted output. These methods have been applied to detect false news [9], plan treatments after diagnosing diseases [10–12], credit risk analysis and mitigation [13,14], road and traffic management and control [15,16], emergency help systems and services [17,18], safety and security systems [19,20], disability detection and support services [21,22], child care and early education services [23,24], military and defence protocols and security [25,26] etc. While these methods provide beneficial analyses, they fail in ensuring the accuracy, reliability, and stability of a given interpretation in scenarios when the decision output being interpreted for a time period is dependent on what happens in previous time periods. To explain this further, we consider that an ML model's decision output can be classified in one of two broad categories, namely *static* and *progressive*. A static decision output is one in which the input features $\{f_1, f_2....f_n\}$ that influence the decision output {d} for a time slot $\{t_z\}$ are given manually by the decision maker at the beginning of the same time slot $\{t_z\}$. Examples of a static decision output are fake news detection [9] or spam classification [27]. On the other hand, a progressive decision output is one in which the inputs $\{f_1, f_2....f_n\}$ to the ML model are given at $\{t_z\}$ for it to recommend a {d} at a future time slot $\{t_{z+n}\}$. In such cases, it is up to the ML model to first determine how each of the $\{f_1, f_2....f_n\}$ evolve from $\{t_z\}$ to $\{t_{z+n}\}$ before determining the {d} at $\{t_{z+n}\}$. Examples of a progressive decision output are risk assessment [28] or a financial position [29] at a future time period.

Current XAI models are local explanation methods with LIME [30], SHAP [31], Anchors [32], BayesLIME [33], and BayesSHAP

* Corresponding author.
E-mail addresses: s.nimmy@adfa.edu.au (S.F. Nimmy), o.hussain@adfa.edu.au (O.K. Hussain), r.chakrabortty@adfa.edu.au (R.K. Chakrabortty), farookh.hussain@uts.edu.au (F.K. Hussain), morteza.saberi@uts.edu.au (M. Saberi).

https://doi.org/10.1016/j.knosys.2023.110552

[34] leveraging perturbations of individual instances to construct interpretable local approximations whereas LINDA-BN using conditional probabilities. This means that the accuracy, reliability, and stability of their interpretations will be high and therefore confidence will be high when the ML model's decision output is static. This is because both the inputs and the output relate to the same time slot and the inputs are given by the experts, therefore they are confirmed as being current. However, when the ML model's decision is progressive, XAI models first need to consider how $\{f_1, f_2...f_n\}$ evolve from $\{t_z\}$ to $\{t_{z+n}\}$ before interpreting the {d} at $\{t_{z+n}\}$. In such scenarios, slight variations to the values of $\{f_1, f_2...f_n\}$ in a time slot will result in a different output at $\{t_{z+n}\}$ and thus a different explanation. Therefore, the explanations given in a progressive type of output may be unstable [35] as different runs on the same dataset but with slightly changed input parameter settings can yield wildly divergent results [36] bringing the trustworthiness of the interpretations comes into question. To address the aforementioned gaps so that XAI approaches can be applied on progressive types of decision outputs, in this paper, we explain our proposed approach which:

1. uses causal links to model how the values of the features in a progressive decision evolve over time.
2. uses a Belief-Rule-Based (BRB) approach to determine which input features from the given set are strongly correlated to the different types of decision output classes. This analysis assists the decision maker to confirm if the interpretation given by an XAI method in a time slot is consistent with the expected output.
3. compares the output of the proposed approach with three existing XAI approaches from the literature (SHAP, LIME, and LINDA-BN). We then use this analysis to determine how each of the XAI approaches meets the XAI requirements discussed in the literature.

The remainder of the paper is structured as follows: Section 2 presents the related work from the literature. Section 3 details the proposed architecture and explains the different modules. This section also presents the results and explains them in a step-by-step interpretable way. It also graphically represents the optimized output showing the connected features with the output. Section 4 compares the proposed BRB approach to an improved version of LINDA-BN. This section reports on enhancements to LINDA-BN to enable it to consider changes in the feature values over time. Section 5 compares the output of the proposed BRB approach against other XAI approaches in the XAI requirements discussed in the literature. Finally, Section 6 summarizes the research findings and concludes the paper with a discussion on future work.

2. Related works

Core AI models like deep neural networks and ensemble models among other mathematical and statistical classifiers have been applied as XAI approaches because of their universality [37,38]. For example, regression is used in LIME [39], whereas the Shapley value is employed in SHAP [40]. LINDA-BN uses a game theory concept, Bayesian network and Markov model [8]. These models work on a post-hoc basis and explain the reasoning behind a certain decision by including visual aids. Because of their usability, they have been applied in different domains. For example, Alharbi et al. [9] interpret false news identification models to identify which significant aspects contribute to the prediction of a model from an explainable machine learning perspective using Captum, LIME, and SHAP. This sheds light on how detection models work and the extent to which they can be relied upon. However, a primary drawback of this system is that post-hoc explainers like LIME and SHAP, despite giving an output, do not explain how trustworthy their outputs are. Tree-LIME, a model-agnostic method [41] is a revised LIME technique established on local interpretation by applying decision tree regression. To illustrate the significance of fidelity in the regression explanation problem, mean absolute error (MAE) is used. The methodology can improve the fidelity of the interpreter which provides more authentic reasoning for explicit events and delivers a better visual presentation of the tree structure in real supply chain forecasting applications. A drawback of the model is that due to flat features classification it is unable to prioritize the features with similar values. To resolve this drawback, designing hierarchical representations that can prioritize (rank) the features, model the interconnectedness between them and explain the output in a glass-box way may be a good choice. Szczepa et al. [27] developed an innovative explainable method to better understand false information detection. LIME and Anchors, which are two XAI strategies to explain fake information, were deployed and assessed on fake tweets or headline data. A drawback with Anchors as an explainer is that it is not consistently capable of giving an interpretation. LIME is a post-hoc explainer and in a dynamic platform like Twitter or online news, it may not highlight meaningful patterns which brings into question the trustworthiness and accuracy of the decision.

Matin et al. [17] proposed an earthquake-induced building-damage map using a multilayer perceptron (MLP) and SHAP. A single after-occurrence satellite image was used as the input of MLP to classify buildings as collapsed or non-collapsed. SHAP was used to interpret the effect of the components on the model outcome. To design such a system, it is essential to ensure a time-series dataset evaluation technique that can connect the outcomes of all of the time sequences and describe the condition of the building, the effect of the earthquake, the damage sustained and how it will propagate over time. However, as discussed in Section 1, for progressive-based decisions, MLP and SHAP are not appropriate. Petsis et al. [18] used an expert system and XAI to predict and analyse the Emergency Department's (ED) visitors. As an AI technique, the XGBoost algorithm which uses data from patient visits, time-based data, dates of holidays, special events, and weather data was used to anticipate the frequency of ED visits. SHAP was applied to explain the approach's output. However, the problem's inputs are data-driven and strongly connected. This means that a minor change in a single feature can affect the entire assessment. To ensure the assessment is trustworthy, there is a need to design and assess the dependency of the dataset and use this in further analysis. Northcote et al. [21] designed an Alzheimer's disease (AD) patient investigation employing gene expression and an image dataset. CNN and SpinalNet techniques predicted the AD categories from an MRI which is an image of a brain scan. k-nearest neighbours, support vector machines and XGBoost classifiers were used to classify the AD categories from the gene analysis dataset. Finally, LIME was used to obtain a better understanding of the responsible AD genes. However, gene expression data has a specific pattern and is interconnected. To capture these interconnections, an explainer that can capture them (for example, LINDA-BN) is better than those that cannot (for example, LIME). Adak et al. [42] analysed emotions using simple and hybrid deep learning (DL) procedures in the food delivery service discipline and explained the forecasts by employing LIME and SHAP. The training and testing procedure was undertaken on the client feedback dataset. The drawback of this study is that it can misunderstand consumer feedback resulting in incorrect decisions. A possible resolution to this issue can be using a glass-box explainer instead of post-hoc explainers so it explains what the input was, how it was processed, and why the output was produced. Areti et al. [43] employed machine

learning, namely the XGBoost algorithm and XAI, namely SHAP, to predict house prices using open government data. The XGBoost algorithm was used to create the predictive model and SHAP was used to explain the model's decisions. House price criteria depend on many other domain criteria and change if the connected domain criteria change the impact that they will have on others. If these dependencies are not captured correctly, the decision will be biased and will not be trustworthy to the users.

So, from the above summary, while it can be seen that XAI approaches have been applied in different domains, they have limitations as to what type of datasets they can be best applied. For example, while the XAI approaches work well on static and progressive datasets, they are limited in terms of capturing the dynamic nature of progressive datasets. This means that even though they give an output, the trustworthiness of these outputs is not guaranteed. This needs to be addressed to increase users' confidence of using XAI analysis widely in different automated systems. In the next section, we propose a BRB approach that attempts to interpret why a decision has been reached in a glass-box manner. This analysis also assists users to increase their trustworthiness in the generated output.

3. An optimized BRB approach to ensure the trustworthiness of interpreted outcomes

Fig. 1 shows our proposed BRB framework to ensure the trustworthiness of the interpreted decisions. The framework is built on three modules, namely *Knowledge Graph Module (KGM), Knowledge Propagation Module (KPM)*, and *Feature Evaluation Module (FEM)*. KGM is a semi-automated module that displays a graphical representation between the input features that lead to the decision output in the form of a knowledge graph (KG). KPM propagates and updates how the values of the features evolve over time based on the interconnections modelled by KGM. KPM also considers how changes in the value of each feature in a time period impact the other interdependent features and the resultant output class. FEM uses BRB to ascertain the output decision class of a time slot and determine the consistent features leading to that output class before graphically visualizing them. In the next sections, we further explain the working of each module.

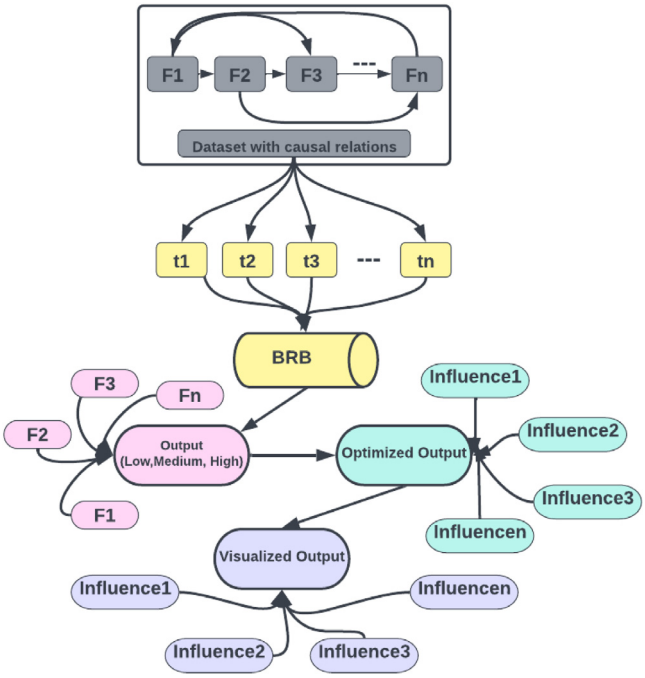


Fig. 1. Working of the BRB approach to interpret the logic of the decision outputs in a glass-box manner.

3.1. Knowledge Graph Module (KGM)

KGM creates a KG based on the causal relationship between the input features. The primary goal of a KG is to find out how different features are related and how the interconnection between them affects the system. Any two features will have a causal connection if due to a change in the first feature, the second feature also changes [44]. In this case, the first feature is called the cause for the change and the second feature is called the effect of the change. To have a cause-and-effect connection, the link should be credible and non-spurious, the cause feature must occur before the affected feature is impacted, and the affected feature is updated after the cause feature changes [45]. Analyzing the causes and their consequences is the goal of causal analysis. This technique reveals the facts that lead to a specific situation which is different to merely focusing on reasoning the symptoms that lead to the cause [46]. KGM collects knowledge about the features and uses this to determine the interconnections between them. After establishing the connections between the features, it analyzes the relationships to determine the effects of the links on the interlinked features.

To explain with an example, from here on we consider the domain of asset maintenance where an asset manager wants to determine the chances of an asset failing over a time period. We consider *time space* as the total time period over which the analysis is done (for example, six months or one year). The time space is divided into different non-overlapping periods of time (for example, one month) known as the *time slots*. In each time slot, the objective of the asset manager is to determine the risk of the asset failing in the following four output decision classes, *Safe, Low Risk*, *Medium Risk* and *High Risk*. This decision is influenced by 23 input features that are given at the beginning of the time space. KGM determines the causal links between the features using the knowledge of the relationships between them as shown in Fig. 2. Based on this causal relationship, Fig. 3 shows how the feature "Efficiency" is connected with other features. Fig. 4 shows the design of a KG based on the causal links between all the features.

3.2. Knowledge Propagation Module (KPM)

KPM propagates and updates how the values of the features evolve in the different time slots of the time space based on the interconnections modelled by KGM. The first step of this module is to quantify the lower and upper bound ranges of the input features. This is done by studying the knowledge and information about the linked features, as shown in Fig. 2. This process identifies the metrics and the range over which each input feature spans. This leads to the second step in which the determined range of the input features is divided into the corresponding output classes. For example, if we consider five input features, namely *pressure, friction, crushing, shearing and entanglement*, as shown in Table 1, this step represents the range to which each feature corresponds to each output class. In the third step of this module, the value of each input feature in each time slot of the time space is determined. This information is used further in the next module.

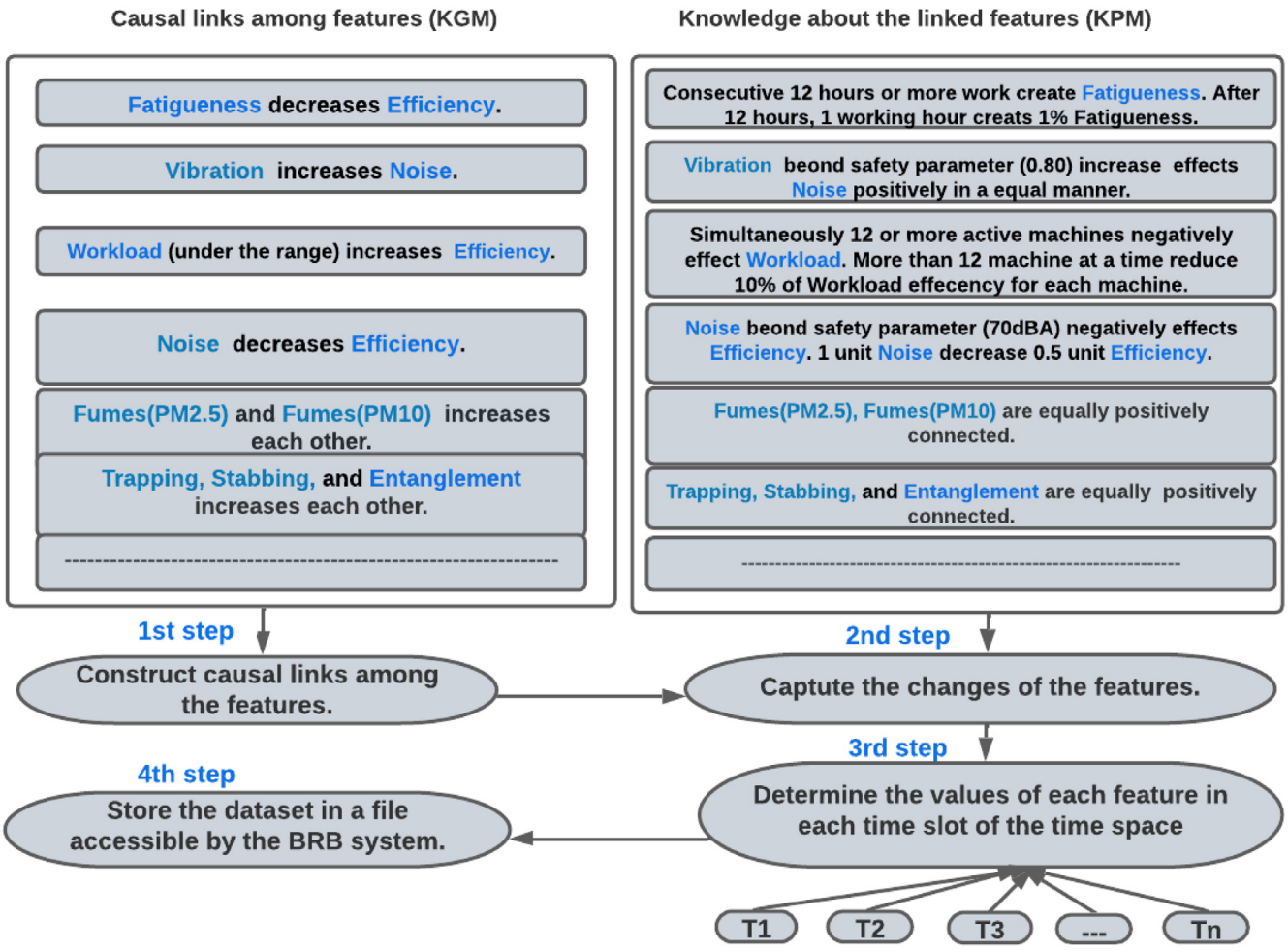


Fig. 2. Capturing the features and the interconnectedness between them.

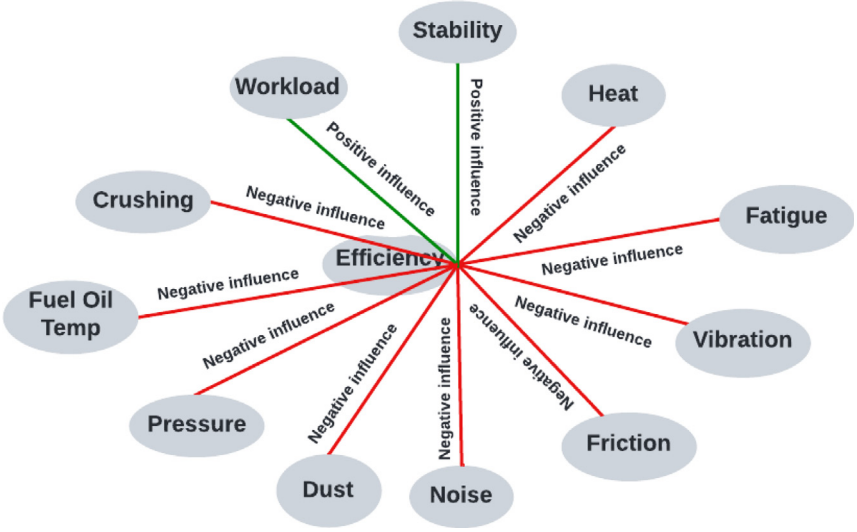


Fig. 3. Connection among the features using causal relationship.

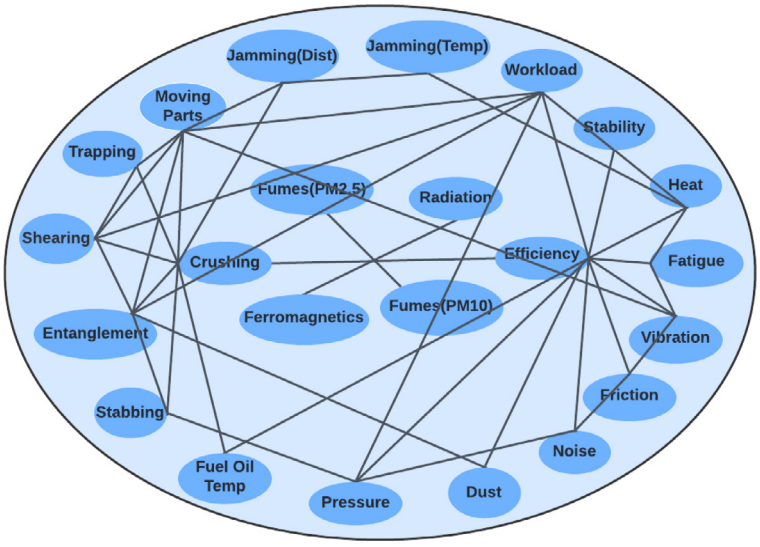


Fig. 4. Relationship between features that contribute to the output class.

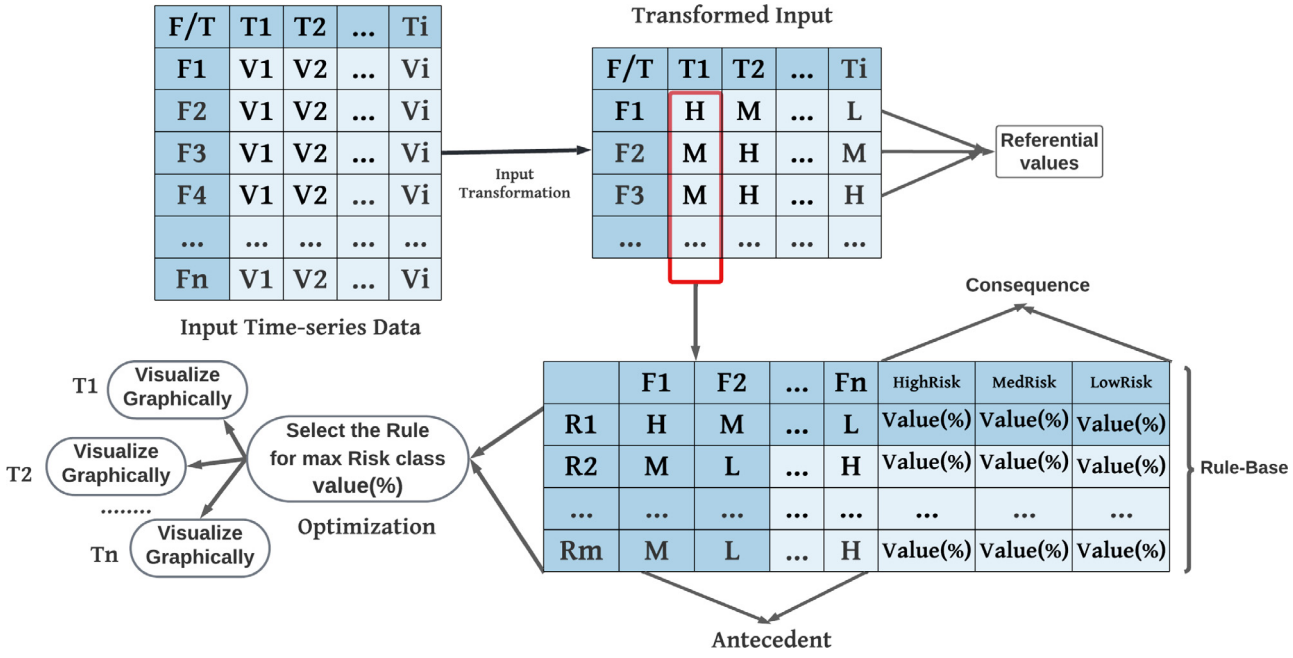


Fig. 5. Steps of a BRB Network.

Table 1
Feature value distribution among the time slots.

Key features/output risk classes	Pressure	Friction	Crushing	Shearing	Entanglement
Safe	29.6 to 30.2	1 to 70	6 to 5	8 to 10	30 to 20
Low risk	30.3 to 30.6	71 to 75	4.9 to 4	11 to 15	19 to 15
Medium risk	30.7 to 31.2	76 to 84	3.9 to 2	16 to 24	14 to 10
High risk	31.3 to 32	85 to 99	1.9 to 1	25 to 30	09 to 01

3.3. Feature evaluation module (FEM)

FEM builds on the analysis from KGM and KPM using BRB to ascertain the output decision class of a time slot along with determining the features leading to that output class. Fig. 5 shows the steps of the BRB algorithm. The following are the steps taken by FEM to achieve this aim:

- **Step 1: Transform the values of the input features to referential values in the different time slots**

KPM determines the value of each input feature in each time slot of the time space. In this step, these values are transformed to referential values. Referential values are fuzzy linguistic values that show the degree of influence of a feature on the output risk class in that time slot. To determine the referential value of a feature, the first task is to determine the linguistic terms across

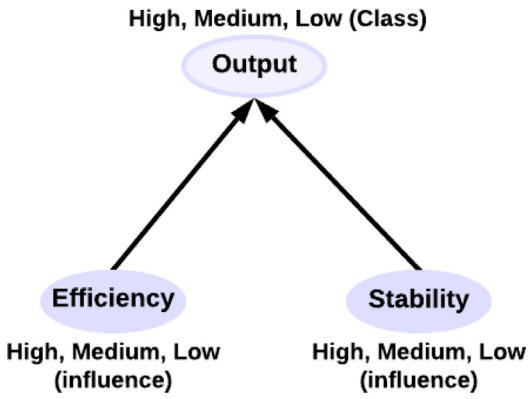


Fig. 6. Sample BRB Network.

which the influence of an input feature should be mapped. In our asset management example discussed in Section 3.1, we consider that the influence of each input feature is mapped across the three linguistic terms of 'low', 'medium' and 'high', as shown in Fig. 6. Once these are determined, the next task is to determine which range of the input features should map to each linguistic term. Continuing with our example, in this task, we determine that for the input feature *efficiency*, if the input is ≥50% and ≤74%, then it maps to the 'medium' referential value. Similarly, if the input value is ≥1% and ≤49%, then it maps to the 'low' referential value whereas if it is between ≥75% and ≤100%, it maps to the 'high' referential value. In the third task, the midpoint of each linguistic range is determined using Eq. (1):

$$Mid = \frac{H + L}{2} \tag{1}$$

where, H is the high range value and L is low range value. This is to determine the influence of each input feature to the output risk class in that time slot. So, for the input feature *efficiency*, where, L = 1, H = 100, Mid = $\frac{100+1}{2} = 50.5$ by applying Eq. (1)

- **Step 2: Determine the influence of each input feature to the output risk class in that time slot**

In this step, the influence of each input feature on the output risk class in that time slot is determined. The aim of this step is achieved in two tasks. In the first task, the influence or utility of the highest and the least contributing input feature is determined. This is determined by considering those features that have the highest and the lowest values for the fuzzy linguistic terms of 'high' and 'low', respectively. To calculate the highest and the least utility for a particular time slot, Eq. (2) is used

$$D_n = a + \sum X_n b_n \tag{2}$$

where $n = 1$ *to* K (K is the number of features). Here D_n is regarded as the nth preference value (utility factor). X_n is the nth referential value of the antecedent X, b_n is the subsequent attribute weight of the antecedent X, and a is an arbitrary value most of the time neglected as zero. For example, let us consider that a time slot has two input features, namely *efficiency* and *stability* (Both have feature values in the range of ≥1 to ≤100). The highest influence of these features on the output class as determined by Eq. (2) is D1 = 0 + 100 ∗ 1 + 100 ∗ 1 = 200. Similarly, the lowest influence of these features on the output class as determined by Eq. (2) is D3 = 0 + 1 ∗ 1 + 1 ∗ 1 = 2. To measure the intermediate degree of belief (D2), Eq. (3) is used.

$$D_i = \frac{D_1 * m + D_k * n}{m + n} \tag{3}$$

where i is the number of utilities, m and n are the ratio. Continuing with the example, D2 with a ratio 1:1 using Eq. (3) is $D2 = \frac{200*1+2*1}{1+1} = 101$

- **Step 3: Calculating the degree of belief of each output class**

After determining the influence values of each input feature, Eq. (4) is used to ascertain the degree of belief of each output risk class using the RIMER methodology [47].

$$Y_p = a + \sum X_p^E b_E \tag{4}$$

where $E = 1, 2, 3, \ldots$ *and* (X_E^p is the referential value of Eth antecedent in pth rule). Here Y_p is regarded as the input value for pth rule. X_E^p is the referential value of Eth antecedent for pth rule and b_E is the subsequent attribute weight of the antecedent E, and a is an arbitrary value most of the time neglected as zero. As shown in Fig. 6, using Eq. (4) and RIMER [47] input transformation technique, IF efficiency is high and stability is high (with feature value 75 for both of them), then output is

Input $Y = 0 + 75 * 1 + 75 * 1 = 150$ [applying Eq. (4)]

Now, according to the rule of input transformation:

$$D_{2(transformed)} = \frac{Y - D_{2(Ref\,value)}}{Y - X_{(Ref\,value)}} \tag{5}$$

$D2 = \frac{150-101}{150-75} = 0.65$, [Applying Eq. (5)]

Again,

$$D_1 + D_2 = 1 \tag{6}$$

$D1 = 1 - 0.65 = 0.35$ [Applying Eq. (6)]

To continue with the asset management example discussed in Section 3.1, we consider that the time space over which the chances of an asset failing has to be determined is made up of 50 equal non-overlapping time slots ($t_1 - t_{50}$). The values of the 23 input features in each time slot is determined before applying FEM. The values of the features in t_1 are used as its starting point and BRB evaluates the values to define the influence on the occurrence of each output class in each time slot of the time space. Fig. 7 provides a snapshot of the BRB output in a specific time slot, showing that for a specific time slot, nPr different output set is possible where 'r' is the options (high, medium, low, safe) from a group of 'n' number of features (for 5 features). The formula to find nPr is n!/(n-r)! which represents the different combinations of output that will be generated [48]. From this, the rule which has the highest value for the risk classes is selected as the output for this time slot. For ($t_7 - t_{18}$), Table 2 shows the most probable output risk class occurring along with the features that are influencing its occurrence. From that table it can be seen that the most probable output class in t_7 - t_{11} is *low risk*, after which it moves to *medium risk* till t_{17} followed by *high risk* from t_{18} onwards.

After evaluating the most probable output in each time slot, the system graphically represents the output with the influence of each feature. Fig. 8 represents the *low risk*, *medium risk*, and *high risk* output as well as the underlying contributing features. Once the analysis for each time slot has been undertaken and the main features associated with a given output risk class are determined, Fig. 9 shows the most consistent features towards an output risk class. This analysis is determined by using the analysis of Table 2 which shows the common features influencing the output risk class. As the figure shows, ten features of the 23 consistently change their influence as the output class changes. These 10 features demonstrate consistent behaviour

Rule	Pressure	Friction	Crushing	Shearing	Entanglement	HighRisk	MedRisk	LowRisk
0	H	H	H	H	H	1	0	0
1	H	H	H	H	L	0.783018868	0	0.216981132
2	H	H	M	M	M	0.216981132	0.783018868	0
3	M	M	H	H	L	0.301886792	0.608490566	0.08962265
4	M	M	L	L	H	0.174528302	0.216981132	0.608490566
5	M	M	L	L	L	0	0.608490566	0.391509434
....	....	....	....	....	....	....	....	....

Fig. 7. Optimized output selection using BRB.

Table 2
Influence of the features on the most probable output risk class in each time slot.

T-slots	Output	Influence of the features on the most probable output risk class.
t7	LR	Jamming (Dist), Efficiency, Fumes (PM2.5), Stability, Fuel Oil, Fumes (PM10), Dust, Workload, Heat, Radiation, Stabbing, Fatigue, Crushing, Jamming (Temp), Pressure, Ferromagnetic, Noise.
t8	LR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Jamming (Dist), Fumes (PM2.5), Radiation, Ferromagnetic, Jamming (Temp).
t9	LR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Radiation, Jamming (Temp), Vibration, Entanglement, Shearing.
t10	LR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Moving Parts, Vibration, Entanglement, Fumes (PM2.5), Fuel Oil, Jamming (Temp).
t11	LR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Moving Parts, Vibration, Entanglement, Fuel Oil, Shearing.
t12	MR	Radiation, Stabbing, Stability, Fuel Oil, Crushing, Noise, Vibration, Entanglement, Heat, Workload, Moving Parts, Efficiency, Friction, Shearing, Fatigue, Dust, Pressure.
t13	MR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Jamming (Temp), Radiation, Entanglement, Jamming (Dist), Trapping.
t14	MR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Radiation, Jamming (Dist), Entanglement, Fuel Oil, Trapping, Jamming (Temp).
t15	MR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Vibration, Entanglement, Fuel Oil, Trapping, Moving parts.
t16	MR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Vibration, Jamming (Dist), Fuel Oil, Trapping, Jamming (Temp), Friction.
t17	MR	Pressure, Noise, Dust, Crushing, Stability, Fatigue, Stabbing, Workload, Heat, Efficiency, Friction, Shearing, Jamming (Dist), Entanglement, Vibration, Jamming (Temp).
t18	HR	Jamming (Dist), Stabbing, Stability, Trapping, Crushing, Noise, Vibration, Entanglement, Heat, Workload, Moving Parts, Efficiency, Friction, Shearing, Fatigue, Dust, Pressure.

with the corresponding output class. To manage the occurrence of an output risk class, this analysis can help determine which features are the most significant ones to the output risk class. This analysis can also be used by the risk manager to determine which features need to be managed to make the output risk class *safe* to minimize the risk.

4. Comparison of the output of the BRB approach with the modified version of LINDA-BN

To validate the trustworthiness of the interpreted decisions, we compare the output of our proposed BRB framework with the output of modified LINDA-BN on the same asset management dataset on which the BRB framework is applied. The aim is to compare the output risk class occurring in a time slot and then determine the trustworthiness of the explained features leading to that decision as given by these two different approaches. The reason for using a modified version of LINDA-BN as opposed to the standard LINDA-BN is that LINDA-BN cannot automatically process the time-series datasets of a dynamic system. So, to apply LINDA-BN in a time-series dataset, we used a KG and system dynamics (SD) as shown in Fig. 10 to capture the relationship between the features and determine how they evolve over time to apply it to LINDA-BN. This modified version of LINDA-BN considers all of the time slots of the time space and gives the output in the risk classes of *low*, *medium* and *high*. Fig. 11 shows the output of LINDA-BN for the output risk class of *low* along with the contributing features. Similar output is produced for the output risk classes of *medium* and *high* which we do not show in this paper due to space constraints. These contributing features to an output risk class for a time slot will be compared with the output of the BRB approach. A comparison of the approaches is as follows:

4.1. Ordering among the features has an impact on the output

The interpretable BRB approach produces an output based on the glass-box interpretation of the impact of the features leading to it whereas LINDA-BN explains the output after producing them. So, in LINDA-BN, the sequence in which the features are presented as input can have an impact on the output being shown in terms of their influence. For example, Fig. 12 shows that if the ordering of the features changes, their influence on the output class also changes. Before changing position, the influence of "stability" was 94.2%, and the influence of "workload" was 92.4% on the output class (as shown in Fig. 12(a)) whereas it decreased to 92.4% and increased to 94.2% respectively after changing the position (as shown in Fig. 12(b)). This brings the trustworthiness of the interpreted decision into question. The interpretable BRB approach attempts to address this by the way it presents the outputs. As

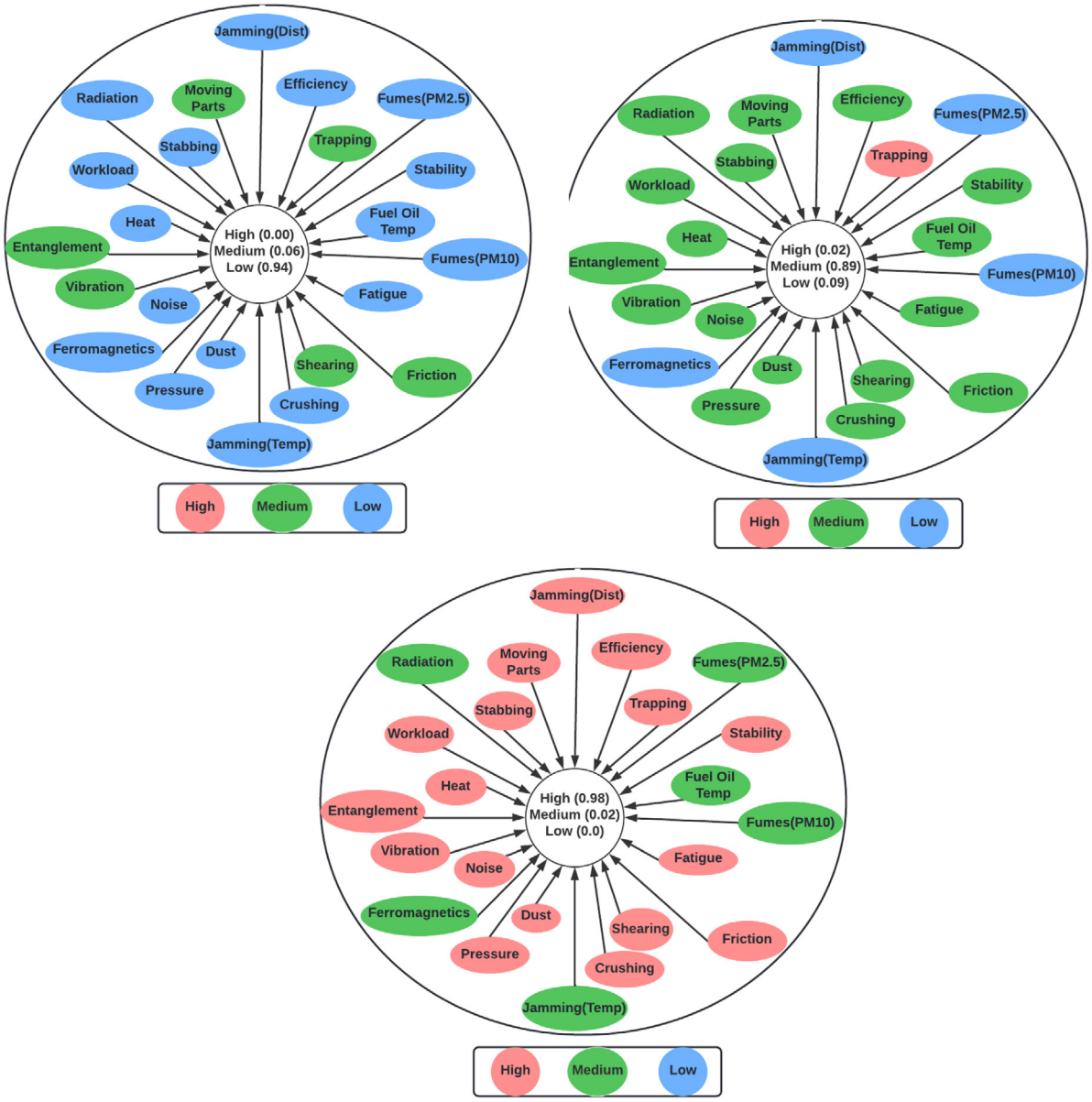


Fig. 8. Low, medium and high outputs of BRB approach.

shown in Fig. 8, the output of the BRB approach shows the possibility of each output class occurring along with the features that are contributing to them. So, in the model that has 23 features as input, the BRB model for a time slot determines the nPr = n!/(n-r)! different permutations of the inputs along with their influence values to ascertain the most predictable output risk class in that time slot. In the above analysis, *n* shows the number of features whereas *r* shows the number of output classes. When the problem has 50 different time slots, the approach produces 10626*50 = 531,300 different outputs. For each time slot, the output is then optimized according to the maximum possibility of occurrence of the output risk class before determining which input features along with their influence contribute to it. This assists the BRB approach to determine the most optimized output from the different possible sets as opposed to the one output that post-hoc explainers like LINDA-BN produce. By determining this analysis for all the time slots, as shown in Fig. 9, the BRB output represents the set of features that consistently change along with the change in the output risk class.

4.2. Output depends on what features are present as input to that time slot

Post-hoc explanation methods such as LINDA-BN commonly return a value for the relevance of a given feature. These relevance numbers show which features have the most significant impact on the forecasted output, whether positive or negative. This analysis is used as the most significant piece of information when evaluating the outcomes. A drawback here is that if a differing list of features is given, it will result in a different output, bringing the trustworthiness of the interpreted decision into question. For example, in Fig. 13(b), it can be seen that a number of features influences the output. Fig. 13(a) shows that when "pressure" is included in the top features, the influence of "trapping" on the output is 74.7%. However, when "pressure" is excluded from the features in Fig. 13(b), the influence of "trapping" on the output increased to 84.1%. So, the question in terms of the integrity of the presented output is which one of these is correct? In other words, LINDA-BN does not guarantee to the end user that the output it generates considers all the input

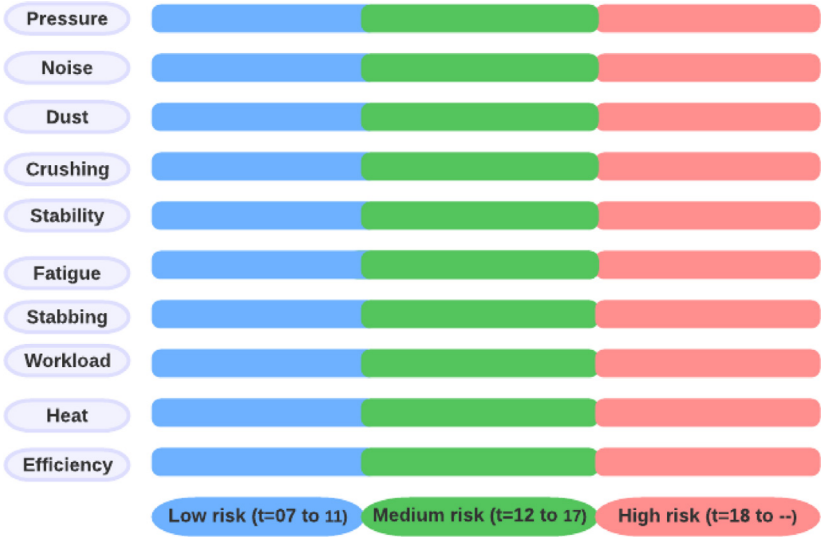


Fig. 9. Updating the features in each time slot.

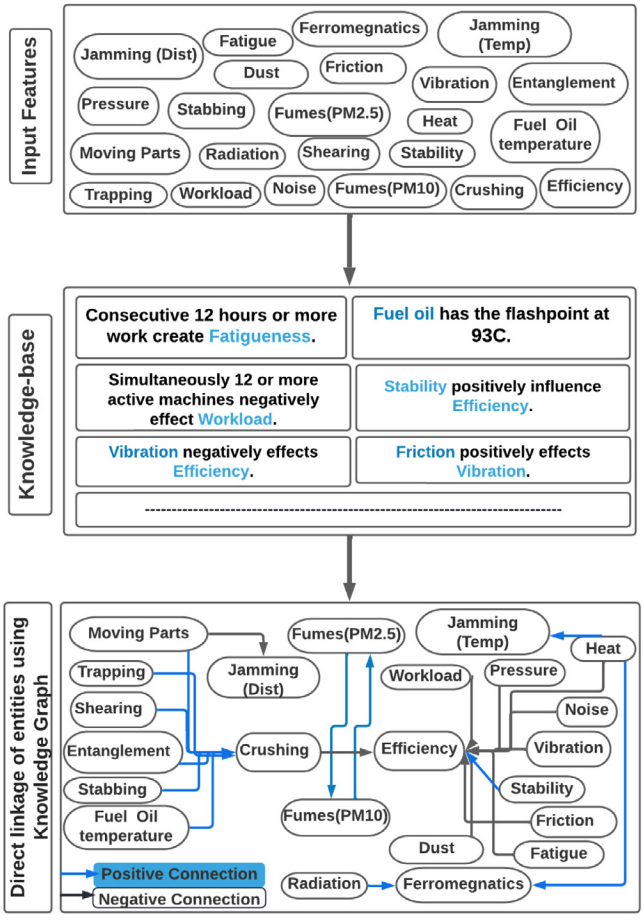


Fig. 10. Using KG to determine the changes in the feature values .

features for a time slot. The proposed BRB approach addresses this drawback by identifying inconsistencies in both the input dataset and the output class [49] by finding the most consistent and inconsistent features, as shown in Figs. 9 and 14 respectively, which strongly trigger the given output risk class in that time slot. Using this analysis, the risk manager can determine if the interpretation given by an approach to an output risk class considers the features that are strongly consistent with it. It also assists in ensuring that the output which is shown captures all the required inputs for it. This is discussed further in the next point.

4.3. Does not capture if the output shown considers all the required inputs

A key aspect to guarantee the trustworthiness of an interpreted output is to ensure that all the features that lead to that output are considered. LINDA-BN does not guarantee this and thus cannot guarantee the trustworthiness of its outputs. The proposed BRB approach does this by ascertaining if the inputs that are considered are such that it completes the occurrence of each output class. In other words, it detects incompleteness (missed and undefined) in the input dataset [50]. For an ideal (complete) dataset, the sum of all the output classes is 1 whereas for an incomplete dataset, it is less than 1. Fig. 8 shows that for the considered input dataset for the BRB approach, the total value of the output risk classes is 1. Therefore, the input dataset is said to be complete, i.e., it considers all the inputs that are required to make an output. Fig. 15 shows the output of the same time slot of Fig. 8 (High Risk) in which we do not consider the most consistent features of that output risk class. As a result, the output shows that the input dataset is incomplete which assists in producing (42% + 39% = 81%) of the output and for the rest (100%–81% = 19%), the input features are unknown. The BRB approach can also identify and resolve if any feature is vague (value or influence in confusing) or imprecise (value or influence in undefined) in a specific time slot [51,52].

4.4. Polarity of the contribution of the input features towards an output

The polarity of a feature's contribution (i.e., whether the feature contributes positively or negatively to the projected class) is an essential piece of information to consider while interpreting the output. LINDA-BN does not show this as it only represents a link between the features. The BRB approach addresses this by representing and showing only those features that contribute to the occurrence of each output risk class in a time slot.

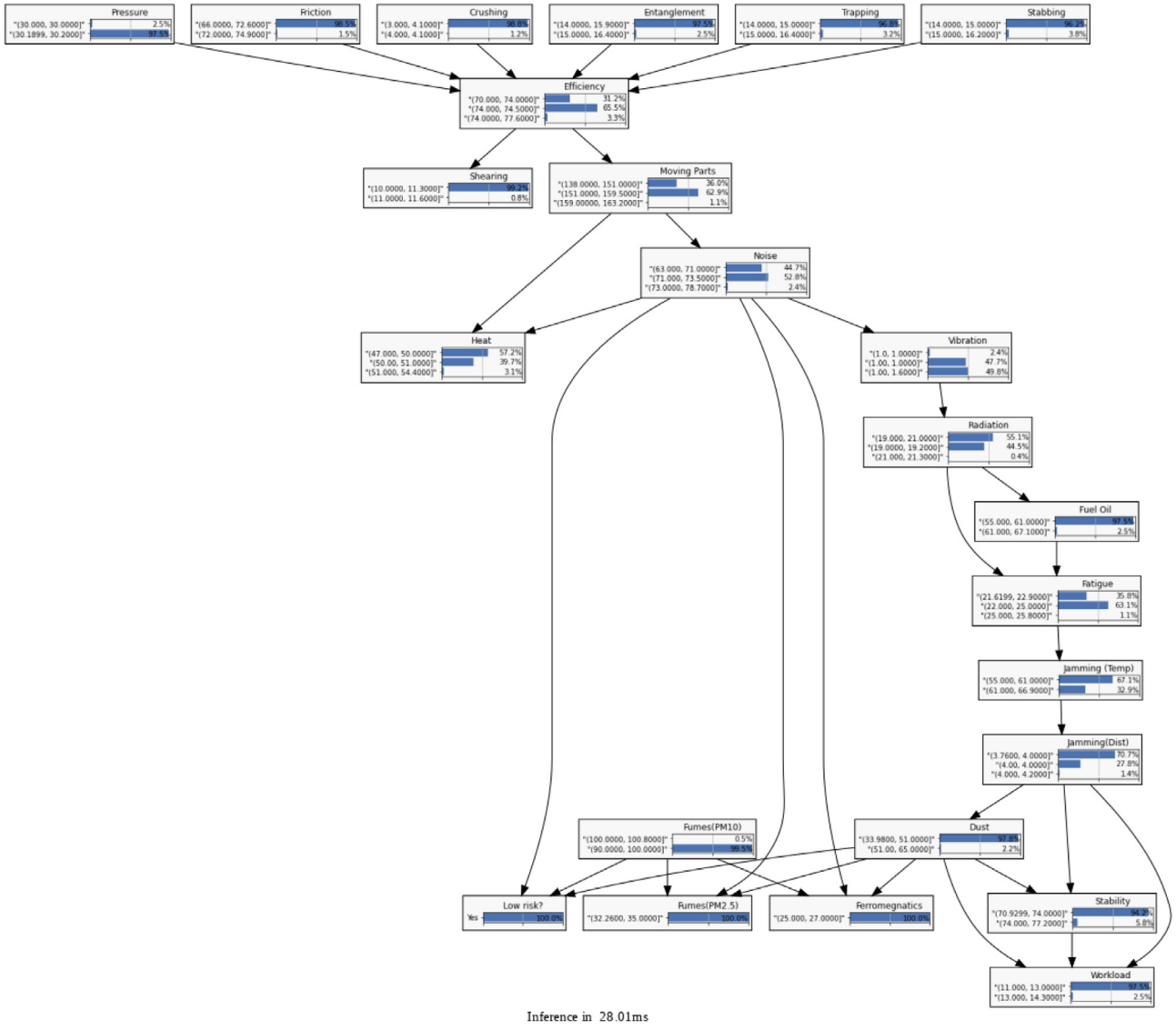


Fig. 11. Low risk output of LINDA-BN.

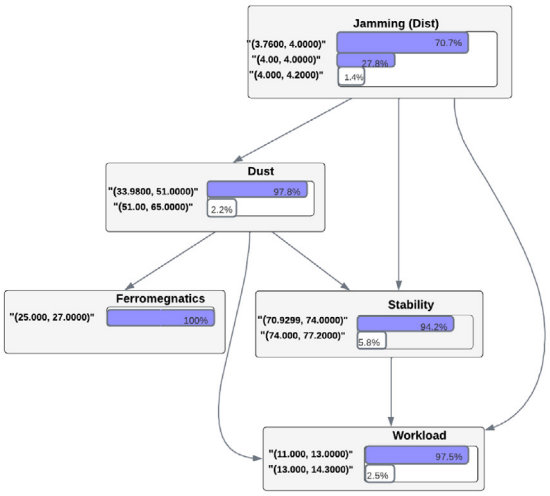
5. Comparing the outputs of the different XAI methods against the expected requirements from XAI approaches

In this section, we compare the outputs of the different XAI approaches and their ability to meet the different requirements of XAI from the users' perspective. The intention here is to determine from a human observer's perspective how well they can understand why the model is explaining a given output that has led it to make the judgement (or forecast) [53]. This is important to gain the users' trust which will form the primary motivation for them to use a model [54]. When users have faith in a system, they feel more confident and comfortable with it [55]. To have a consistent understanding of interpretability, there are established universal and objective requirements that XAI approaches should meet [56–58]. In this section, we determine the extent to which the proposed BRB framework, SHAP (modified with KG to consider changes in the feature values over a time period), LIME (modified with KG to consider changes in the feature values over a time period), LINDA-BN, and Modified LINDA-BN (modified with KG to consider changes in the feature values over a time period) meets the defined requirements. We evaluate the requirements from the perspective of four different categories of users, namely *end users (AI novices)*, *data experts*, *AI experts*, and *decision makers*. From this analysis, the degree of explainability of a method and the requirements to fill the identified gaps will be easily understood.

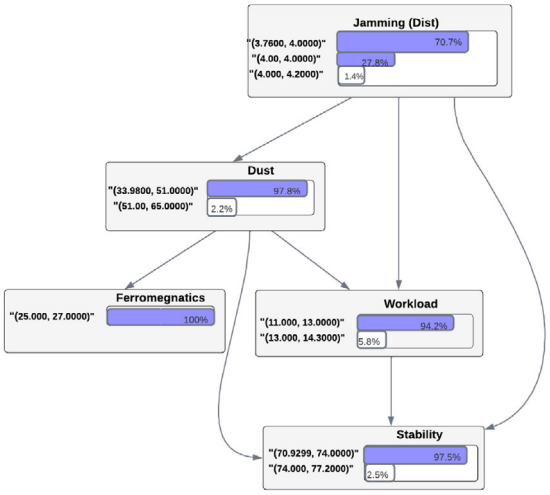
5.1. End users (AI novices)

End users are those who have slight or no familiarity with machine learning techniques yet still utilize AI products in their daily lives. Personalized mechanisms, e-commerce websites and social media are some of the most common applications used by end users. Intelligent and context-aware user interfaces of these applications rely on machine learning algorithms for many of their underlying operations and APIs [30]. Previous work has shown that end-users' adoption of technology increases by improving the intuitive interface and interaction design [38]. From the perspective of XAI, the following requirements (XAIRs) need to be met:

($XAIR_1$) **Transparency in algorithms:** An algorithm's widespread trust in its ability to behave sensibly [59]. It concerns how much of a system's inner workings can be understood "in theory". This may also involve making computational models and decisions that the user can understand [60]. The confidence in



(a) Output before changing the position of "Workload".



(b) Output after changing the position of "Workload".

Fig. 12. Shows how changing position of the feature can effect the output.

a method is proportional to the clarity with which its decision-making algorithms are presented.

($XAIR_2$) **Trust and reliance of users:** The ability of a method to choose truly relevant features is measured in terms of explainability [61]. The term "algorithmic trust" describes the general public's opinion that computers can be relied upon to safely process their personal information just as effectively as if a human were in charge [62]. Algorithms are well-known to operate within a strict set of codes and programmes, and it can be challenging to get them to deviate from their intended purposes.

($XAIR_3$) **Bias mitigation:** Bias mitigation in an algorithm is that which causes systematic and recurring errors that can lead to unfair outcomes, such as giving preference to one random group of users over others [63]. The conclusions of qualitative research might be distorted due to the presence of bias, which also leads to the collection of skewed data, which undermines the validity and reliability of the systematic study [64].

($XAIR_4$) **Privacy and security awareness:** The capacity of a system to achieve a security measure in all conceivable scenarios [65]. The primary focus of security is on keeping information safe, while the primary focus of privacy is on keeping individual identities secret [66]. The specific distinctions are, however, more complex, and there is a possibility for overlap. Specific to data, "security" means ensuring only authorized people can access it.

LIME and SHAP seek to interpret outputs at a local level by permutating the input and determining which variations are primarily considered to alter the concluded output [67]. The major drawback of these explainers is that they are post-hoc and it is not possible to alter or update the influence of the features automatically in different time periods. So, if these approaches are applied to progressive systems, they may give biased outputs if the features and their influence on the outcomes are not appropriately defined at the beginning of the system (input level). LINDA-BN uses conditional probability to identify the interconnected features [8]. Thus, it may give a biased or inappropriate output if the features and their conditional dependencies are not correctly captured. The same logic is applicable to Modified LINDA-BN as shown in Figs. 12 and 13(b) where biased and wrong outcomes are generated if the inputs are incorrect. However, the proposed BRB system produces all possible combinations of features and their influence for a specific input set by generating the different BRB rules (Fig. 7). It then selects the one which has the highest chance of occurrence. Furthermore, it also determines if the input dataset is complete or not. So, it is unlikely to produce a biased or wrong output.

5.2. Data experts

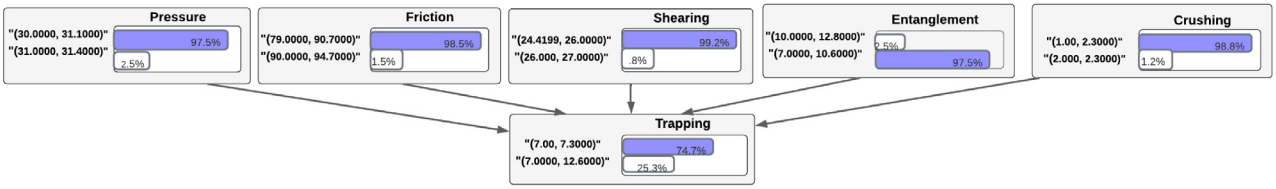
Data experts are those who specialize in the data profession in their respective fields. This category of users employs machine learning for computation, determination, or analysis [68]. Data experts explore information in different varieties and disciplines, such as medicine [69], cybersecurity [70], text [71], business [2] and image processing [72]. These users may be specialists in specific domains or specialists in public information technologies. However, in our classification, we presume that those in the data expert class have special technical training in AI algorithms and uses data analytics tools or visual analytic models to obtain insights into AI. So, the key requirements for the XAI model to meet to satisfy the needs of data experts are as follows:

($XAIR_5$) **Model Visualization and Inspection:** The goal of model visualization and inspection is for data experts to identify and understand model failures, as well as to increase model transparency and trustworthiness while reducing uncertainty [73].

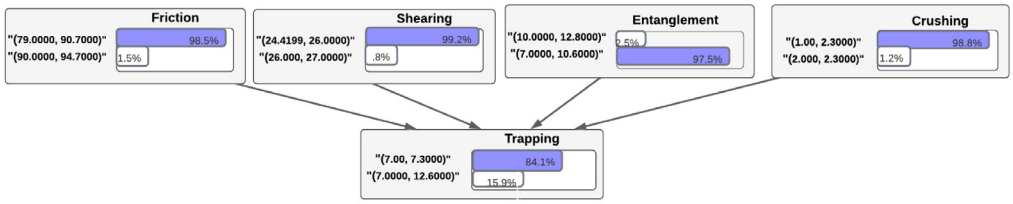
($XAIR_6$) **Model Tuning:** Model tuning allows data analysts to examine numerous methods and choose the most appropriate model for the desired data [74]. As an example, visual analytics tools improve the ability of deep neural network builders to alter networks, provide better training, and compare multiple networks [75].

($XAIR_7$) **Model Monotonicity:** An increase in the predictor's value creates a change in the probability of an instance belonging to the class, which in turn changes the relationship between the predictor and the predicted class [76]. In problems with monotonic objective and constraint functions, a monotonicity analysis can be used as a pre-optimization method. The analysis is used to obtain specific relationships among the decision variables of the optimal solutions without actually completing any optimization job [77]. Analysis for monotonicity seeks to prevent misleading interpretations.

As previously mentioned, LIME, SHAP, LINDA-BN, and Modified LINDA-BN are post-hoc explainers and provide a single output using a single set of inputs. Post-hoc explainers do not



(a) Influence of the top features on the feature "Trapping"



(b) Update output of "Trapping" after changing the top features.

Fig. 13. Shows the influence of top features on the output.

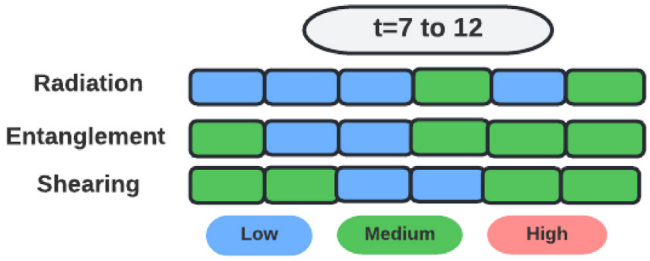


Fig. 14. Inconsistent influences of the features.

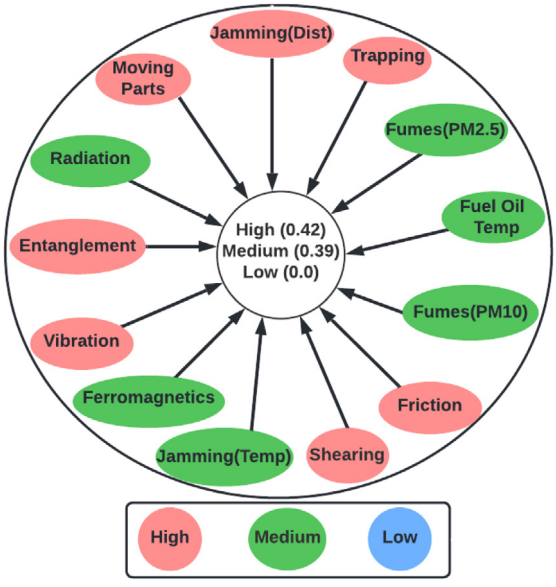


Fig. 15. Output provided by incomplete dataset.

provide any option to question the output, tune the model or validate it in different circumstances [78]. However, as shown in Fig. 5, the proposed interpretable BRB system is a glass-box system which shows the multiple decision options from which it selects one to be recommended. This option gives the advantage of tuning and validating the model, thereby ensuring more correctness and trust in the system.

5.3. AI experts

Our definition of AI experts includes researchers and engineers who focus on AI algorithms and interpretable approaches. XAI methods either produce model interpretations or model illustrations and include inherently explainable methods [79], deep methodology interpretations [80], visualizations of internal systems [81] etc., in the literature. Tools for interactively inspecting internal model variables, monitoring, and controlling the training process also helps AI experts [82]. The XAI requirements of AI experts are more precise, technical, core, in-depth algorithms and are as follows:

($XAIR_8$) **Model Explainability:** Model explainability provides transparency, justification, informativeness, and uncertainty estimation, which is the ability of a model to explain how it arrived at a specific response [83]. The degree of explainability quantifies the reliability of the prediction [84].

($XAIR_9$) **Model Debugging:** Model debugging enables the explanations to be examined to enhance system performance through different quality engineering techniques, namely detecting dataset bias [85], model failure prediction [86], adversarial attack detection [87] etc., to provide more accurate and precise output.

($XAIR_{10}$) **Post-hoc Explanation:** Post-hoc explainers are concerned about explaining the output and may apply the black-box approach to generate output [88]. Due to the constraints inherent in post-hoc approaches, they cannot be relied on as the primary mechanism to ensure the fairness of model outcomes when high-stakes decisions are being made [89].

As previously explained, LIME, SHAP, LINDA-BN and Modified LINDA-BN are post-hoc explainers that explain the outputs after producing them. This means that users will have no access to how these methods have generated an output nor will they know whether it is correct or not. Another thing to consider is that the proposed BRB approach is a glass-box system where users can see and verify each step from the inputs to the output. This provides transparency and enhance trust and confidence in the recommended decision.

Table 3
XAI requirements and their measures.

Different XAI approaches/parameters of XAI	SHAP	LIME	LINDA-BN	Modified LINDA-BN	Interpretable BRB
End users (AI Novices)					
$XAIR_1$	✓	✓	✓	✓	✓
$XAIR_2$	✓	✓	✓	✓	✓
$XAIR_3$	×	×	×	×	✓
$XAIR_4$	✓	✓	✓	✓	✓
Data experts					
$XAIR_5$	✓	✓	✓	✓	✓
$XAIR_6$	×	×	×	×	✓
$XAIR_7$	✓	✓	✓	✓	✓
AI experts					
$XAIR_8$	✓	✓	✓	✓	✓
$XAIR_9$	×	×	×	×	✓
$XAIR_{10}$	✓	✓	✓	✓	×
Decision makers					
$XAIR_{11}$	×	×	×	×	✓
$XAIR_{12}$	×	×	✓	✓	×
$XAIR_{13}$	×	×	✓(conditional)	✓(conditional)	✓
$XAIR_{14}$	×	×	×	✓	✓
$XAIR_{15}$	✓	✓	✓	✓	✓
$XAIR_{16}$	×	×	×	×	✓
$XAIR_{17}$	×	×	×	✓	✓
$XAIR_{18}$	×	×	×	×	✓

5.4. Decision makers

Another important category of users that need XAI systems is decision makers. Different measures are needed to verify and validate explanations so that decision-makers can confidently make decisions in different circumstances. XAI evaluations use a variety of supervised in-lab and research projects to gather input from experts in the field while they undertake high-level cognitive tasks using evaluation tools. [57]. Although this is needed, decision makers or users of XAI also need evaluation measures that use interpretable algorithms so they can quickly determine the completeness and precision of explanations. [90]. So the requirements which XAI approaches need to meet in this category are as follows:

($XAIR_{11}$) **Does the system need special hardware and processing speed support?:** This requirement specifies the degree of professional and technical skills needed to implement a system. LIME, SHAP, LINDA-BN and Modified LINDA-BN give a single output for a single input set. However, the proposed interpretable BRB approach produces nPr = n!/(n-r)! number of outputs for a single input set where n denotes the quantity of the input features. When interpretable BRB is applied in a large dataset, special hardware and speed support are needed to run the system smoothly.

($XAIR_{12}$) **Does the system change its behaviour for datasets of different sizes and shapes?:** This requirement defines the degree of relevancy of the method in diverse domains and applications. One of the significant drawbacks of LINDA-BN is that it is not applicable for a large dataset. Furthermore, an error generated in a time slot propagates and becomes more prominent over the time space when applied to a time series dataset [8]. These drawbacks are addressed by the BRB approach.

($XAIR_{13}$) **Is it possible to visualize all the features participating in the assessment (average dataset)?:** This requirement detects the degree of explainability and accuracy in terms of visualization. LIME and SHAP visualize only the features that are associated with the output. LINDA-BN can visualize all the features if the dataset is small (features equal to or less than 30) whereas Modified LINDA-BN does it for features equal to 23. In comparison, the interpretable BRB approach shows all of the features that are participating in the system.

($XAIR_{14}$) **Does the system capture the interconnections between the features that are participating in the output?:** This requirement portrays the degree of accuracy, explainability and overall acceptability in terms of the data processing capability of the approach. Modified LINDA-BN and the proposed BRB approach uses a knowledge graph (causal links) to capture the interconnections between the features. However, LIME and SHAP does not do this.

($XAIR_{15}$) **Does the system visualize all the features responsible for the output?:** This requirement describes the degree of output interpretation of the system. This requirement is one of the pivotal features of XAI methods. LIME, SHAP, LINDA-BN, Modified LINDA-BN and the interpretable BRB approach meets this requirement.

($XAIR_{16}$) **Does the system have the ability to provide all the possible output:** This requirement determines if the XAI approaches quantify the different possible outputs in a time slot. As shown in the workings of LINDA-BN and modified LINDA-BN, they only show the most probable output occurring in a time slot, as does SHAP and LIME. However, the proposed BRB approach as shown in Fig. 7 shows all the possible outputs in a time slot.

($XAIR_{17}$) **Does the system capture the changes in each step when applied to a time-series dataset?:** This requirement estimates the degree of comprehensive applicability of the approach. Modified LINDA-BN applies a KG and SD and the proposed BRB approach employs a knowledge graph (causal links) to capture the changes in the features in each time slot of the time series dataset.

($XAIR_{18}$) **Is the system glass-box?:** This requirement depicts the degree of interpretability, trustworthiness and accuracy of the entire model. The proposed BRB approach is a glass-box system but the other approaches are not as they use a post-hoc explanation of the output.

Table 3 in a tabular form shows how each of the considered XAI approaches performs against the aforementioned XAI requirements. As seen from the analysis, the proposed BRB approach performs better than the other approaches in terms of meeting the XAI requirements. In doing so, it also better ensures

the trustworthiness of the given interpretations compared to the other approaches.

6. Conclusion and future work

In this paper, we proposed a BRB framework that interprets the associated features along with their influences due to which a decision is reached. We instantiated this framework to obtain a risk class that pointwise estimates the importance of the features leading to it and gives an explanation behind the association. To apply this model to progressive decisions, we designed a knowledge graph that can capture the causal relationships between the features to estimate the level of influence. BRB was then applied to determine the most probable risk class to occur and the impact of the contributing features. Compared to popular post-hoc explainers, we demonstrated the advantages of our proposed method and show that it instils confidence in decision makers in terms of the trustworthiness of the generated outputs. Using these characteristics, the proposed BRB approach can detect impreciseness, vagueness, incompleteness and inconsistency in the input dataset as well as in the outputs generated. In doing so, we have also identified one of the critical challenges to using post-hoc explanations in practice. In our future work, we will further investigate this issue so that users can have greater trust in their recommended outputs.

CRediT authorship contribution statement

Sonia Farhana Nimmy: Co-conceptualization of the idea, Formal analysis, Co-development of the algorithms, programming, Data curation, Experiments, Initial and final draft version of the manuscript. **Omar K. Hussain:** Co-conceptualization of the idea, Formal analysis, Co-development of the algorithms, Experiments, Initial and final draft version of the manuscript. **Ripon K. Chakrabortty:** Writing – review & editing, Fedback towards finalization. **Farookh Khadeer Hussain:** Writing – review & editing, Fedback towards finalization. **Morteza Saberi:** Writing – review & editing, Fedback towards finalization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgements

Sonia Farhana Nimmy acknowledges the financial support from The University of New South Wales, Canberra to support her time on this project.

References

[1] J. Bunn, Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI), Rec. Manag. J. (2020).
[2] S.F. Nimmy, O.K. Hussain, R.K. Chakrabortty, F.K. Hussain, M. Saberi, Explainability in supply chain operational risk management: A systematic literature review, Knowl.-Based Syst. 235 (2022) 107587.
[3] C.T. Wolf, Explainability scenarios: towards scenario-based XAI design, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 252–257.
[4] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, D. Capuzzo, Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models, J. Oper. Res. Soc. (2020) 1–11.
[5] W. Zeng, A. Davoodi, R.O. Topaloglu, Explainable DRC hotspot prediction with random forest and SHAP tree explainer, in: 2020 Design, Automation & Test in Europe Conference & Exhibition, DATE, IEEE, 2020, pp. 1151–1156.
[6] G. Plumb, D. Molitor, A.S. Talwalkar, Model agnostic supervised local explanations, Adv. Neural Inf. Process. Syst. 31 (2018).
[7] R. Guidotti, Evaluating local explanation methods on ground truth, Artificial Intelligence 291 (2021) 103428.
[8] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models, Decis. Support Syst. 150 (2021) 113561.
[9] R. Alharbi, M.N. Vu, M.T. Thai, Evaluating fake news detection models from explainable machine learning perspectives, in: ICC 2021 - IEEE International Conference on Communications, 2021, pp. 1–6.
[10] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, Artif. Intell. Med. 94 (2019) 42–53.
[11] M.S. Kamal, N. Dey, L. Chowdhury, S.I. Hasan, K. Santosh, Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning, IEEE Trans. Instrum. Meas. 71 (2022) 1–9.
[12] Q. Ye, J. Xia, G. Yang, Explainable AI for COVID-19 CT classifiers: An initial comparison study, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2021, pp. 521–526.
[13] V. Moscato, A. Picariello, G. Sperlí, A benchmark of machine learning approaches for credit score prediction, Expert Syst. Appl. 165 (2021) 113986.
[14] A. Adak, B. Pradhan, N. Shukla, Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: Systematic review, Foods 11 (10) (2022) 1500.
[15] A. Nascita, A. Montieri, G. Aceto, D. Ciuonzo, V. Persico, A. Pescapé, XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures, IEEE Trans. Netw. Serv. Manag. 18 (4) (2021) 4225–4246.
[16] C.S. Hernandez, S. Ayo, D. Panagiotakopoulos, An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools, in: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference, DASC, IEEE, 2021, pp. 1–10.
[17] S.S. Matin, B. Pradhan, Earthquake-induced building-damage mapping using explainable AI (XAI), Sensors 21 (13) (2021) 4489.
[18] S. Petsis, A. Karamanou, E. Kalampokis, K. Tarabanis, Forecasting and explaining emergency department visits in a public hospital, J. Intell. Inf. Syst. (2022) 1–22.
[19] M.A. Rahman, M.S. Hossain, M.M. Rashid, S. Barnes, E. Hassanain, Ioevchain: a 5G-based secure inter-connected mobility framework for the internet of electric vehicles, IEEE Netw. 34 (5) (2020) 190–197.
[20] H. Suryotrisongko, Y. Musashi, A. Tsuneda, K. Sugitani, Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intelligence sharing, IEEE Access 10 (2022) 34613–34624.
[21] M.S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R.G. Crespo, E. Herrera-Viedma, Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes, IEEE Trans. Instrum. Meas. 70 (2021) 1–7.
[22] M.H. Lee, D.P. Siewiorek, A. Smailagic, A. Bernardino, S. Bermúdez i Badia, An exploratory study on techniques for quantitative assessment of stroke rehabilitation exercises, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 303–307.
[23] A. Oliver-Roig, J.R. Rico-Juan, M. Richart-Martínez, J. Cabrero-García, Predicting exclusive breastfeeding in maternity wards using machine learning techniques, Comput. Methods Programs Biomed. 221 (2022) 106837.
[24] M. Mahmud, M.S. Kaiser, M.A. Rahman, T. Wadhera, D.J. Brown, N. Shopland, A. Burton, T. Hughes-Roberts, S.A. Mamun, C. Ieracitano, et al., Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder, in: International Conference on Human-Computer Interaction, Springer, 2022, pp. 356–370.
[25] A. Warren, A. Hillas, Friend or frenemy? The role of trust in human-machine teaming and lethal autonomous weapons systems, Small Wars Insur. 31 (4) (2020) 822–850.
[26] C. Maathuis, On explainable AI solutions for targeting in cyber military operations, in: International Conference on Cyber Warfare and Security, Vol. 17, 2022, pp. 166–175.
[27] M. Szczepański, M. Pawlicki, R. Kozik, M. Choraś, New explainability method for BERT-based model in fake news detection, Sci. Rep. 11 (1) (2021) 1–13.
[28] A. Mondlane, K. Hansson, O. Popov, ICT for flood risk management strategies a GIS-based MCDA (M) approach, in: 2013 IST-Africa Conference & Exhibition, IEEE, 2013, pp. 1–9.
[29] E. Kallina, Delegating agency? the effects of XAI, personality traits, and the moral significance of the application on the reliance on autonomous systems: a user study.

[30] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

[31] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI Conference on Artificial Intelligenc, Vol. 32, 2018.

[33] D. Slack, A. Hilgard, S. Singh, H. Lakkaraju, Reliable post hoc explanations: Modeling uncertainty in explainability, Adv. Neural Inf. Process. Syst. 34 (2021).

[34] G. Laberge, Y. Pequignot, F. Khomh, M. Marchand, A. Mathieu, Partial order: Finding consensus among uncertain feature attributions, 2021, arXiv preprint arXiv:2110.13369.

[35] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.

[36] S. Upadhyay, S. Joshi, H. Lakkaraju, Towards robust and reliable algorithmic recourse, Adv. Neural Inf. Process. Syst. 34 (2021).

[37] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, C. Mooney, Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review, Appl. Sci. 11 (11) (2021) 5088.

[38] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[39] D. Garreau, U. Luxburg, Explaining the explainer: A first theoretical analysis of LIME, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1287–1296.

[40] A. Messalas, Y. Kanellopoulos, C. Makris, Model-agnostic interpretability with shapley values, in: 2019 10th International Conference on Information, Intelligence, Systems and Applications, IISA, IEEE, 2019, pp. 1–7.

[41] H. Li, W. Fan, S. Shi, Q. Chou, A modified LIME and its application to explain service supply chain forecasting, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2019, pp. 637–644.

[42] A. Adak, B. Pradhan, N. Shukla, A. Alamri, Unboxing deep learning model of food delivery service reviews using explainable artificial intelligence (XAI) technique, Foods 11 (14) (2022) 2019.

[43] A. Karamanou, E. Kalampokis, K. Tarabanis, Linked open government data to predict and explain house prices: the case of scottish statistics portal.

[44] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, D. Sontag, Learning a health knowledge graph from electronic medical records, Sci. Rep. 7 (1) (2017) 1–11.

[45] U. Jaimini, A. Sheth, CausalKG: Causal knowledge graph explainability using interventional and counterfactual reasoning, 2022, arXiv preprint arXiv:2201.03647.

[46] B. Zhou, X. Shen, Y. Lu, X. Li, B. Hua, T. Liu, J. Bao, Semantic-aware event link reasoning over industrial knowledge graph embedding time series data, Int. J. Prod. Res. (2022) 1–18.

[47] J.-B. Yang, J. Liu, J. Wang, H.-S. Sii, H.-W. Wang, Belief rule-base inference methodology using the evidential reasoning approach-RIMER, IEEE Trans. Syst. Man Cybern. A 36 (2) (2006) 266–285.

[48] Y.-M. Wang, L.-H. Yang, Y.-G. Fu, L.-L. Chang, K.-S. Chin, Dynamic rule adjustment approach for optimizing belief rule-base expert system, Knowl.-Based Syst. 96 (2016) 40–60.

[49] B.-C. Zhang, G.-Y. Hu, Z.-J. Zhou, Y.-M. Zhang, P.-L. Qiao, L.-L. Chang, Network intrusion detection based on directed acyclic graph and belief rule base, Etri J. 39 (4) (2017) 592–604.

[50] L. Wei, Z. Lv, D. Peng, E. Qi, Integrated energy systems security assessment based on belief rule base model, in: 2021 6th International Conference on Power and Renewable Energy, ICPRE, IEEE, 2021, pp. 1460–1465.

[51] J.-B. Yang, J. Liu, D.-L. Xu, J. Wang, H. Wang, Optimization models for training belief-rule-based systems, IEEE Trans. Syst. Man Cybern. A 37 (4) (2007) 569–585.

[52] Y. Cao, Z. Zhou, C. Hu, W. He, S. Tang, On the interpretability of belief rule-based expert systems, IEEE Trans. Fuzzy Syst. 29 (11) (2020) 3489–3503.

[53] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[54] R.R. Hoffman, A taxonomy of emergent trusting in the human–machine relationship, in: Cognitive Systems Engineering: The Future for A Changing World, CRC Press Boca Raton, FL, 2017, pp. 137–164.

[55] E. Thelisson, Towards trust, transparency and liability in AI/AS systems, in: IJCAI, 2017, pp. 5215–5216.

[56] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106.

[57] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4) (2021) 1–45.

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

[59] H.K. Dam, T. Tran, A. Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, 2018, pp. 53–56.

[60] M. Strobel, Aspects of transparency in machine learning, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019, pp. 2449–2451.

[61] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, Adv. Neural Inf. Process. Syst. 31 (2018).

[62] A. Schmitt, T. Wambsganss, M. Söllner, A. Janson, Towards a trust reliance paradox? Exploring the gap between perceived trust in and reliance on algorithmic advice, in: International Conference on Information Systems, ICIS, 2021.

[63] A. Ortega, J. Fierrez, A. Morales, Z. Wang, M. de la Cruz, C.L. Alonso, T. Ribeiro, Symbolic AI for XAI: Evaluating LFIT inductive programming for explaining biases in machine learning, Computers 10 (11) (2021) 154.

[64] W. Sun, O. Nasraoui, P. Shafto, Evolution and impact of bias in human and machine learning algorithm interaction, Plos One 15 (8) (2020) e0235502.

[65] A. Kuppa, N.-A. Le-Khac, Black box attacks on explainable artificial intelligence (XAI) methods in cyber security, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–8.

[66] N. Papernot, P. McDaniel, A. Sinha, M.P. Wellman, Sok: Security and privacy in machine learning, in: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2018, pp. 399–414.

[67] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, Lancet Digit. Health 3 (11) (2021) e745–e750.

[68] A.G. Hoepner, D. McMillan, A. Vivian, C. Wese Simen, Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective, Eur. J. Finance 27 (1–2) (2021) 1–7.

[69] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2020) 4793–4813.

[70] J.R. Goodall, E.D. Ragan, C.A. Steed, J.W. Reed, G.D. Richardson, K.M. Huffer, R.A. Bridges, J.A. Laska, Situ: Identifying and explaining suspicious behavior in networks, IEEE Trans. Vis. Comput. Graphics 25 (1) (2018) 204–214.

[71] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, S. Pan, An uncertainty-aware approach for exploratory microblog retrieval, IEEE Trans. Vis. Comput. Graphics 22 (1) (2015) 250–259.

[72] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognit. 120 (2021) 108102.

[73] D. Sacha, H. Senaratne, B.C. Kwon, G. Ellis, D.A. Keim, The role of uncertainty, awareness, and trust in visual analytics, IEEE Trans. Vis. Comput. Graphics 22 (1) (2015) 240–249.

[74] E. Alexander, M. Gleicher, Task-driven comparison of topic models, IEEE Trans. Vis. Comput. Graphics 22 (1) (2015) 320–329.

[75] N. Pezzotti, T. Höllt, J. Van Gemert, B.P. Lelieveldt, E. Eisemann, A. Vilanova, Deepeyes: Progressive visual analytics for designing deep neural networks, IEEE Trans. Vis. Comput. Graphics 24 (1) (2017) 98–108.

[76] A.A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD Explor. Newsl. 15 (1) (2014) 1–10.

[77] S. Wang, M. Gupta, Deontological ethics by monotonicity shape constraints, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2043–2054.

[78] P. Blunsom, O.-M. Camburu, J. Foerster, E. Giunchiglia, T. Lukasiewicz, Can I trust the explainer? Verifying post- hoc explanatory methods, 2019, CoRR.

[79] V. Arya, R.K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S.C. Hoffman, S. Houde, Q.V. Liao, R. Luss, A. Mojsilovic, et al., AI explainability 360: An extensible toolkit for understanding data and machine learning models, J. Mach. Learn. Res. 21 (130) (2020) 1–6.

[80] M. Wu, M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, F. Doshi-Velez, Beyond sparsity: Tree regularization of deep models for interpretability, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[81] P. Angelov, E. Soares, Towards explainable deep neural networks (xDNN), Neural Netw. 130 (2020) 185–194.

[82] M. Liu, J. Shi, K. Cao, J. Zhu, S. Liu, Analyzing the training processes of deep generative models, IEEE Trans. Vis. Comput. Graphics 24 (1) (2017) 77–87.

[83] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International Conference on Machine Learning, PMLR, 2018, pp. 2668–2677.

[84] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (3) (2017) 50–57.

[85] Q. Zhang, W. Wang, S.-C. Zhu, Examining cnn representations with respect to dataset bias, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[86] A.K.M. Nor, S.R. Pedapati, M. Muhammad, V. Leiva, Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data, Mathematics 10 (4) (2022) 554.
[87] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.
[88] A.M. Sharif, Z. Irani, P.E. Love, Integrating ERP using EAI: a model for post hoc evaluation, Eur. J. Inf. Syst. 14 (2) (2005) 162–174.
[89] D. Curran-Everett, H. Milgrom, Post-hoc data analysis: benefits and limitations, Curr. Opin. Allergy Clin. Immunol. 13 (3) (2013) 223–224.
[90] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D.A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, IEEE, 2019, pp. 4197–4201.