Journal of Manufacturing Systems 75 (2024) 102–115

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

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

Technical paper

Ensemble Bayesian Network for root cause analysis of product defects via learning from historical production data

Karen Wang [a], Chao Liu [b], Yuqian Lu [a,*]

[a] Department of Mechanical and Mechatronics Engineering, The University of Auckland, Auckland, 1010, New Zealand
[b] College of Engineering and Physical Sciences, Aston University, Birmingham, B47ET, United Kingdom

ARTICLE INFO

Keywords:
Root cause analysis
Bayesian network
Ensemble learning
Product quality

ABSTRACT

Root Cause Analysis (RCA) of product defects is crucial to improving manufacturing quality and productivity. However, current efforts to localize root causes are prone to limitations in the aspect of robustness, causality discovery, knowledge representation, stochasticity, and sample size. Therefore, we propose a product-wise Ensemble Bayesian Network (EBN) to provide a robust, intelligent and human-interpretable probabilistic reasoning method for RCA. BN is adopted to enable interpretable probabilistic reasoning under uncertainty. We developed various structure learning algorithms, a parameter learning algorithm, and a Bayesian inference algorithm for BN to learn the root causes of product quality issues from historical product defect records. Our Ensemble Learning (EL) techniques enhance BN base learners with bootstrapped re-sampling and combine the predictions from multiple structure learning algorithms, ensuring a robust performance of BN. The framework structure is modularized by products to reduce the sample size and achieve high efficiency. We proved our method achieved good performance in acquiring causal knowledge, identifying the root cause with probabilities, and predicting quality risks in production, from implementation and extensive experimental testing on real-world data collected from the plastic industry.

1. Introduction

Root Cause Analysis (RCA) is essential to identifying product quality issues and improving manufacturing excellence. However, RCA is a difficult and time-consuming engineering problem [1]. The current RCA methods, including statistical and machine learning methods, are prone to various limitations, such as heavy reliance on human knowledge, lack of interpretability, inefficiency, lack of causality discovery, and coverage of uncertainty. While Bayesian Network (BN) seems to bridge the gaps with its probabilistic reasoning ability to discover the root cause under uncertainty [2–4], BNs could be computationally expensive [3] and unrobust due to data size, data sparsity [5], and the selection of BN structure learning algorithms [6]. Inspired by the success of ensemble learning in machine learning, we believe that fusing multiple learning algorithms with ensemble learning techniques could provide more robust RCA results. Besides, adopting a product-wise framework to build BN models per product type could help improve BN model accuracy as product defect root causes may vary between product types.

Therefore, we present a product-wise Ensemble BN (EBN) where BN is adopted as the fundamental learner for RCA to accommodate the stochastic nature of manufacturing process variations [3] and to predict the probability of the potential product defect root causes. Ensemble learning techniques are integrated to address the lack of robustness in single BN learners. The framework is modularized by product type to reduce the size of BN, increasing computational efficiency. Our contributions are fourfold:

1. Developed an interpretable, data-driven, and probabilistic reasoning solution for RCA using BN, allowing manufacturers to engage the causal knowledge visually and efficiently.
2. Incorporated ensemble learning methods to address the robustness issues in existing BN models, via aggregating multiple BNs learned from bootstrapped samples.
3. Designed a quantitative performance evaluation method for assessing probabilistic reasoning results
4. Compared the performance of our model against individual structure learning algorithms under different knowledge sources. The results offer a direction for model strengthening and model selection.

The rest of the paper is as follows. Section 2 summarizes knowledge gaps in the literature. We detail the RCA problem to be solved and

* Corresponding author.
E-mail addresses: kwan528@aucklanduni.ac.nz (K. Wang), c.liu16@aston.ac.uk (C. Liu), yuqian.lu@auckland.ac.nz (Y. Lu).

https://doi.org/10.1016/j.jmsy.2024.06.001
Received 25 November 2023; Received in revised form 3 May 2024; Accepted 3 June 2024
Available online 14 June 2024
0278-6125/© 2024 The Author(s). Published by Elsevier Ltd on behalf of The Society of Manufacturing Engineers. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

present our overall Ensemble BN solution in Section 3. Section 4 presents the detailed methods and algorithms. Section 5 verifies the effectiveness of our solution via real-world case studies. Section 6 concludes the work.

2. Related work

This section reviews RCA methods and identifies the knowledge gaps. Root cause analysis of product defects refers to investigating the causal factors that lead to quality deviations [7] [8]. The purpose of RCA is threefold: (a) to identify the root cause of a problem, (b) to learn and understand the underlying mechanics of the issue, and (c) to identify appropriate corrective action to rectify the situation systematically [9]. The typical challenges and enablers of RCA have been thoroughly investigated in [10,11]. Generally, RCA methods can be categorized into two groups — statistical techniques and machine learning methods.

2.1. Statistical techniques

Statistical approaches exploit the statistical features in the data to assist the RCA process, including Principal Component Analysis (PCA) [12], Partial Least Squares (PLS) [12], Fisher Discriminant Analysis (FDA) [12], Dynamic Principal Component Analysis (DPCA) with minimax distance classifier [13], and Discriminant Partial Least Squares (DPLS) [12]. However, these methods alone are not sufficient to perform sophisticated RCA; they often need to integrate with other classifiers [13], such as multivariate statistical control charts [14,15] or feature selection techniques [16]. Moreover, as the data size decreases, their performance deteriorates [13]. A statistical method that fuses Dynamic Principal Component Analysis (DPCA) and minimax distance classifier was implemented to simultaneously monitor and diagnose an automatically controlled process, though undesired performance was observed with smaller data samples [13]. In the meantime, DPLS was used for RCA on the failure in the Tennessee Eastman chemical plant by maximizing covariance between the predictors [12]. The root cause was successfully detected, whereas the assumption of multivariate Gaussian distribution for the control limits of the PCA or PLS-based monitoring indices restricted their validity and adaptability to realistic process data [17]. Abdelrahman and Keikhosrokiani [18] applied a statistical RCA on assembly line anomaly detection by studying the minimum and maximum values of each attribute in the data. They used Pareto chart to visualize the frequency of the possible causes and its cumulative occurrence. Though the level of impact of each individual root cause was calculated, their complex interdependencies were not analyzed. An application of FDA for RCA in the chemical processing industry revealed FDA's shortcoming in capturing nonlinear behavior in the data [16]. It led to poor performance with an overall misclassification rate of 38%. With the aid of feature selection algorithms, the misclassification rate dropped to 17% [16].

Statistical RCA approaches can generally assist RCA in a short run time. However, their performance is not the most competitive [12]. They struggle with non-linear relationship modeling [16], the requirement of large data size [13] and support algorithms [7,16,19] and lack of interpretability. Therefore, more integrated and automated RCA methods are required.

2.2. Machine learning methods

Machine learning techniques enable automated RCA by pattern mapping and knowledge acquisition from historical product defect records. Algorithms, such as decision tree, Support Vector Machine (SVM), Artificial Neural Network (ANN) and BN, have been leveraged to automatically identify the root cause from the historical production data under faulty situations [20]. Decision tree is popular thanks to its nature of generating human-interpretable results [21]. Chen [22] presented a decision tree learning approach to diagnose failures in large Internet sites. Detzner proposed an improved method of the interactive decision tree to combine experts' domain knowledge into the pattern recognition process in the automotive industry [21]. However, decision tree did not seem to be a well-performed classifier as it required a longer sampling time [1] and was incompetent in handling scarce datasets [22]. Other research has shown that SVM outperforms conventional classification technologies on root cause diagnosis. Chiang [16] used SVM to determine the root cause of the observed out-of-control status in the chemical processing industry, outperforming FDA by three times on the misclassification rate. SVM also tends to run faster [23] and to have a stronger generalization capability with small sample learning problems [24]. Unfavorably, the recognition accuracy of SVM degrades severely when the two crucial structural parameters, penalty factor, and kernel function parameter, are not tuned desirably [16,23]. ANN has proven effective for RCA to recognize patterns in the data easily, even over distorted inputs whilst yielding relatively high accuracy [18] and flexibility [25] in classifying the root causes. On the flip side, ANN is subjected to a long training time and the risk of poor convergence with increasing layers [26]. It is also criticized for its black box feature, omitting a logical explanation between inputs and outputs [27]. Accordingly, manufacturers are reluctant to employ it in the real world.

In general, the aforementioned machine learning methods face various issues, such as prediction accuracy, data scarcity, and causality discovery. Moreover, most methods model the RCA problem as a classification problem in determining whether the reason for defects belongs to a class. Such approaches do not account for the probability of multiple root causes nor the distinct strength of their causal influences on product quality. However, probabilistic reasoning explains the causal influence of the potential root causes with stochasticity. It is an important attribute for RCA in industrial practice as it includes uncertainty and supports decision-making for on-site staff.

Another common limitation among the existing RCA methods is the lack of interpretability. Even though the final root causes are identified, it is intractable to explain the causality of the root cause (i.e., why the identified root cause contributes to the issue) without a human-interpretable knowledge representation. For example, Lee et al. [28] developed an attention mechanism-based LSTM model for RCA in semiconductor yield enhancement considering the order of manufacturing processes. Relying solely on data and deep learning method, their approach showed obvious limitations in interpreting the causality of the root causes due to lack of knowledge integration. This limits the manufacturer's ability to find corresponding actions to solve the problem in the real-world scenario, as the results are not explainable and not visualized. Therefore, there is an urgent need for a robust, intelligent, human-interpretable probabilistic reasoning method for RCA.

BN [4,29–31] has emerged with sheering benefits [7,32], in particular for addressing uncertainty [2,3] where multiple root causes can contribute to the occurred product failure with various probabilities. The probabilities of different root causes inferred from BN can also quantify the strength of their causal influences on the quality issue. Causal relationships can be discovered in BN through modeling conditional dependencies between different variables, making it powerful for reasoning in RCA [7,32,33]. BN is also a powerful tool for knowledge representation as it displays the relationship amidst different features [4]. Acceptable results can be obtained by BN even with incomplete data [4]. For example, Kirchhof et al. [34] developed a large-scale, cross-process Bayesian Failure Network for RCA in lithium-ion battery production based solely on knowledge from process experts. Though acceptable prediction results were obtained, the lack of integration of production data on failures has significantly limited their model's performance. Every coin has two sides; implementation of BN is an NP-hard problem. The computational expense increases, as the network size goes up [3]. Furthermore, BN's performance can

be worsened without any prior knowledge [2]. The robustness of BN is sensitive to data sparsity [6], and the choice of structure learning algorithms [35]. BN models tend to lose robustness for big datasets. The sparsity from big data can make BNs have low independent validation accuracy and be overfit [6]. Moreover, the prediction accuracy of an individual BN model has been demonstrated to be dependent on the selection of the BN structure learning algorithm [35]. This means that BN learned from a single structure learning algorithm can be insufficient and unrobust for identifying root causes. Comparatively, fusing multiple learning algorithms via ensemble learning [36,37] can improve the robustness of the models. Thus, we are inspired to develop a more robust and accurate BN-based RCA framework by incorporating an ensemble learning philosophy, as the first of its kind for solving product defects issues via learning from historical production signals.

In summary, while statistical techniques and traditional machine learning methods have enhanced the accuracy and efficiency of RCA, they often treat RCA as a simplistic classification problem, typically binning outcomes into 'defect' or 'no defect' categories. This approach fails to account for the complex interplay of multiple root causes and their probabilistic impacts on product quality. Furthermore, such methods often lack the interpretability needed for practical application in real-world manufacturing settings. In contrast, BN has the ability to perform human-interpretable probabilistic learning, dynamically incorporating new data to update its understanding of root causes, and effectively modeling causal relationships that reveal how various factors interact to impact product quality. This approach not only supports the integration of expert knowledge alongside data-driven insights but also provides a flexible and responsive framework suitable for the variable conditions of real-world manufacturing. However, the robustness of BNs can be sensitive to the availability of data and the selection of appropriate structure learning algorithms. Therefore, there is an urgent need for an enhanced BN approach that integrates ensemble learning to mitigate data scarcity and overfitting issues, thereby providing a more accurate, interpretable, and robust RCA method.

3. Problem formulation and solution

This section formulates the RCA problem to be solved and our EBN method Table 1.

3.1. Problem formulation

Fig. 1 shows the overall RCA problem via learning causal relations from historical production records. Generally, a factory monitors each job's key production status and product defects along the process. Given a factory, during the past period, has produced M distinct types of products $Pd = \{Pd_1, Pd_2, \dots, Pd_M\}$ from N production jobs $J = \{J_1, J_2, \dots, J_N\}$, resulted in a historical production dataset D. This dataset tracks I features — production-related factors that may signal product defect issues (such as the operator, equipment used, and cycle time), in $X = \{X_1, X_2, \dots, X_I\}$. Each feature is indexed as X_i or (X_i, X_j) in a pair. Let R_n be the product quality of job J_n. It is a binary variable, $R_n \in \{0, 1\}$; 1 indicates that job J_n is problematic with quality issues; and 0 indicates no quality issue. The occurred root causes, $Y = \{Y_1, Y_2, \dots, Y_C\}$ where C denotes the total number of occurred root causes, in the records are regarded as the root-cause candidates leading to product quality issues (i.e., $R = 1$). Each root cause Y_c has a corresponding probability P_c implying the likelihood of root cause Y_c causing the identified quality issues.

With all the elements involved in RCA defined, the three questions in RCA can be formulated mathematically as follows.

Q1. Given a historical production data D with features $X = \{X_1, X_2, \dots, X_I\}$, find a function h, such that it satisfies all the components in (1), where "1" represents there is causality between the pair and "0" otherwise. h is developed to detect the existence of causal relationships between job feature X_i and X_j, $h(X_i, X_j) \to \{0, 1\}$; the existence of causal relationships between job feature X_i and root cause Y_c, $h(X_i, Y_c) \to \{0, 1\}$; the existence of causal relationships between job feature X_i and the observation R, $h(X_i, R) \to \{0, 1\}$; and the existence of causal relationships between root cause Y_c and the observation R, $h(Y_c, R) \to \{0, 1\}$.

$$h(X_i, X_j) \to \{0, 1\}; h(X_i, Y_c) \to \{0, 1\}; h(X_i, R) \to \{0, 1\}; \\ h(Y_c, R) \to \{0, 1\}, X_i, X_j \in X, i \neq j, X_i, Y_c \in Y \tag{1}$$

Q2. Given a finished job J_n with quality issues $R_n = 1$, what are the non-empty set of root-cause variables $Z = \{Y_1, Y_2, \dots, Y_{C'}\}$ for $C' \leq C$, $Z \subseteq Y$, and what are their corresponding probabilities $P = \{P_{Y_1}, P_{Y_2}, \dots, P_{Y_{C'}}\}$ based on its job feature vector X_n?

Q3. Given a job $J_{n'}$ that has not been operated yet, what is the value of $R_{n'}$, $R_{n'} \in \{0, 1\}$?

Table 1
Table of notations.

Symbol	Definition
Indices	
m	Index of produced product types
n	Index of jobs
i	Index of job features
c	Index of root cause variables
k	Index of structure learning algorithms
s	Index of bootstrapped samples
Sets	
Pd	Set of products, $Pd = \{Pd_1, Pd_2, \dots, Pd_M\}$
J	Set of jobs, $J = \{J_1, J_2, \dots, J_N\}$
X	Set of job features, $X = \{X_1, X_2, \dots, X_I\}$
Y	Set of root cause variables, $Y = \{Y_1, Y_2, \dots, Y_C\}$
Y'	Set of root cause variables, $Y' = \{Y_1, Y_2, \dots, Y_{C'}\}$ for given job J_n
P	Set of probabilities for root cause variables, $P = \{P_{Y_1}, P_{Y_2}, \dots, P_{Y_{C'}}\}$
$P_{m,k}$	Set of root cause probabilities for BN structure $\mathcal{G}_{m,k}$
X^i	Job features for job i
D	Set of historical job records for all given product types Pd, $D = \{D_1, D_2, \dots, D_M\}$
A	Set of structure learning algorithms, $A = \{A_1, A_2, \dots, A_K\}$
$\mathcal{V}$	Set of vertices in a BN structure $\mathcal{G}$
$\mathcal{E}$	Set of arcs in a BN structure $\mathcal{G}$
p	Set of conditional probabilities for vertices $\mathcal{E}$ in a BN structure $\mathcal{G}$
E	Set of evidence of a to-be-predicted job J_n to be input into a BN structure $\mathcal{G}$ for Bayesian inference, $E \leftarrow X_n$
Constants	
M	Number of product types
N	Number of jobs
I	Number of features
C	Number of total occurred root causes
K	Number of implemented structure learning algorithms
S	Number of total bootstrapped samples
Variables	
J_n	nth production job
X_i	ith production feature of the jobs
Y_c	cth root cause contributing to the quality issues of the jobs
P_c	The probability of root cause Y_c contributing to quality issues of the jobs
R_n	Binary variable: 1 indicating job J_n has quality issues; 0 otherwise
D_{m_s}	sth even subset of historical production data for mth product
$\mathcal{G}_{m_s,k}$	The learnt BN structure from subset D_{m_s} via algorithm A_k
$\mathcal{G}_{m,k}$	The aggregated BN structure for mth product via algorithm A_k
v_i	A vertex in a BN structure $\mathcal{G}$, $v_i \in \mathcal{V}$
e_{v_i,v_j}	An arc connecting vertex v_i and v_j
$w_{\mathcal{G}_{m,k}}$	The weight assigned to algorithm A_k for product Pd_m in for fusing predictions from different BN models
f_k	The optimal frequency for algorithm A_k in RCA predictions
MAE	The mean of absolute difference between predicted and observed probabilities for a group of root causes
ε_{rank}	The ranking difference between prediction and observation on an ordered sequence

3.2. Proposed solution

This section presents the proposed product-wise Ensemble BN for solving the above three RCA problems.

R	Y	X										
isRejectFail	rrnDescription	rejQuantity	ToolID	matID	jobID	jobRun	equID	jobStartUserPk	...	jobStartTime	ProductCount	ActCycleTime
0	Setup Reject	205	1213	PPB00450	129	1	BM06	23	...	11/12/2014	7646	24.85
0	Setup Reject	202	1213	PPB00450	569	1	BM07	26	...	2/16/2015	6720	26.18
0	Setup Reject	256	1213	PPB00450	728	1	BM07	23	...	3/9/2015	11194	25.89
1	Yield Update Adjustment	187	1213	PPB00450	969	1	BM01	6	...	4/20/2015	13145	25.61
0	Setup Reject	91	1235	PPB00450	1121	1	BM07	6	...	5/15/2015	13670	26.14
0	Yield Update Adjustment	679	1235	PPB00450	1121	1	BM07	6	...	5/15/2015	13670	26.14
1	Setup Reject	4	1235	PPB00450	1228	1	BM07	30	...	6/11/2015	13538	26.02
0	Yield Update Adjustment	634	1213	PPB00450	1228	1	BM07	30	...	6/11/2015	13538	26.02
0	Setup Reject	73	1213	PPB00450	1383	1	BM01	23	...	7/5/2015	12892	25.7
0	Yield Update Adjustment	1209	1213	PPB00450	1383	1	BM01	23	...	7/5/2015	12892	25.7
0	Setup Reject	49	1213	PPB00450	2914	1	BM01	23	...	3/23/2016	5476	25.14
0	Yield Update Adjustment	267	1213	PPB00450	2914	1	BM01	23	...	3/23/2016	5476	25.14

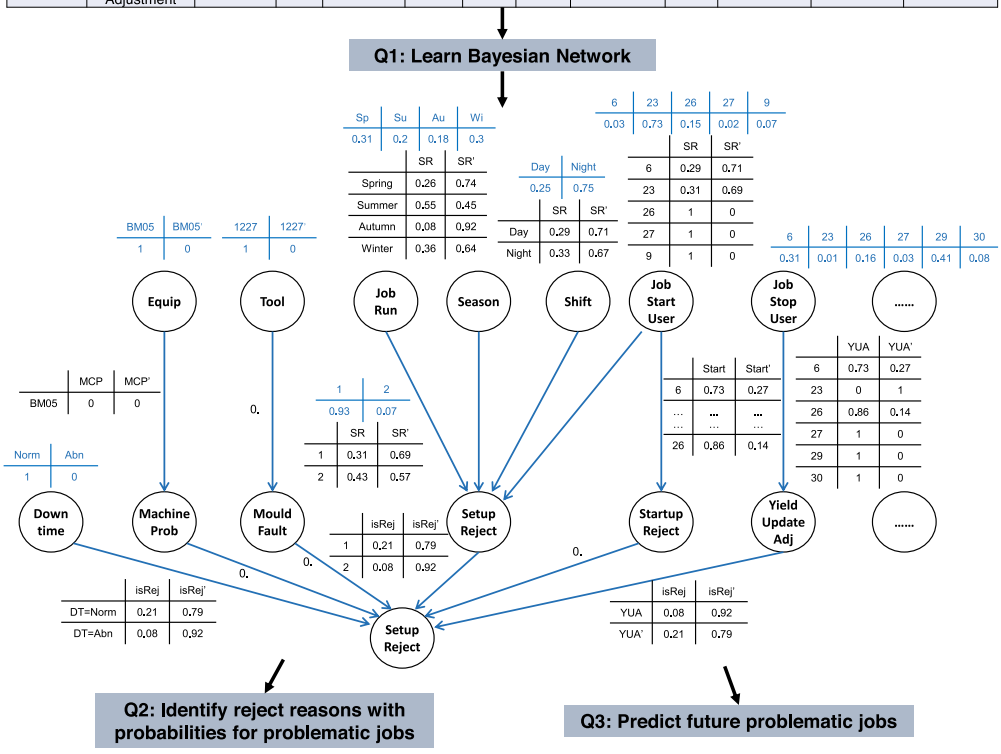


Fig. 1. Overall learning task from raw production records to answering three root cause analysis questions.

Our method employs a product-wise Ensemble BN model to find the causal relationships between job features X, potential root causes Y, and observation of product quality R for a product Pd_m based on its historical production data D_m. Once the causal network is discovered, the job features of the to-be-predicted job J_n, X_n will be introduced as external evidence E into the causal network to allow probability inferencing and risk prediction. The proposed framework consists of three steps outlined in Fig. 2: (I) — Modularize data by products, (II) — Construct bagged BN models, and (III) — Combine BN predictions using the Weighted Average Ensemble Learning (WAEL) technique.

In the first step, given a job J_n producing product Pd_m, the proposed method modularizes the historical data D into M (i.e., the number of product types) small product-wise data samples $\{D_1, D_2, \ldots, D_M\}$. Each dataset captures the historical job records that produce the corresponding product. This strategy reduces the sample size for BN. Hence, it increases BN learning efficiency and avoids sparsity that often occurs in big datasets. As a result, the historical data sample for job J_n will be D_m according to its product Pd_m.

Based on the modularized historical record D_m, in the second step, the structures of BN will be learned using a structure learning algorithm A_k with bagging ensemble technique. In total, there are K different structure learning algorithms, $\{A_1, A_2, \ldots, A_K\}$, implemented to learn BN models. Initially, D_m is bootstrapped evenly into S subsets for all the K algorithms as shown in Part II of Fig. 2, with each subset being D_{m_s}, $D_{m_s} \subseteq D_m$. Then, the structure learning algorithm A_k, $A_k \in \{A_1, A_2, \ldots, A_K\}$, will learn a BN model $\mathcal{G}_{m_s,k}$ for each subset D_{m_s}, resulting in a total of S models $\{\mathcal{G}_{m_1,k}, \mathcal{G}_{m_2,k}, \ldots, \mathcal{G}_{m_S,k}\}$ for algorithm A_k. Implementing different structure learning algorithms A_k for learning BN model $\mathcal{G}_{m_s,k}$ will be illustrated in Section 4.2, where the directed acyclic graphs (DAG) properties of the learned BN structures $\mathcal{G}_{m_s,k}$ are also defined. At the end of Step 2, S learned BN models are aggregated into one BN structure $\mathcal{G}_{m,k}$ using bagging ensemble technique. The bagging ensemble technique is integrated to account for sample variations and to reduce the risk of scarce data by fusing the models from the S bootstrapped samples, enabling robust BN structure learning. The detailed implementation of model aggregation using bagging ensemble is shown in Section 4.1.

After the bagged BN structure $\mathcal{G}_{m,k}$ is constructed with distinct structure learning algorithm A_k, parameter learning and inferencing of BN structure $\mathcal{G}_{m,k}$ need to be conducted in Step 3 to infer a set of root-cause probabilities, $P_{m,k} = \{P_{m,k_1}, P_{m,k_2}, \ldots, P_{m,k_C}\}$. As shown in Part III of Fig. 2, Step 3 starts with BN parameter learning where $p(v_i)$ for each v_i from all the vertices $\mathcal{V}_{m,k}$ in model $\mathcal{G}_{m,k}$ is estimated, $v_i \in \mathcal{V}_{m,k}$. The explicit parameter learning method will be elaborated in Section 4.3. Then, inference introduces the job features of the to-be-predicted job J_n as external evidence E into the BN model $\mathcal{G}_{m,k}$ (i.e., $E \leftarrow X_n$). Bayesian inference then updates the belief distribution

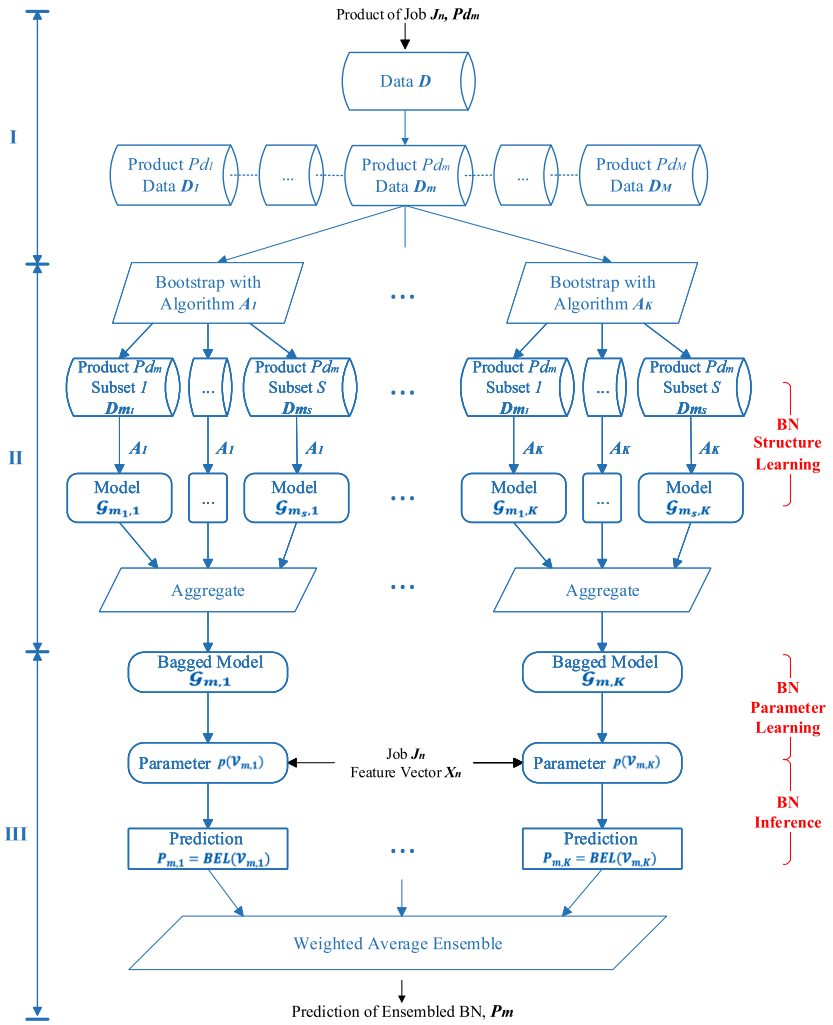


Fig. 2. The proposed product-wise Ensemble BN framework for finding the root cause relations between job features X, potential root causes Y and the observation of product quality R for product Pd_m based on its historical data D_m.

$BEL(v_i)$ for each $v_i \in \mathcal{V}_{m,k}$, based on the new evidence E using junction tree algorithms (explained in Section 4.4). $BEL(v_i)$ contains the probability of root cause Y_c contributing to quality failures, P_c, when v_i corresponds the root-cause variable Y_c. Similarly, the probabilities can be inferred for a set of potential defect root causes Y, obtaining $P_{m,k} = \{P_{m,k_1}, P_{m,k_2}, \ldots, P_{m,k_C}\}$ in the context of using structure learning algorithm A_k based on dataset D_m. Finally, the sets of predicted probabilities $\{P_{m,1}, \ldots, P_{m,K}\}$ from different BN models $\{\mathcal{G}_{m,1}, \mathcal{G}_{m,2}, \ldots, \mathcal{G}_{m,K}\}$ learned by distinct algorithms $\{A_1, A_2, \ldots, A_K\}$ are fused into a single set of root-cause probabilities $P_m = \{P_{m_1}, P_{m_2}, \ldots, P_{m_C}\}$ through WAEL technique to enhance prediction accuracy and robustness, which is further explained in Section 4.5. The resulting root-cause probabilities P_m can answer Q2 for RCA. The same procedures are taken to predict the quality risk $R_{n'}$, of a job $J_{n'}$ to answer Q3.

In summary, the proposed solution comprises two functional modules, BN models and ensemble learning techniques. BN is the fundamental model of our proposed method. It is a probabilistic graphical model for reasoning under uncertainty [23]. BN development process goes through three procedures, namely, structure learning, parameter learning, and inference, to provide probabilistic graphical reasoning. Structure learning algorithms for BN uncover the causal relationships between the variables $\mathcal{V} \subseteq \{X, Y, R\}$ from the historical data D and construct human-interpretable graphical networks. Parameter learning estimates the conditional probability tables (CPT) for each vertex v_i, which is an important attribute for inference. Inference updates the belief in the network by passing messages regarding probability distributions throughout the network to infer the probabilities of different root causes leading to the event of interest. As a result, intelligent and human-interpretable probabilistic reasoning is achieved by BN. On the other hand, ensemble learning techniques are incorporated to reinforce the robustness of the constructed BN models. Bagging ensemble techniques are applied during BN structure learning to counter BN's sensitivity to data sparsity. The weighted average ensemble learning technique is integrated after the BN inferencing. It fuses the predictions from the structures of BN models learned by different learning algorithms to alleviate the deficiencies in the accuracy and stability of a single BN model, ensuring robustness.

4. Methods

This section presents all the required methods and algorithms, in the sequence of bagging ensemble learning, BN structure learning, parameter learning, Bayesian inference, and WAEL, which follows the workflow of the proposed RCA solution in Section 3.

4.1. Bagging ensemble learning

Bagging ensemble learning technique is developed first so that it can be used to reinforce the structure learning process of BN in the following section. Bagging technique, also known as bootstrap aggregating, accounts for sample variations and reduces the chance of a

Algorithm 1: Bayesian Network Structure Learning with Bagging Ensemble

Input : Training data for product Pd_m, D_m, Structure learning algorithm A_k; Bootstrap rounds
Output: Bagged Bayesian Structure $\mathcal{G}_{m,k}$
for *each* D_{m_s} *in* $\{D_{m_1}, D_{m_2}, \ldots, D_{m_S}\}$ **do**
  $\mathcal{G}_{m_s,k} = A_k(D_{m_s})$
end
$\mathcal{G}_{m,k} = \frac{1}{S}\sum_{s=1}^{S} \mathcal{G}_{m_s,k}$

poor BN model induced by sparse data, enabling robust BN structure learning. As shown in Algorithm 1, it resamples the original data with replacement and implements homogeneous learners on the varying resulting samples.

4.2. Structure learning

The structure of BNs will then be learned by adopting the bagging ensemble technique. This sub-section presents all the different knowledge sources and structure learning algorithms used to learn the BN structures in our work.

4.2.1. Knowledge sources

This sub-section discusses different knowledge sources from which the BN structures are learned. The learned BNs are DAG, defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ denotes vertices (v_i, $v_j \in \mathcal{V}$), corresponding with the variables selected from the historical job records, including job features X, root-cause variables Y and quality risk indicator R, $\mathcal{V} \subseteq \{X, Y, R\}$. $\mathcal{E}$ denotes a set of arcs e_{v_i,v_j} in $\mathcal{G}$, signifying the conditional dependency between the connected random variables (v_i, v_j). If a causal relationship exists between v_i and v_j, an arc linkage e_{v_i,v_j} is established in the corresponding BN model, also referred to as causal knowledge. If the arc linkage e_{v_i,v_j} in the network is known ahead of structure learning, e_{v_i,v_j} is called prior knowledge. They are normally discovered by human instinct or experience. Structure learning is the process of deducing the structure of BN from the dependency relations in the data, aided by any prior knowledge as constraints. Our solution injects three different knowledge sources for BN structure learning, namely "Data", "Hybrid" and "Human". When a structure is purely learned from the data without any prior knowledge, its knowledge source is labeled as "Data"; When the entire structure is solely built on human knowledge without any structure learning algorithm, it is classified as "Human" knowledge source; When structure learning algorithms learn the BN structure with prior knowledge, it is defined as a "Hybrid" knowledge source as it involves both prior knowledge obtained from human and structure learning from the data. In our study, prior knowledge is derived from a combination of sources in the case factory. While data analysts play a crucial role in identifying and structuring this knowledge, it is also significantly enriched by surveys conducted with operators. These surveys capture valuable firsthand observations and experiences of the operators, providing insights into causal influences from job features to the root-cause variables. This collaborative approach ensures that our analysis integrates both the detailed analytical perspectives of the data analysts and the practical, operational insights of the operators, leading to a more accurate and comprehensive understanding of the manufacturing processes.

As a result, we will build BNs using these three different knowledge sources. Particularly, "Data" and "Hybrid" knowledge sources can be integrated with different structure learning algorithms.

4.2.2. Structure learning algorithms

In BNs, predicted probabilities are inferred based on a causal structure. However, the structure of a BN is not often known. In manufacturing, normally, only a few arcs, representing causal relationships, in the network could be established based on expert knowledge. Therefore, automatic structure learning is needed to deduce the structure of BN from the dependency relations in the data, aided by any prior knowledge as constraints. The structure learning task for a dataset of D_m for product Pd_m can be defined as determining a set of directed arcs $\mathcal{E}$ for the DAG to achieve some criterion used for evaluating the goodness of fit of the model. We adopted hill-climbing (hc), tabu search, pc.stable, grow-shrink (gs), incremental association Markov blanket (iamb), max–min hill-climbing (mmhc) and chow.liu algorithms as the base learners for learning the BN model. Hill-climbing and tabu search allow a greedy search without trapping in local optimum. However, they tend to overfit for large datasets. Grow-shrink and incremental association Markov blanket converge quickly, but they are subject to poor node connection with sparse or small data sets. Hybrid learning offers a sound skeleton identification process and parameter tuning. Chow.liu algorithm is time efficient, whereas its spanning tree structure and direction assigning method limits its use on non-tree network. Due to the varying practicality of different learning models in different cases, the learned models will then be aggregated via (6).

4.3. Parameter learning

After the structures of BN are learned, the strength of the conditional dependency between each pair of connected variables in BN needs to be obtained through CPTs to allow probability inferencing. Parameter learning entails computing the CPTs of each node given its parent nodes. We apply Bayesian parameter estimation to estimate the values of CPTs. $p(v_i)$ denotes the probability density function of an observable variable v_i, reflecting its contribution to a quality issue. With its distribution depending on the unknown parameter θ, $p(v_i|\theta)$ represents the prior probability density function for variable v_i, given θ. When new evidence $E = \{v_1, v_2, \ldots, v_n\}$ is found for variable v_i in the experiment, the goal of parameter learning is to compute $p(v_i|E)$ so that its value approaches the unknown $p(v_i)$ as close as possible. It is noted that θ is modeled as a random variable following distribution $p(\theta)$. Then the probability density function of v_i given a set of evidence E can be inferred as follows:

$$\begin{aligned} p(v_i|E) &= \int p(v_i, \theta|E)\, d\theta \\ &= \int p(v_i|\theta, E) p(\theta|E)\, d\theta \\ &= \int p(v_i|\theta) p(\theta|E)\, d\theta. \end{aligned} \tag{2}$$

As $p(v_i|\theta)$ is known before obtaining new evidence E, the posterior probability density function for parameter θ after E, $p(\theta|E)$, needs to be obtained. This is achieved by adapting Bayes' theorem as shown in:

$$p(\theta|E) = \frac{p(E|\theta)p(\theta)}{p(E)} = \frac{p(E|\theta)p(\theta)}{\int p(E|\theta)p(\theta)\, d\theta}, \tag{3}$$

where $p(\theta)$ is the prior distribution, and $p(E|\theta)$ is the likelihood function. $p(\theta)$ represents the knowledge of the parameter before engaging the information from the data, while $p(\theta|E)$ updates the distribution succeeding the introduction of new evidence. Consequently, the probability distribution of node v_i can be estimated. For a node v_i, $v_i \in \mathcal{V}$, if v_i has a parent, its CPT will be conditional probability distribution based on the states of its parent nodes $p(v_i|pa(v_i))$; else, its CPT will be its own probability distribution $p(v_i)$. These conditional probabilities also embody the strength of causal dependency relations between a pair of variables (v_i, v_j). These learned CPTs can now be used in Bayesian inference to infer the probabilities of reject root causes.

4.4. Inference

Once the parameters of the BN models are learned, Bayesian inference then needs to be performed to obtain the root-cause probabilities

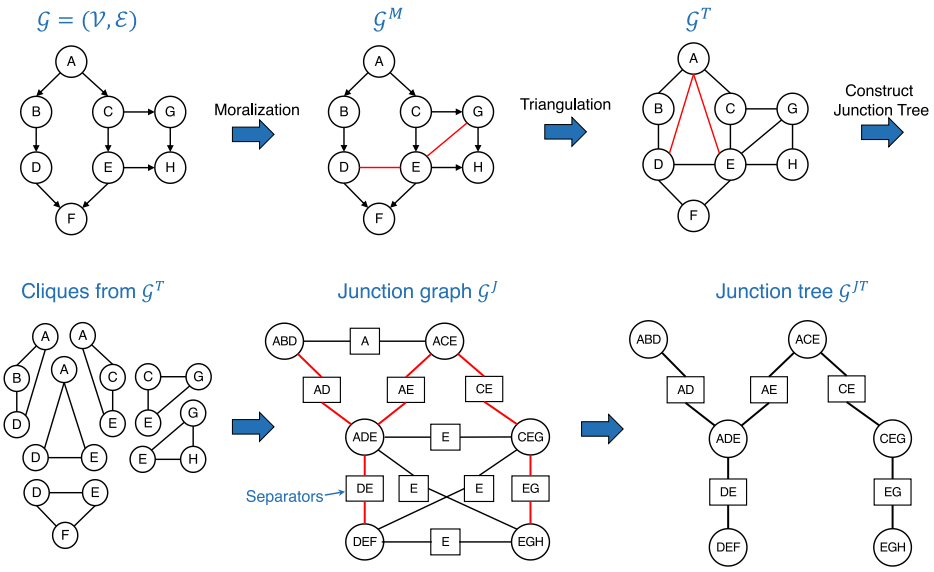


Fig. 3. Junction tree algorithm for converting learned graph to tree structure.

from the learned BN models. The purpose of a Bayesian network is to enable the efficient computation of updated probability distributions for a set of events in the Bayesian network, given the evidence of the newly observed cases. If our Bayesian network is a tree structure, then Belief Propagation (BP) can be applied to infer the probability of interest automatically. However, most learned structures are more complicated than trees. Therefore, Junction Tree (JT) algorithm needs to be implemented in our study to infer the probability of each reject cause from the complex structures that are learned previously. The core idea of the junction tree algorithm is to turn a graph into a tree of clusters amenable to Belief Propagation. We start with a Bayesian structure with its corresponding parameters learned from above, and then undergo the following steps for a JT inference: (a) Moralize the graph (b) Triangulate the graph (c) Build a junction tree (d) Apply Belief Propagation. For presentation purposes, a simplified Bayesian network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is used for demonstrating the process of JT algorithm, as shown in Fig. 3.

(a) Moralization connects each node's parents and drops the directionality of the arcs to allow a uniform treatment of directed and undirected graphs. (b) Triangulation adds chords into the moral graph $\mathcal{G}^M$ such that any cycle of more than three vertices short in the graph is cut short. (c) Junction tree is built by forming a maximal spanning tree from the cliques. When two cliques intersect, they are joined in the junction graph by an edge labeled with their intersection, called separators. A junction tree $\mathcal{G}^{JT}$ is then extracted from the junction graph such that the tree contains all the cliques (spanning tree) and satisfies the running intersection property: For each pair of clusters $c^{(i)}$, $c^{(j)}$, every cluster on the path between $c^{(i)}$, $c^{(j)}$ contains $v_c^{(i)} \cap v_c^{(j)}$.

Finally, the probability distributions of the cliques (nodes) and separators (edge labels) in the junction tree $\mathcal{G}^{JT}$ need to be transferred from the conditional probability distribution of the original Bayesian network $\mathcal{G}$ using potentials. The process is carried out as follows. For each (conditional) distribution from the BN, create a node potential:

$$P(v_i|par(v_i)) \Rightarrow \phi_i(v_i, par(v_i)), \tag{4}$$

where $par(v_i)$ is the parent of node v_i, ϕ_i signifies the potential between the nodes. Assign each node potential to its associated clique C, and compute the clique potential ϕ_C for C as the product of its assigned node potentials:

$$\phi_C = \prod_{\phi_{v_i}} \phi_{v_i}, \tag{5}$$

such that $\{v_i\} \cup par(v_i) \subseteq v_C$.

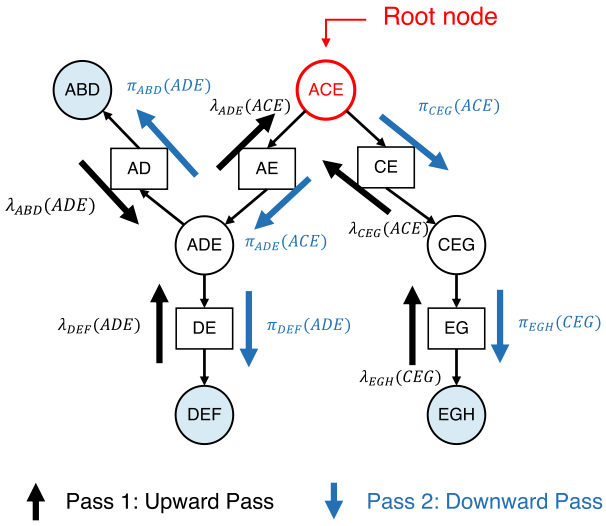


Fig. 4. Belief propagation graph for Bayesian Network inference.

(d) Belief Propagation, also known as sum–product message passing, is a message-passing algorithm for inference on tree-like structures. Since the learned Bayesian networks have been converted into tree structures by JT algorithms, BP can be used to infer the probability of each reject reason from the junction trees, conditional on any observed nodes with the external evidence. In addition, belief propagation is a generalization of the Forward–Backward method and consists of two passes with Pass 1 (Upward Pass) from the leaf nodes to the root node, and Pass 2 (Downward Pass) from the root node to the leaf nodes, as shown in Fig. 4.

Pass 1 takes the evidence at node v_i and computes the message $\lambda(v_i)$ from its child node if there is any. Then the message is passed upwards to its parent node with matrix multiplication on the conditional probability matrix M_{v_i}, generating information $\lambda_{v_i}(par(v_i))$ until it reaches the root node, following the logic in Fig. 5. On the other hand, Pass 2 starts in the opposite direction. It takes the prior distribution of the current node v_i and the message from its parent nodes to compute the message $\pi(v_i)$ to pass it downwards to its child nodes. Since the node v_i has already captured both messages from its children and parents, the belief probability $BEL(v_i)$ can be calculated by combining these messages,

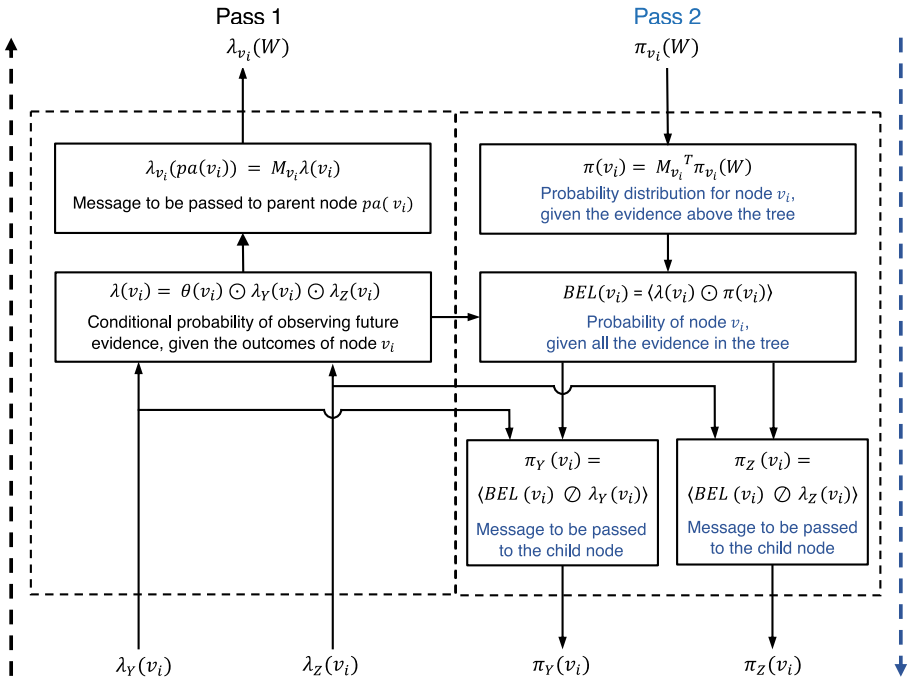


Fig. 5. Messages passed in pass 1 and pass 2 in belief propagation.

$\lambda(v_i)$ and $\pi(v_i)$. $BEL(v_i)$ encapsulates the conditional distribution for a node v_i, given all the associated evidence in the network. The message containing the belief probabilities keeps being passed down to its child nodes with the elimination of duplicate information to update the belief of its child nodes. Once it reaches all the leaf nodes, all the belief probabilities $BEL(v_i)$ in the network have been inferred.

When the inferred variable is a root-cause variable (i.e., $v_i \Rightarrow Y_c$), $BEL(v_i)$ expresses the probability of a root cause Y_c leading to quality failures, given all the evidence in the network, which is our probability of interest for Q2. If the variable is the quality risk indicator $v_i \Rightarrow R$, the belief probability $BEL(v_i)$ encodes the probability of the quality risk indicator R being "1", implying whether the job will have quality issues or not. If the probability exceeds a predefined classification threshold, the job is predicted to be problematic with quality issues. Otherwise, the job is projected to be normal. In this way, Q3 can also be solved. In our case, instead of using predictions from a single BN model to answer Q2 and Q3, we fuse the predictions from a library of BN models into aggregated results using WAEL to answer the research questions.

4.5. Weighted average ensemble learning

To provide a robust and accurate RCA solution, we fuse the predictions from different BN models, using WAEL. WAEL is a voting ensemble method that combines the predictions from multiple models by taking the weighted sum of the predictions for regression models or selecting the class with the largest weighted sum of predicted probabilities for classification models. As an important factor, the choice of weight needs to reflect each model's skill. We chose the weight based on their robustness, which is reflected by their likelihood of being the best classifier. The likelihood is computed as the number of times that a BN model outputs the most accurate prediction (optimal frequency) in ratio to the total number of times that a prediction has been performed (total frequency), as shown in:

$$w_{\mathcal{G}_{m,k}} = f_k / \sum_{i=1}^{K} f_i, \tag{6}$$

where $w_{\mathcal{G}_{m,k}}$ is the weight assigned to Bayesian learning algorithm A_k for Product Pd_m, f_k is the optimal frequency for algorithm A_k, and K is the total number of BN learning algorithms. The optimal frequency of each model can be obtained from experiments on all the BN models learned by different structure learning algorithms and knowledge sources on the entire dataset.

Subsequently, the WAEL technique follows Algorithm 2 to obtain the fused predictions for Ensemble BN. It first gathers the BN structures learned by different learners. Then the predictions are performed through parameter learning and inference on the collected distinct structures. After that, each prediction is weighted according to (6). Lastly, the weighted predictions are summed up to obtain the aggregated results. In this way, WAEL combines the predictions from distinct BN models into one set of root-cause probabilities and a single risk prediction for a job.

4.6. Answering the RCA questions

Now since the predictions from different BN models have been fused into a robust aggregated solution, the three RCA questions raised in 3.1 can be answered. For Q1, the casual relationships have been discovered from the historical production data. For Q2, the parameters of the BN models are learned. For a given problematic job, we can input its job features into the learned BNs for inference. By following the procedures of the junction tree algorithm, the probabilities of potential reject causes can be inferred from different BNs. Then the root-cause probabilities predicted from different BN models are fused into an aggregated set of probabilities using WAEL to output a robust solution.

Algorithm 2: Bayesian Network Prediction with Weighted Average Ensemble Learning

Initialise Testing data for product Pd_m, D_m;
Initialise $\{\mathcal{G}_{m,1}, \ldots, \mathcal{G}_{m,K}\}$ for Pd_m;
Initialise weights of each bagged model, $\{w_{\mathcal{G}_{m,1}}, \ldots, w_{\mathcal{G}_{m,K}}\}$;
for *each* $\mathcal{G}_{m,k}$ *in* $\{\mathcal{G}_{m,1}, \ldots, \mathcal{G}_{m,K}\}$ **do**
  $P_{m,k} = \mathcal{G}_{m,k}(D_m)$
end
$P_m = \sum_{k=1}^{K} w_{\mathcal{G}_{m,k}} P_{m,k}$

Table 2
Dataset variables.

Variable name	Variable type	Data description
jobRun	Feature	The number of operations performed for a job
jobStartTime	Feature	The timestamp the job starts
Tool	Feature	The tool a job runs with
Equip	Feature	The machine a job runs on
jobStartUser	Feature	The operator who starts the job
jobStopUser	Feature	The operator who stops the job
ProductionTime	Feature	The time it takes to complete a job
DownTime	Feature	The downtime rate of a job
CycleTime	Feature	The cycle time of a job
jobPLCSetupTime	Feature	The time it takes to set up a job
RejWeight_kg	Feature	The weight of rejected products in kg
rrnDescription	Root cause	The names of the potential reject reasons
isRejectFail	Observation	Binary; 1 indicates a job with quality issues, 0 otherwise

For Q3, the job features of a future job are input into various BN models to predict the likelihood of it being a problematic job.

5. Experimental results

This section presents an experimental study on a real-world factory to validate our solution performance, with discussions on the dataset, performance evaluation metrics, and comparative results.

5.1. Dataset

The case company is a plastic manufacturer specializing in producing drug packages. In their production, a combination of manual data logging and factory monitoring system tracks all the production processes for each production batch. Information such as machines, raw material, operators, production parameters, and job quality signals are tracked. Our dataset includes all this information for 6791 production jobs across 199 product types. Table 2 lists all the tracked parameters in the raw dataset. Note that the BN topology developed based on this dataset is presented in Fig. 1, reflecting the complex interdependencies among variables detailed in Table 2. This visualization serves as a foundational component of our methodology, illustrating how each variable within the network is intricately connected to facilitate a comprehensive understanding of the manufacturing processes.

5.2. Evaluation metrics

We design the below evaluation metrics for assessing our model's performance for predicting the probabilities of the reject causes for a given job and classifying the quality risk of scheduled future jobs.

5.2.1. Evaluation methods for root cause prediction

Q2 is to identify a list of potential reject causes and their corresponding probabilities for a job with quality issues. This means that the predicted result of an instance will be a sequence of probabilities. The ordering of the root causes in the sequence is important, indicating which reject causes are the dominant reasons. Therefore, both the accuracy of each predicted probability and their ranking in the sequence need to be evaluated to have a comprehensive judgment on the predicted lists of probabilities. Therefore, our prediction error metric includes Mean Absolute Error (MAE) and ranking error. MAE is to quantify the difference between the predicted and observed probabilities for a group of root causes on average, and ranking error is to identify the ranking difference between prediction and observation as sequences. MAE is calculated by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{7}$$

where $\hat{y}_i$ is the predicted probability for root cause i in the sequence, y_i is the observed probability, and n is the length of the probability list (i.e., the number of identified root causes in the list for a job).

Before computing the ranking error of the predicted sequence, all the noises in the probabilities need to be removed so that the ranking metric will not be oversensitive toward considerably small probabilities. This is achieved by truncating the probabilities to 2 significant figures. Then, each probability in the list is assigned a rank according to the magnitude of the predicted probabilities. The ranking difference between the observed and predicted sequences can be quantified. Eventually, the ranking difference in ratio to the possible maximum ranking difference is computed as ranking error. The ranking error can be obtained via:

$$\varepsilon_{rank} = \begin{cases} \frac{\sum_{i=1}^{k} |\widehat{rank}_i - rank_i|}{k^2/2}, & \text{if } k \text{ is even} \\ \frac{\sum_{i=1}^{k} |\widehat{rank}_i - rank_i|}{k^2/2}, & \text{otherwise;} \end{cases} \tag{8}$$

where k is the length of the predicted list. Then, MAE and ranking error are combined by a weighted sum to comprise the prediction error, as in (9). The parameter w can be used to interpolate between the accuracy of predicted root cause probabilities and their ranking in sequence. In this study, we set $w = 0.5$.

$$\varepsilon = w \times MAE + (1 - w) \times \varepsilon_{rank} \tag{9}$$

5.2.2. Evaluation methods for risk prediction

Q3 is a classification problem where the jobs with quality issues must be predicted before execution. We adopt common classification problem metrics, i.e., accuracy, sensitivity, and specificity together.

In manufacturing, we are more concerned with Type I Error (False Positives). Since the system should not distract the staff often with a false alarm, it needs to guarantee all the alarms that went off are correct and worth taking note of. Otherwise, the alert system will lose its credibility among the factory workers. As a result, the evaluation system will emphasize False Positive Rate (FPR), which is the proportion of identified positives (i.e., jobs predicted to be problematic) among the normal jobs. This is also defined as 1-specificity. The Receiver Operating Characteristics (ROC) curve provides a good way to visualize the true positive rate (or sensitivity on the y-axis) against the false positive rate (or "1-specificity" on the x-axis). It also gives a picture of the classifier's performance across all possible probability thresholds. Additionally, the Area Under the ROC Curve (AUC) provides an aggregate measure of performance across the whole spectrum of classification thresholds. One way of interpreting AUC is the ability of a classification model to distinguish 1s from 0s. Lastly, the robustness of different algorithms will be assessed by the likelihood of generating the worst predictions.

5.3. Data preparation and experimental setup

It is well known that correlated features in regression analysis can lead to an inflation of Type I Error, whereas such an issue is inclined to persevere in BN [38]. Therefore, we carried out a pairwise correlation test against all the features to eliminate the highly

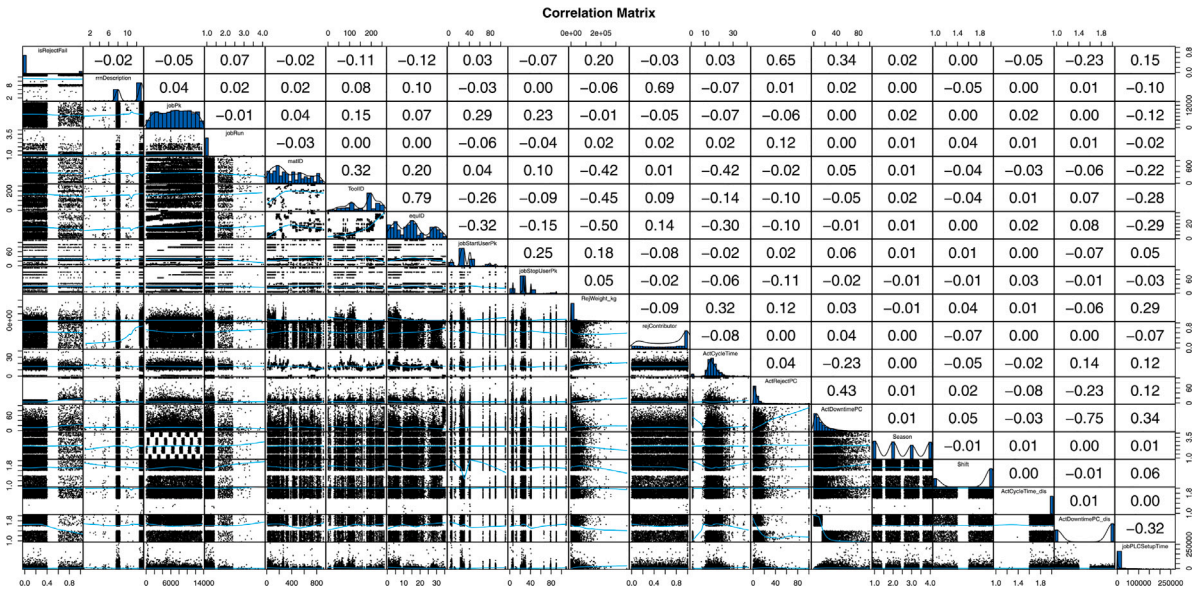


Fig. 6. Correlation matrix of all job features.

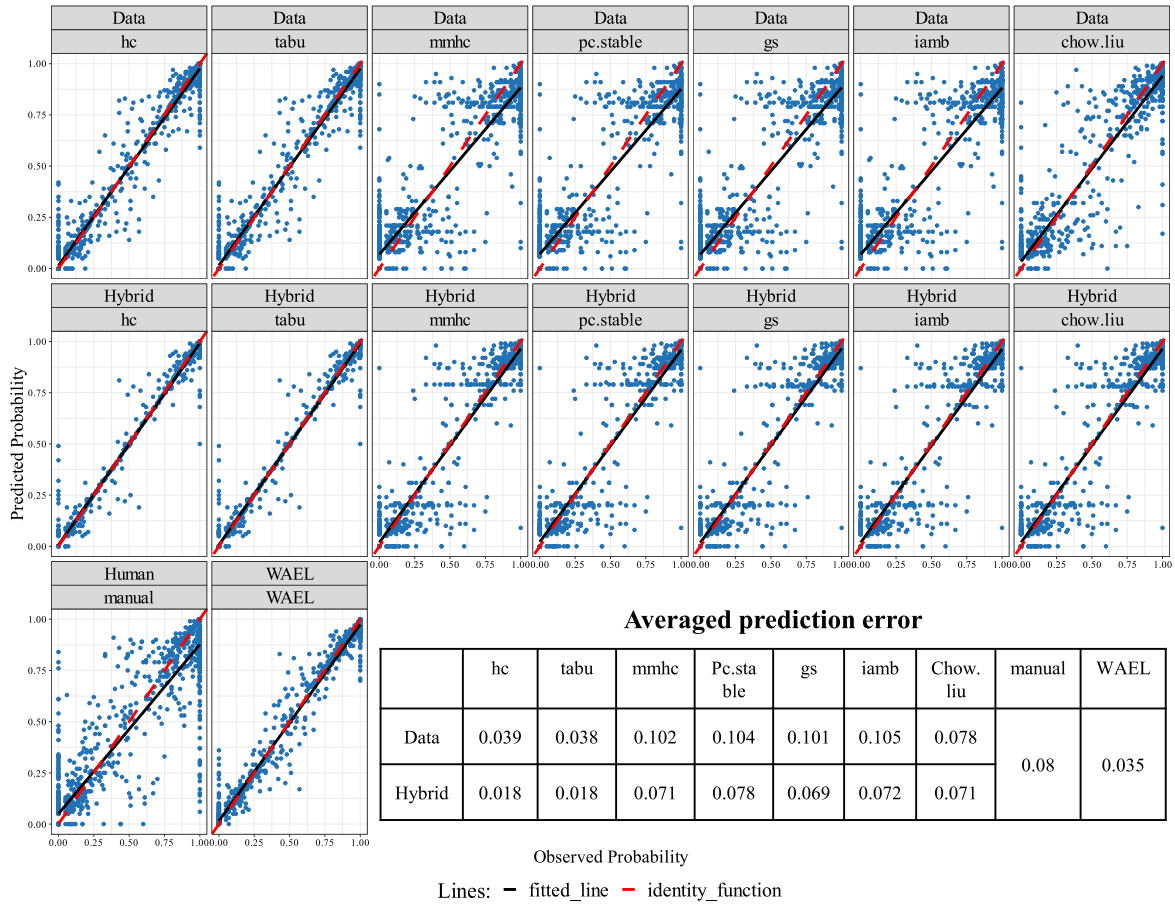
correlated feature "Tool". Fig. 6 presents the correlation matrix of all the 19 job features (shown in the diagonal, including: isRejectFail, rrnDescription, jobPk, jobRun, matID, ToolID, equID, jobStartUserPk, JobStopUserPk, RejWeight, rejContributor, ActCycleTime, ActRejectPC, ActDowntimePC, Season, Shift, ActCycleTime, ActDowntimePC, jobPLCSetupTime), which includes both numerical correlation coefficients and visual scatter plots. The upper right half of the matrix displays the correlation coefficients, ranging from −1 to 1, indicating the strength and direction of the linear relationships between pairs of features. A strong positive correlation (values near 1) suggests that as one feature increases, the other tends to increase as well; whereas a strong negative correlation (values near -1) indicates that one feature tends to decrease as the other increases. Values near 0 imply little to no linear correlation. Conversely, the lower left half of the matrix shows scatter plots for each pair of features, providing a visual insight into the nature of their relationships. These plots help to identify not only linear but also potential non-linear relationships or distributions without clear trends, which are critical for structuring our BN in root cause analysis, ensuring that it captures the true interactions and dependencies between variables effectively. In the case study, we performed 10-fold cross validation on the 6791 data samples with 90% for training BN algorithms and 10% for testing RCA and defect risk prediction, for each of the folds across 199 product types. When calculating the worst model frequency, the tied models are counted multiple times for each sample which causes the total to be greater than the sample size.

5.4. Predicted probabilities of reject causes for RCA

This section shows the results for question Q2, predicting the probabilities of each root cause for each job. The ground truth for defect causes in this work was established through a collaborative effort with industry experts, who brought extensive practical experience to the process. This process involved detailed discussions where each potential defect instance was thoroughly validated and categorized, ensuring the reliability and applicability of our ground truth. Such a method not only strengthens the validity of our analysis but also ensures that our BN model is grounded in accurately categorized real-world data. Fig. 7 shows the predicted probabilities of reject causes against the ground truth for all the tested knowledge sources and structure learning algorithms. The table in Fig. 7 lists the Averaged prediction error of all tested methods. While hc and tabu methods exhibit slightly lower prediction error of 0.018, our WAEL method shows the second lowest prediction error of 0.035. Notably, the WAEL method does not necessarily achieve the lowest error but distinguishes itself significantly in other crucial aspects. The clustering of probabilities at the spectrum's ends, particularly noted in the compared structure learning methods, indicates a strong model confidence in cases where defects are either very likely or unlikely. This polarization suggests that those methods are effectively distinguishing clear cases of defects and non-defects. Conversely, the more evenly distributed clustering observed in our WAEL method is attributed to its integration of ensemble techniques and diverse knowledge sources, which moderate model predictions to reflect a wider range of probabilities. This distribution pattern, especially closer alignment to the identity function (red dashed line), indicates higher prediction accuracy as it indicates that predictions closely match actual observations. The visual proximity of points to the identity function shown in our WAEL approach highlights its effective calibration in predicting probabilities that mirror real-world outcomes. This visualization not only supports the robustness and accuracy of WAEL but also highlights the beneficial impact of our ensemble learning strategy in providing a balanced predictive performance across the spectrum of probabilities. Such balanced outputs are crucial in RCA, where understanding the likelihood of defect causes can significantly aid in decision-making and prioritization of corrective actions.

By shifting the focus onto the effect of different knowledge sources on RCA performance (Table 3), we found a distinctive favor on the models learned with hybrid knowledge as their mean prediction error is significantly lower than models with other knowledge sources. This aligns with the expectation that extra knowledge provides insights and guidance to structure learning so resulting in a better prediction. However, BN constructed purely from expert knowledge exhibited poorer predictive performance compared to those developed from hybrid knowledge. This finding contrasts with the conventional expectation that more expert guidance would inherently lead to better predictions. The root of this discrepancy lies in the inherent design limitations of networks created solely based on human expertise. Such networks often rely on established theories and experiences and tend to miss critical features and causal relationships that are only evident through detailed data analysis. This rigidity restricts the network's ability to learn from new data, limiting both parameters learning and the inference processes. Therefore, the performance of these networks is often poorer than expected.

Conversely, hybrid models, which blend empirical data with expert insights, demonstrate better performance in RCA. These models mine



Averaged prediction error

	hc	tabu	mmhc	Pc.stable	gs	iamb	Chow.liu	manual	WAEL
Data	0.039	0.038	0.102	0.104	0.101	0.105	0.078	0.08	0.035
Hybrid	0.018	0.018	0.071	0.078	0.069	0.072	0.071		

Fig. 7. Faceted scatter plots of predicted vs observed probabilities for different reject reasons by distinct knowledge sources and structure learning methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison of root cause analysis over different structure learning methods and different knowledge sources on testing dataset.

Structure learning method	Knowledge source	Avg prediction error	Worst model frequency
hc	Data	0.039	440
hc	Hybrid	**0.018**	535
tabu	Data	0.038	433
tabu	Hybrid	0.018	534
mmhc	Data	0.102	878
mmhc	Hybrid	0.071	912
pc.stable	Data	0.104	740
pc.stable	Hybrid	0.078	838
gs	Data	0.101	912
gs	Hybrid	0.069	999
iamb	Data	0.105	857
iamb	Hybrid	0.072	944
chow.liu	Data	0.078	732
chow.liu	Hybrid	0.071	944
manual	Human	0.080	970
WAEL	WAEL	0.035	**0**

causal relationships primarily from the data, while still incorporating crucial expert knowledge. This approach allows hybrid BNs to maintain the interpretability of human insights while also adapting to new information and complex data interactions, thus leading to improved predictive accuracy and robustness.

Residuals are also inspected to check if the model is appropriate and trustworthy for the data. They are the estimates of experimental error obtained by subtracting the observed probabilities from the predicted probabilities for the root cause. Fig. 8 illustrates the faceted residual histogram plot for different structure learning methods. From the plot, we can see that the overall patterns of the residuals for all the models approach a bell shape, signifying a normally distributed variance. Thus, the normality assumption is likely to be true.

To assess the robustness of the models, we counted the occurrence of each model generating the worst prediction for each job in the testing dataset. Table 3 shows that the proposed WAEL method is the most stable technique among all the algorithms. It has never appeared to be the worst model for any job instance. This is mainly due to its voting nature of putting more weight on the more stable algorithms. This weighting method alleviates the risk of the existing deficiencies in accuracy and stability in stand-alone algorithms. As a result, it achieves our goal of providing a robust RCA model.

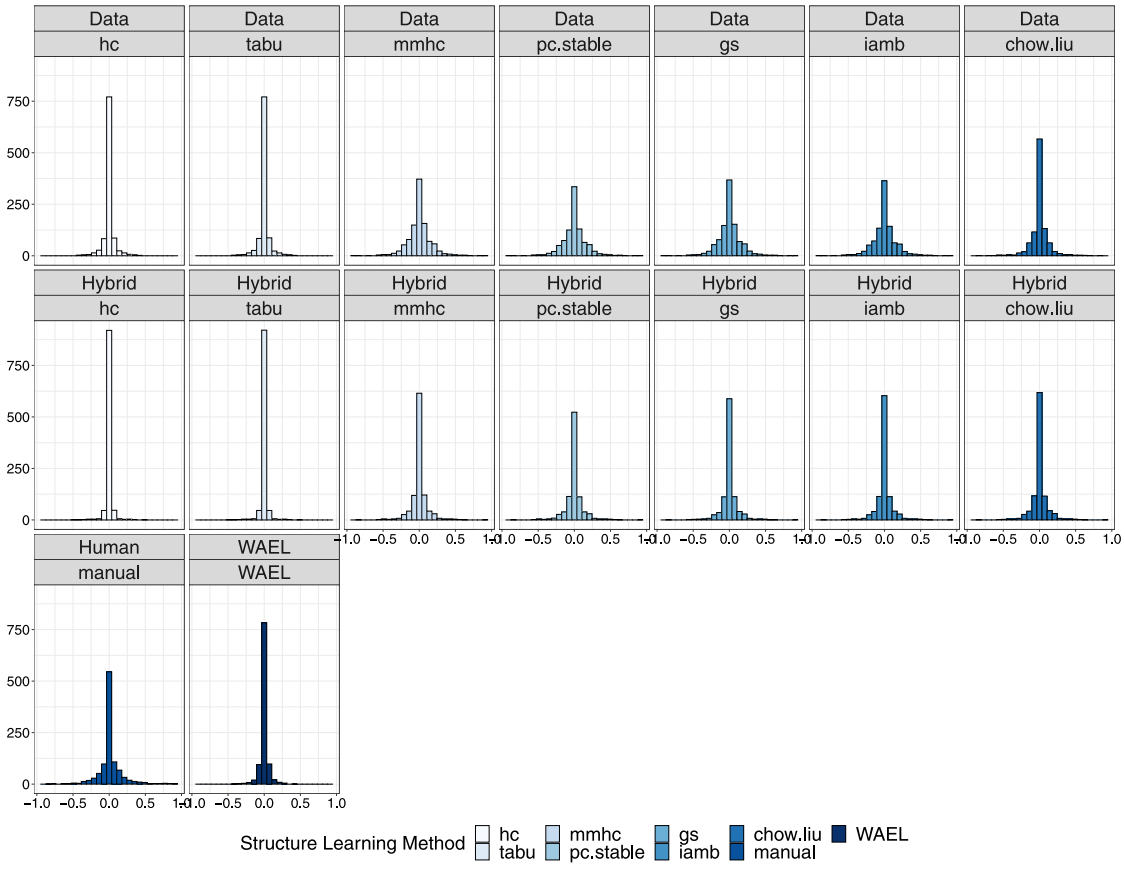


Fig. 8. Residual histogram plot for RCA over different structure learning methods from different knowledge sources.

Table 4
Comparison of job risk prediction over different structure learning methods and different knowledge sources on testing dataset.

Structure learning method	Knowledge source	Prediction accuracy	Worst model frequency
hc	Data	0.97	225
hc	Hybrid	0.97	227
tabu	Data	0.97	216
tabu	Hybrid	0.97	212
mmhc	Data	0.92	536
mmhc	Hybrid	0.92	552
pc.stable	Data	0.89	613
pc.stable	Hybrid	0.88	635
gs	Data	0.94	371
gs	Hybrid	0.93	482
iamb	Data	0.94	384
iamb	Hybrid	0.92	491
chow.liu	Data	0.89	714
chow.liu	Hybrid	0.89	718
manual	Human	0.87	873
WAEL	WAEL	**0.97**	**140**

5.5. Defect risk prediction

This section shows the results for question Q3, predicting whether a scheduled job will be a risk in the aspect of product quality. Table 4 shows the accuracy of defect classification using different structure learning algorithms based on different knowledge sources. All methods exhibit a high prediction accuracy, among which WAEL is the highest. The overall accurate predictions of all the algorithms also contribute to the excellent performance of the WAEL method as it applies the weighting to the probabilities of the predicted class obtained by various algorithms and determines the riskiness of the job using the weighted sum of the likelihood against the classification threshold of 0.5.

The ROC curve (Fig. 9) reveals the performance of Bayesian network classifiers developed by different structure learning methods and knowledge sources at all discrimination thresholds. Similar to the outcome from the accuracy chart, the networks built solely from data bring about more excelling performance than other knowledge sources, except for score-based algorithms. In addition, tabu search and hill-climbing continue to dominate the model performance amid all structure learning methods across all possible classification thresholds with an AUC of 0.994. Whilst human-built model leads to poor performance at an AUC of 0.672. Peculiarly, WAEL rises a relatively large AUC of 0.988 following the score-based algorithms. This hints that WAEL is an accurate and robust classifier across a wide range of classification cut-offs.

Surprisingly, hybrid knowledge source and data knowledge contribute similar prediction accuracy. Such a phenomenon occurs possibly

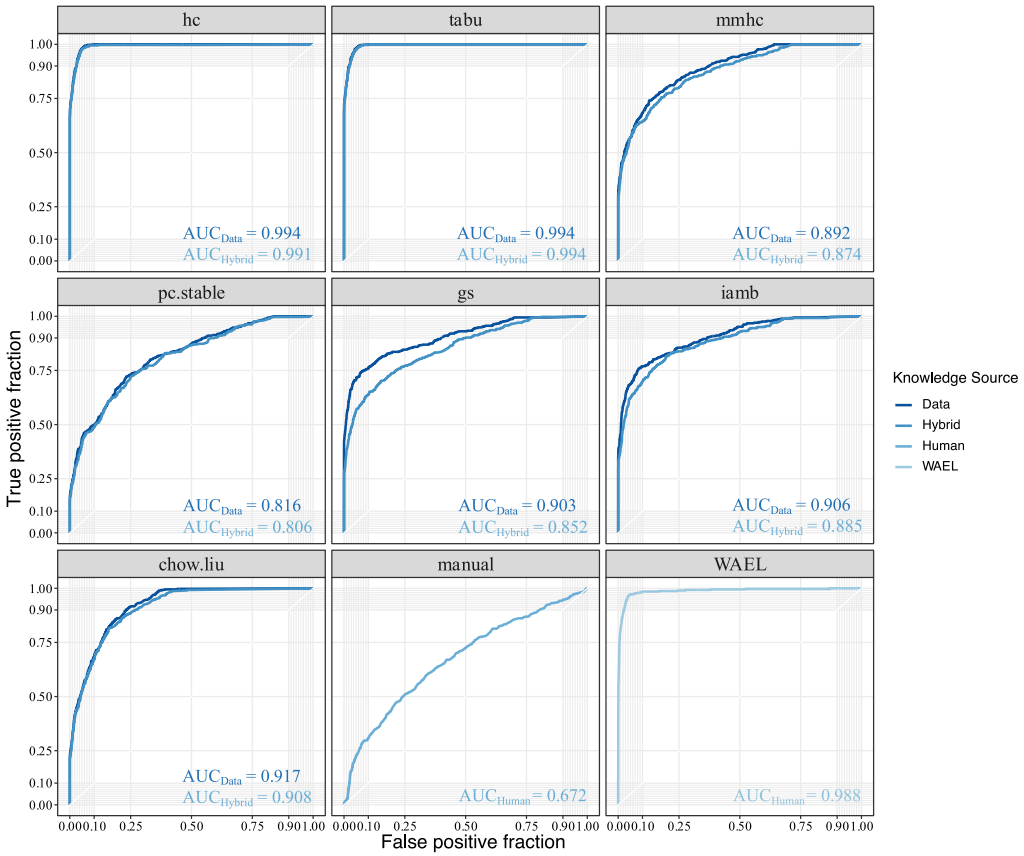


Fig. 9. ROC over different structure learning methods from different knowledge sources for risk prediction.

because of the lack of evidence during inferencing for the nodes in the added knowledge substructure (prior knowledge) from the hybrid knowledge source. In the case of risk prediction, the jobs are yet to be run, so there will be no external information introduced to the observational variables, such as the occurrence of reject cause reasons. This makes the added links (i.e., knowledge) from the reject cause nodes to job failure nodes redundant in the hybrid models.

To assess the robustness of the models, the models of interest are put through vigorous testing across the data sample. Here, the model robustness is reflected in the number of times each model produces an incorrect classification. As shown in Table 4, WAEL remains the most robust model.

6. Conclusions

This paper addresses the problem of root cause analysis of product quality from historical production data. We proposed a product-wise Ensemble Bayesian Network to learn the causal relationship between job features, root causes and defect signals from big unstructured historical production data, with high accuracy and robustness while providing human-interpretable probabilistic reasoning capabilities. Our methods overcame the lack of robustness, sensitivity to data distribution, and ignorance of prior knowledge in existing Bayesian network-based work, using ensemble learning techniques. From comprehensive experiments on a sample of 6791 real-world production jobs from a case company, it has been proven that our model exhibits sound accuracy and robustness amidst all the models in inferring the probabilities of root causes for problematic jobs and predicting future high-risk jobs. In summary, our work provides a robust, accurate, and interpretable probabilistic reasoning method for RCA to support manufacturers with data-driven decision-making under the circumstance of quality failures.

In the context of our case study at a plastic manufacturer, the implementation of the developed EBN model has shown significant practical impacts. The integration of this model into the manufacturer's daily operations has facilitated a more nuanced understanding of the production processes, particularly in identifying and addressing root causes of defects. The use of the EBN, developed from both expert knowledge and empirical data, has enabled the manufacturer to more precisely predict potential failures and inefficiencies. As a result, proactive measures can be taken to mitigate issues before they escalate into more significant problems. This proactive approach has not only reduced downtime but also improved the overall quality of products by decreasing the occurrence of defects. Moreover, the insights gained from the model have led to more informed decision-making processes at the company. For instance, the data-driven nature of the model provides clear indicators of which aspects of the production process are most likely to benefit from adjustments or upgrades. This has guided the manufacturer in allocating resources more effectively, targeting areas that yield the highest returns in terms of quality improvement and cost savings. The successful implementation of this model at the plastic manufacturing site serves as a testament to the potential of advanced data analytics in industrial settings. It highlights how integrating sophisticated data analytics techniques with traditional manufacturing processes can significantly enhance operational efficiency and product quality.

Future work can be focused on introducing additional sensors to enrich more direct quality signals, thereby enhancing the comprehensiveness of knowledge incorporated into our model. Currently, our sensors primarily capture basic operational metrics which may not detect subtle anomalies that may cause defects. Implementing additional sensors such as vibration and acoustic emission sensors could allow for earlier detection of potential machine/process failures, thus significantly enhance our model's sensitivity. These enhancements would enable our BN to more accurately model complex interactions between variables, leading to more effective root cause analysis and predictive

maintenance strategies. More delicate parameter calibration models can also be explored to fine-tune the parameters for BN and ensemble learning techniques such as the resampling time of bagged BN and the weights of WAEL.

CRediT authorship contribution statement

Karen Wang: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Chao Liu:** Data curation, Writing – review & editing. **Yuqian Lu:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

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

The authors do not have permission to share data.

Acknowledgments

This work was supported by Callaghan Innovation R&D Fellowship Grant, New Zealand. The authors would like to thank Aspect Productivity Technology colleagues, Bob Dedekind for sharing domain knowledge and arranging client visits, Chris Rauch for providing data processing support. Special gratitude also goes to the anonymous case company for sharing data and expert knowledge.

References

[1] Rokach L, Hutter D. Automatic discovery of the root causes for quality drift in high dimensionality manufacturing processes. J Intell Manuf 2012;23(5):1915–30.
[2] Yu J, Rashid MM. A novel dynamic Bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. AIChE J 2013;59(7):2348–65.
[3] Dey S, Stori J. A Bayesian network approach to root cause diagnosis of process variations. Int J Mach Tools Manuf 2005;45(1):75–91.
[4] Alaeddini A, Dogan I. Using Bayesian networks for root cause analysis in statistical process control. Expert Syst Appl 2011;38(9):11230–43.
[5] Marcot BG, Penman TD. Advances in Bayesian network modelling: Integration of modelling technologies. Environ Model Software 2019;111:386–93.
[6] Weidl G, Madsen AL, Israelson S. Applications of object-oriented Bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes. Comput Chem Eng 2005;29(9):1996–2009.
[7] Lokrantz A, Gustavsson E, Jirstrand M. Root cause analysis of failures and quality deviations in manufacturing using machine learning. Procedia Cirp 2018;72:1057–62.
[8] Leonhardt V, Claus F, Garth C. PEN: Process estimator neural network for root cause analysis using graph convolution. J Manuf Syst 2022;62:886–902.
[9] Baier L, Frommherz J, Nöth E, Donhauser T, Schuderer P, Franke J. Identifying failure root causes by visualizing parameter interdependencies with spectrograms. J Manuf Syst 2019;53:11–7.
[10] Ito A, Hagström M, Bokrantz J, Skoogh A, Nawcki M, Gandhi K, et al. Improved root cause analysis supporting resilient production systems. J Manuf Syst 2022;64:468–78.
[11] e Oliveira E, Miguéis VL, Borges JL. Automatic root cause analysis in manufacturing: An overview & conceptualization. J Intell Manuf 2023;34(5):2061–78.
[12] Chiang LH, Russell EL, Braatz RD. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. Chemometrics Intell Lab Syst 2000;50(2):243–52.
[13] Tsung F-g. Statistical monitoring and diagnosis of automatic controlled processes using dynamic PCA. Int J Prod Res 2000;38(3):625–37.
[14] Raich A, Cinar A. Multivariate statistical methods for monitoring continuous processes: Assessment of discrimination power of disturbance models and diagnosis of multiple disturbances. Chemometrics Intell Lab Syst 1995;30(1):37–48.
[15] AlGhazzawi A, Lennox B. Monitoring a complex refining process using multivariate statistics. Control Eng Pract 2008;16(3):294–307.
[16] Chiang LH, Kotanchek ME, Kordon AK. Fault diagnosis based on Fisher discriminant analysis and support vector machines. Comput Chem Eng 2004;28(8):1389–401.
[17] Yu J. A particle filter driven dynamic Gaussian mixture model approach for complex process monitoring and fault diagnosis. J Process Control 2012;22(4):778–88.
[18] Abdelrahman O, Keikhosrokiani P. Assembly line anomaly detection and root cause analysis using machine learning. IEEE Access 2020;8:189661–72. http://dx.doi.org/10.1109/ACCESS.2020.3029826.
[19] Wee YY, Cheah WP, Tan SC, Wee K. A method for root cause analysis with a Bayesian belief network and fuzzy cognitive map. Expert Syst Appl 2015;42(1):468–87.
[20] Papageorgiou K, Theodosiou T, Rapti A, Papageorgiou EI, Dimitriou N, Tzovaras D, et al. A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing. Front Manuf Technol 2022;2:972712.
[21] Detzner A, Rückschloß R, Eigner M. Root-cause analysis with interactive decision trees. In: 2020 24th international conference information visualisation. IV, IEEE; 2020, p. 322–7.
[22] Chen M, Zheng AX, Lloyd J, Jordan MI, Brewer E. Failure diagnosis using decision trees. In: International conference on autonomic computing, 2004. proceedings. IEEE; 2004, p. 36–43.
[23] Han T, Jiang D, Zhao Q, Wang L, Yin K. Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Trans Inst Meas Control 2018;40(8):2681–93.
[24] Yang J, Zhang Y, Zhu Y. Intelligent fault diagnosis of rolling element bearing based on svms and fractal dimension. Mech Syst Signal Process 2007;21(5):2012–24.
[25] Xu L, Chow M-Y. A classification approach for power distribution systems fault cause identification. IEEE Trans Power Syst 2006;21(1):53–60.
[26] Wang B, Li Y, Luo Y, Li X, Freiheit T. Early event detection in a deep-learning driven quality prediction model for ultrasonic welding. J Manuf Syst 2021;60:325–36.
[27] Chen C, Liu C, Wang T, Zhang A, Wu W, Cheng L. Compound fault diagnosis for industrial robots based on dual-transformer networks. J Manuf Syst 2023;66:163–78.
[28] Lee MY, Choi YJ, Lee GT, Choi J, Kim CO. Attention mechanism-based root cause analysis for semiconductor yield enhancement considering the order of manufacturing processes. IEEE Trans Semicond Manuf 2022;35(2):282–90.
[29] Cai B, Huang L, Xie M. Bayesian networks in fault diagnosis. IEEE Trans Ind Inform 2017;13(5):2227–40.
[30] Cai B, Kong X, Liu Y, Lin J, Yuan X, Xu H, et al. Application of Bayesian networks in reliability evaluation. IEEE Trans Ind Inf 2018;15(4):2146–57.
[31] Correa M, Bielza C, Pamies-Teixeira J. Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. Expert Syst Appl 2009;36(3):7270–9.
[32] Abele L, Anic M, Gutmann T, Folmer J, Kleinsteuber M, Vogel-Heuser B. Combining knowledge modeling and machine learning for alarm root cause analysis. IFAC Proc Vol 2013;46(9):1843–8.
[33] Kasper D, Weidl G, Dang T, Breuel G, Tamke A, Wedel A, et al. Object-oriented Bayesian networks for detection of lane change maneuvers. IEEE Intell Transp Syst Mag 2012;4(3):19–31.
[34] Kirchhof M, Haas K, Kornas T, Thiede S, Hirz M, Herrmann C. Root cause analysis in lithium-ion battery production with fmea-based large-scale bayesian network. 2020, arXiv preprint arXiv:2006.03610.
[35] Li M, Zhang R, Liu K. A new ensemble learning algorithm combined with causal analysis for bayesian network structural learning. Symmetry 2020;12(12):2054.
[36] Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdiscipl Rev: Data Min Knowl Discov 2018;8(4):e1249.
[37] Yu W, Zhao C. Online fault diagnosis for industrial processes with Bayesian network-based probabilistic ensemble learning strategy. IEEE Trans Autom Sci Eng 2019;16(4):1922–32.
[38] Bae H, Monti S, Montano M, Steinberg MH, Perls TT, Sebastiani P. Learning Bayesian networks from correlated data. Sci Rep 2016;6(1):1–14.