Computing and Informatics, Vol. 44, 2025, 800–827, doi: 10.31577/cai_2025_4_800

DATA-DRIVEN BAYESIAN NETWORK FOR RISK ANALYSIS OF TELECOM FRAUD

Binzhou Si, Haichun Sun*, Mengyuan Shao

College of Information and Network Security
People's Public Security University of China
Beijing, 102206, China
e-mail: 2022211421@stu.ppsuc.edu.cn, sunhaichun@ppsuc.edu.cn, 1972085006@qq.com

Abstract. Given the widespread occurrence of global telecom fraud, the development of proactive measures for crime prevention and control has become increasingly crucial. This study introduces a data-driven Bayesian Network (BN) model, which incorporates D-S evidence theory to integrate prior knowledge for fraud risk analysis. Through the examination of real-world case data, the study identifies key risk-influencing factors (RIFs) and uncovers causal relationships by comparatively evaluating three structure learning algorithms: Peter-Clark (PC), Bayesian Search (BS), and Greedy Thick Thinning (GTT). A robust Directed Acyclic Graph (DAG) is then constructed, and the Expectation-Maximization (EM) algorithm is employed to estimate conditional probability distributions. The proposed model effectively captures the causal relationships and nonlinear complexities among RIFs. To validate the model's applicability, scenario reasoning and sensitivity analysis are conducted, confirming its effectiveness in prioritizing RIFs and supporting informed decision-making. This research presents a novel and practical framework for public security agencies to develop proactive strategies for telecom fraud prevention and control.

Keywords: Telecom fraud, data-driven, Bayesian network, risk analysis

Mathematics Subject Classification 2010: 68U01

* Corresponding author

1 INTRODUCTION

With the rapid development of the digital economy and network technologies, telecom fraud has diversified into various forms [1], inflicting not only economic losses on victims but also posing severe threats to societal trust and the financial security system [2]. While some traditional methods for detecting telecom fraud have been relatively successful in identifying and preventing fraudulent activities, the exponential increase in fraud case data and the relentless evolution of fraudulent tactics present considerable challenges [3, 4, 5]. For example, fraudsters can even leverage artificial intelligence technologies, such as deepfakes, to create highly targeted fraud schemes tailored to the specific characteristics of victims, making it exceedingly difficult to prevent and control fraud.

Traditional risk assessment methods, such as rule-based systems and statistical models, have been employed to detect fraudulent activities. However, these methods often face challenges in keeping up with the fast-paced evolution of fraud tactics and the rapid increase in fraud-related data. Rule-based systems rely heavily on predefined patterns, making them ineffective in detecting novel or complex fraud schemes. Moreover, these methods are limited in capturing nonlinear dependencies and interactions among various risk-influencing factors (RIFs), which are crucial for understanding the fundamental causes of fraud [3].

In response to these challenges, machine learning (ML) and artificial intelligence (AI) technologies have become powerful tools for fraud detection [6]. Notably, deep learning models have demonstrated remarkable potential in processing large datasets and recognizing complex fraud patterns. Hu et al. [7] proposed a sparse graph fraud detection framework called BTG, which relies on reconstructing user behavior graphs. The process involves reconstructing the graph based on user behavior similarities, followed by employing graph ML techniques to identify fraudulent users, thus linking sparse graph data with graph ML. Wang et al. [8] developed a graph neural network that is aware of feature differences, effectively combating the challenges posed by imbalanced fraud datasets. Similarly, Raghavan and Gayar [9] explored the use of deep learning methods to detect fraud in high-dimensional datasets, highlighting their scalability and adaptability. However, despite the high prediction accuracy of ML models, they often function as "black-box" systems, providing limited interpretability. This opacity limits their application in domains like telecom fraud prevention, where understanding causal relationships is crucial for designing effective intervention strategies.

To overcome these limitations, Bayesian networks (BNs) have gained recognition thanks to their ability to model probabilistic dependencies and causal relationships among variables. As a graphical model, BNs provide a transparent framework for fraud risk analysis, making them ideal for understanding the complexity of fraud mechanisms. Byun and Song [10] demonstrated the utility of BNs in system reliability analysis, highlighting their ability to model interdependencies in structured data.

Hybrid models that integrate auxiliary techniques, such as Dempster-Shafer (D-S) evidence theory, have also demonstrated the potential in improving fraud detection. Liu and Li [11] explored the integration of D-S evidence theory into a credit card fraud detection model to address uncertainty in fraud data and achieve robust risk prediction. Yan et al. [12] introduced a cost-sensitive graph neural network model, aimed at improving detection performance and scalability while tackling data imbalance challenges. These hybrid approaches underscore the potential of combining different techniques to bridge the gap between interpretability and predictive performance.

Despite these advancements, the application of BNs and hybrid approaches in telecom fraud detection remains relatively underexplored. Many existing studies either focus on ML models with high predictive accuracy or rule-based methods with limited adaptability. This research aims to fill the gap by proposing a data-driven BN model for analyzing telecom fraud risk. This model combines various RIFs of fraud, assesses different structure learning algorithms, and establishes correlations among risk-influencing factors (RIFs) to identify and predict fraud risks precisely.

Key contributions of this paper include:

1. Introducing a novel data-driven BN framework for fraud risk analysis, which, for the first time, comprehensively considers the interactive effects of various RIFs on telecom fraud;
2. Skillfully incorporating background knowledge derived from D-S evidence theory into the hybrid model, thereby providing more accurate risk prevention and control measures;
3. Identifying key RIFs specific to different types of victimization through modeling and quantitative analysis, and proposing fraud prevention and intervention policies based on the results of the BN model analysis.

This paper is structured as follows: In Section 2, we examine the current landscape and difficulties in research related to telecom fraud and data-driven methods. Section 3 outlines the steps taken to construct a data-driven BN. In Section 4, we validate the model and interpret the results using fraud cases from City S. Section 5 highlights the model's benefits and puts forth suggestions for fraud prevention strategies. Lastly, Section 6 wraps up the paper.

2 RELATED WORKS

2.1 Studies of Telecom Fraud

The emergence of telecom fraud is a multifaceted phenomenon shaped by both internal and external factors, stemming from the interplay between the fraudulent scenarios devised by fraudsters and the intrinsic traits of the victims. A review of existing research indicates that various factors can influence individuals' susceptibility

to fraud, such as the victims' personality traits [13], the technological tactics employed by fraudsters [14], and the psychological conditions prevalent in society [15]. For instance, numerous studies from a demographic perspective consider aspects like gender, age, and other personal characteristics of victims as RIFs [16, 17]. Chen and Ma [18] processed a telecom fraud dataset using the Bert model and analyzed the potential connection between personality traits (Myers-Briggs Type Indicators) and a group of telecom fraud victims from over 20 000 data points. Ni and Yu [19] used a BN to model and predict fraud risk based on victim characteristics. However, a limitation of this approach is the narrow range of RIFs considered, and the causal relationships established in the BN primarily rely on expert judgement rather than explicitly outlining the specific methods and reasoning utilized by the experts.

In summary, existing literature on susceptibility to telecom fraud is primarily constrained by an absence of systematic mathematical analysis and statistical modeling. A significant portion of the research involves descriptive and theoretical analysis based on case study. Additionally, the focus frequently lies on individual types of fraud, specific victim characteristics, or single instances, with insufficient integration of qualitative and quantitative approaches. There is a notable scarcity of multifactorial and multi-indicator coupled analysis based on objective data, along with empirical studies.

2.2 Research on Data-Driven BN

A BN serves as a graphical reasoning model that adeptly captures uncertain knowledge and uncovers causal relationships among variables [20]. It uses conditional probability tables to assess these causal relationships, thus facilitating the analysis of complex interactions among various factors in telecom fraud. Typically, data-absent BN modeling relies on the intuition or expertise of specialists to construct the model, which includes the creation of conditional probability tables and the determination of parameters, potentially introducing uncertainty and bias [21]. In contrast, data-driven approaches enable the identification and extraction of a more objective and robust BN structure from available datasets. Table 1 presents a summary of relevant studies focused on deriving BN network structures through data-driven methods, including classic constraint-based Peter-Clark (PC) algorithms, score-based Greedy Thick Thinning (GTT) search, and Bayesian Search (BS) algorithms, as well as their corresponding data sizes and parameter learning methods.

In brief, when expert experience is unmanageable or when there is a lack of sufficient systemic knowledge for modeling, data-driven methods provide an alternative. These approaches can identify potential relationships among RIFs. To obtain a more sensible BN structure, we incorporated expert experience derived from D-S evidence theory before employing data-driven techniques, ensuring that certain background knowledge was included as mandatory relationships. Furthermore, we compared the causal models generated by three different structure learning algorithms to evaluate their performance.

No.	Data Source	Number of Data	Number of Nodes	Structure Learning	Parameter Learning	Model Validation	Resources
1	Experiment data	8 000	5	Search and scorebased method	Maximum Likelihood Estimate	\	[22]
2	China Maritime Safety Authority	590	10	BS	Bayesian Estimation	k-fold cross validation	[23]
3	Report and database	235	23	BS, GTT and PC	Expectation-Maximization	k-fold cross validation	[24]
4	Public dataset from organizations	414	20	BS	Bayesian Estimation	TAN	[25]
5	INFORM public dataset	191	21	GTT	\	k-fold cross validation	[26]
6	Flight data monitoring program	\	12	BS, GTT	Expectation-Maximization	k-fold cross validation	[27]

Table 1. Research on data-driven BN algorithm

3 METHODOLOGY

By exploring the potential interrelationships among telecom fraud RIFs, one can gain a dynamic understanding of how risks develop, thus preventing the occurrence of risk-related incidents [21]. In the realm of risk assessment, BNs are commonly seen as an effective method, offering statistically-based, interpretable predictions of risk [28]. A BN illustrates a collection of random variables along with their conditional dependencies using a directed acyclic graph, based on the premise that a set of variables can have their joint probability distribution represented as follows:

$$P\{x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid U_i). \tag{1}$$

In this context, $P\{x_1, x_2, \ldots x_n\}$ represents the joint probability distribution associated with the variable set $X = \{x_1, x_2, \ldots x_n\}$, while $U_i \subseteq \{x_1, x_2, \ldots x_n\}$ denotes the collection of parent nodes for the variable X_i. Additionally, $P(x_i \mid U_i)$ indicates the conditional probability of the variable X_i, given its parent nodes. Within a BN, each node represents a specific variable or a RIF that contributes to the system under study. Nodes are capable of assuming different states or values, and the directed edges between them illustrate causal relationships. Collectively, these nodes establish a framework that facilitates the examination of intricate interdependencies and their effects on the system.

The construction of data-driven BN structures is often constrained by imbalanced data (where positive and negative samples are not evenly distributed) [19]. Therefore, this paper proposed a BN model that integrates data-driven methods with prior background knowledge to automatically learn the correlations among telecom fraud RIFs. The model construction process is shown in Figure 1.

1. Step one involved collecting and processing data, which included gathering information on telecom fraud cases and performing data cleaning and integration to form the Telecom Fraud Case Dataset.

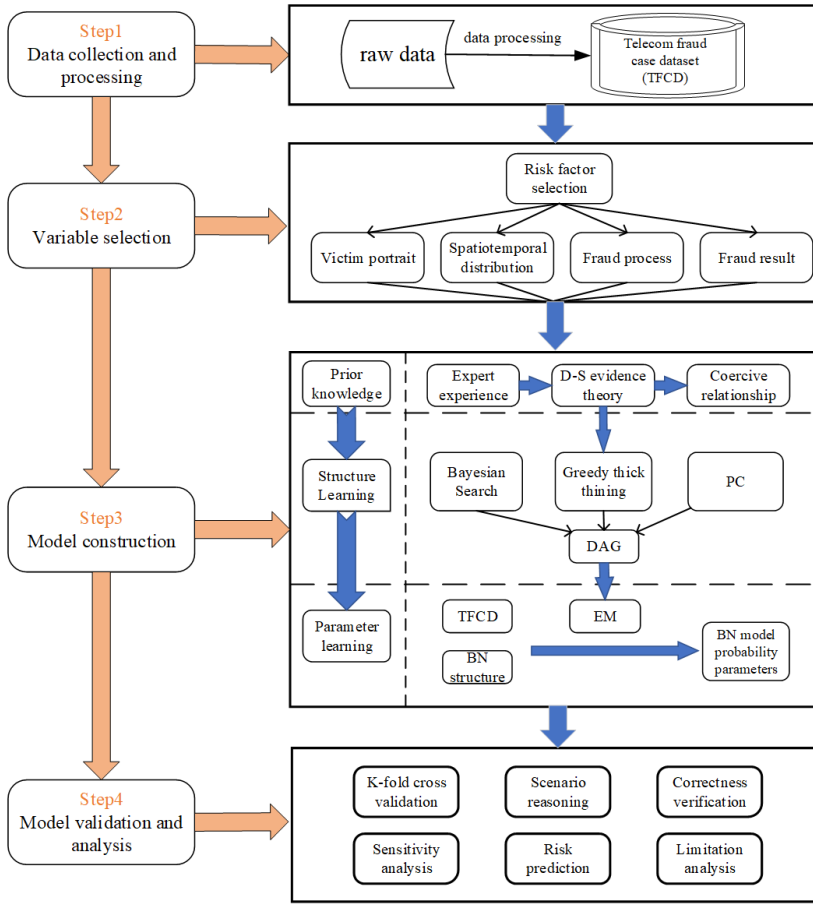


Figure 1. Data-driven BN analysis framework

2. Step two was about selecting model RIFs. By examining the telecom fraud process and combining the collected multisource data, we identified the model RIFs. Also, suitable definitions for node states were established based on the data distribution and the historical frequency of the RIFs.
3. Step three involved constructing the model. We started by using expert experience along with background knowledge from D-S evidence theory to outline mandatory relationships among the RIFs and define some conditional probabilities. Next, we employed three structure learning algorithms-BS, GTT, and PC-to establish a BN model for assessing telecom fraud risk susceptibility. Following this, we performed parameter learning with the EM algorithm to derive a BN model that included node conditional probability distributions. To improve accuracy, we integrated domain knowledge to refine the parameters.

4. Step four was model validation and analysis. The effectiveness of the telecom fraud BN model was evaluated using k-fold cross-validation, scenario reasoning, correctness testing, and sensitivity analysis. Relevant metrics were computed, and the sensitivity RIFs for various fraud types were examined.
5. Finally, the discussion focused on the data-driven BN risk assessment model, examining its benefits and suggesting relevant policies and intervention strategies for preventing telecom fraud.

3.1 Model Construction

Building the BN model involved two main steps: structure learning and parameter learning. Structure learning aims to identify an effective directed acyclic graph (DAG) structure, extracting optimal correlations among RIFs from sample data [29]. These correlations include dependencies and independencies among variables. Previous structure learning for some BNs often utilized the Delphi method through surveys to obtain network structures based on expert experience. However, for complex networks with vast historical data and numerous RIFs, relying solely on expert experience is insufficient. Data-driven modeling focuses on starting from the data, combining data with prior knowledge to uncover potential causal relationships among RIFs [30].

3.1.1 Background Knowledge

In the context of structure learning, when the amount of data is limited or the relationships are complex, incorporating prior knowledge can effectively steer the structure learning algorithms, thereby enhancing both the efficiency of the learning process and the accuracy of the outcomes. The prior relationships that are introduced can consist of mandatory arcs, prohibited arcs, and variables assigned to temporal layers (in dynamic BNs), with the objective of preliminarily evaluating the causal relationships between variables as informed by existing literature and domain expertise, ultimately striving for the optimal DAG. In this study, the D-S theory of evidence, also referred to as belief function theory, was employed to effectively tackle the uncertainty and incomplete information present in the assessment of telecom fraud risks. The D-S theory is especially adept at synthesizing information from multiple sources and serves as a crucial instrument in the domains of risk assessment and decision-making support [31]. The paper involved input from two academic experts specializing in anti-fraud and two police officers with hands-on experience in fraud prevention. The scholars each possess over five years of expertise in telecom fraud and risk assessment, while the police officers bring more than five years of frontline experience in combating telecom fraud. A questionnaire was utilized to explore the interconnections between nodes, and the data collected were thoroughly analyzed using the D-S theory of evidence.

In our model, we established a mass function framework Θ that includes all potential RIFs related to susceptibility to telecom fraud. For each RIF A, the mass

function must satisfy the following conditions:

$$\begin{cases} m(\varnothing) = 0, \\ \sum_{A \subset \Theta} m(A) = 1. \end{cases} \tag{2}$$

In this context, $m(A)$ represents the mass function for event A, reflecting the degree of belief or trust in event A. We utilized the combination rule from the D-S theory of evidence to combine basic belief assignments obtained from various data sources into a unified belief assignment. This combination was achieved using the following formula:

$$(m_1 \oplus m_2 \oplus \cdots \oplus m_n)(A) = \begin{cases} 0, \ A = \varnothing, \\ \frac{1}{K} \sum_{A_1 \cap A_2 \cap \cdots \cap A_n = A} m_1(A_1) \cdot m_2(A_2) \ldots m_n(A_n), \end{cases} \tag{3}$$

where m_1 and m_2 represent mass functions from different sources, while K serves as a normalization constant that adjusts evidence which cannot be directly combined due to complete conflict. The calculation of K is represented by the following formula:

$$\begin{aligned} K &= \sum_{A_1 \cap A_2 \cap \cdots \cap A_n \neq \varnothing} m_1(A_1) \cdot m_2(A_2) \ldots m_n(A_n) \\ &= 1 - \sum_{A_1 \cap A_2 \cap \cdots \cap A_n = \varnothing} m_1(A_1) \cdot m_2(A_2) \ldots m_n(A_n). \end{aligned} \tag{4}$$

Combining expert knowledge with the D-S theory of evidence helped enhance the understanding of intricate relationships among telecom fraud RIFs in the final BN model. This approach also boosted the reliability and thoroughness of the decision-making process, resulting in enhanced computational efficiency and accuracy in structure learning algorithms.

3.1.2 Structure Learning Algorithms

BN structure learning methods can be generally divided into two categories: constraint-based and score-based methods [32]. Constraint-based methods primarily rely on performing repeated conditional independence tests on the dataset to determine whether variables are independent. This approach uses a skeleton graph to derive the network structure. Specifically, we utilized the chi-square statistic and mutual information for testing conditional independence. The chi-square test determined independence between nodes by comparing the observed frequency O_i with the expected frequency E_i, while the mutual information test assessed the dependency between variables by calculating the conditional entropy of two variables

based on a conditioning variable and the differences between them [33].

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Z,Y). \tag{6}$$

Equations (5) and (6) describe the methods for calculating these statistics. Equation (5) uses the chi-square value to test conditional independence, while Equation (6) calculates the conditional mutual information value to assess the dependency between variables. Here, $H(X \mid Z)$ and $H(X \mid Z,Y)$ represent the conditional entropies given the conditioning variable.

Conversely, score-based methods explored the best network structure by optimizing a specific scoring function. This method assessed the network based on its posterior probability, which was derived from the prior probability and the likelihood of the given data. Our aim was to find a network structure maximizing the posterior probability given the data [31].

$$P(G \mid D)\alpha P(D \mid G)P(G), \tag{7}$$

where G represents the network structure, D signifies the data, $P(D \mid G)$ indicates the likelihood of the data given the structure $P(G)$, and $P(G)$ represents the prior probability of the structure.

In our study, we employed the constraint-based algorithm PC [34] along with two heuristic-based approaches: GTT [35] and BS [36] for learning the structure of BNs. These methods not only increased the model's precision but also improved its practicality and effectiveness in evaluating the risks of telecom fraud.

3.1.3 Parameter Learning

Parameter learning for the BN involved quantifying dependencies between node variables and determining their probability distributions. We employed the Expectation Maximization (EM) algorithm [37] to determine these distributions and conditional probabilities among the nodes. The EM algorithm serves as a convenient approximation to maximum likelihood estimation, offering robustness for generating conditional probabilities within the BN structure for telecom fraud. The EM algorithm is shown below:

$$\theta^{(t+1)} = \arg\max_{a} \int_{z} \log P(x,z \mid \theta) \cdot P\left(z \mid x, \theta^{(t)}\right) \mathrm{d}z. \tag{8}$$

In this context, x represents the given data, z denotes the latent variable, $\theta^{(t)}$ refers to the parameter at time t, $P\left(z \mid x, \theta^{(t)}\right)$ indicates the posterior probability, and $\log P(x,z \mid \theta)$ signifies the log joint probability of the complete data.

3.2 Model Validation

K-fold cross-validation is a widely-used for evaluating ML models, including BN models. In this method, data was randomly divided into K similar, mutually exclusive subsets called folds. The model underwent training K times, each time using data from K−1 folds for training and the remaining fold for testing performance [38]. To compare the performance of three structure learning algorithms for telecom fraud risk analysis in BN models, model classification accuracy and the area under the receiver operating characteristic (ROC) curve were utilized as metrics. The ROC curve, derived from the confusion matrix shown in Table 2, graphically combined the True Positive Rate (TPR) and False Positive Rate (FPR). A higher area under the curve (AUC) indicated better model classification performance.

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Table 2. Research on data-driven BN algorithm

Additionally, we validated the effectiveness of the final BN model by utilizing the forward reasoning capabilities of the BN and conducting partial theorem verification.

4 CASE STUDY

4.1 Data Collection and Processing

The primary dataset utilized in this paper was derived from a real-world telecom fraud case dataset [38]. This dataset included 60 000 original samples of fraud data collected over three years from the anti-fraud center in S city, a coastal city in southern China. Additionally, it incorporated open data sourced from platforms such as the Internet and government channels. Each fraud case provided detailed descriptions of the victim's experience, outlining the fraud process, fraud results, and basic victim information. Due to confidentiality and privacy concerns, we are unable to include the data files as appendices. However, we can email the processed data to interested readers upon request. Spanning from 2019 to 2021, these data offered comprehensive insights into various types of fraud events, victims, and other relevant information, establishing a basis for the analysis of telecom fraud.

S City, located in the southern coastal region of G Province, is a prefecture-level city with a diverse and large population. As of 2022, it had a permanent population of 17.662 million, covering an area of 1 997.47 square kilometers and encompassing nine urban districts and one administrative region. As an international metropolis, its vibrant economy, high population density, and extensive use of digital communication make it a prime target for telecom fraud.

Data preprocessing mainly included two steps: identifying RIFs and cleaning and merging data. We identified and described RIFs from fraud cases (Table 3), with the identification of RIFs relying mainly on

1. theoretical analysis of the entire fraud process and
2. the frequency of factor occurrence.

Factors with a lower frequency of occurrence, such as the victim's recent anxiety and other psychological conditions, which only appeared twice in the statistics, were excluded. In the data cleaning phase, we removed incomplete or irrelevant records from the initial database, addressing missing values, inconsistencies, and duplicate entries. A similar number of cases were selected for each type of fraud to ensure balance. All data fields were standardized to maintain format consistency, especially for important variables like victim demographics, fraud types, and economic losses. Finally, we conducted an experimental analysis using 1 876 preprocessed samples.

Based on a thorough analysis of the telecom fraud process and the theory of multi-source heterogeneous data fusion, this paper identified key RIFs related to telecom fraud. These factors were divided into 4 main categories and 14 nodes, as shown in Table 3.

From a fraud process perspective, network nodes can be categorized into four types: victim portraits, spatiotemporal distribution, fraud process, and fraud result. We further identified discrete states for these nodes in a BN, breaking down the primary fraud types into 8 states. Residential areas were defined based on S City's administrative divisions. Additionally, definitions were provided for the methods used to induce telecom fraud and for the contact method.

4.2 Model Construction

4.2.1 Coercive Relationship

The analysis and prediction of target nodes were influenced by correlations among RIFs. To enhance the model's predictive performance, we incorporated expert experience before structure learning. We collected insights from four anti-fraud experts via questionnaires about the relationships among the 14 network nodes. The collected data were then analyzed using the D-S theory of evidence (see Section 3.1.1) to mitigate the impact of subjective expert opinions.

To illustrate how expert experience shapes the relationships between nodes, we examined target nodes A and B. We assumed $m_1(1,2), \ldots, m_4(1,2)$ represented the probabilities given by four experts regarding the connection between these nodes. The value $m_1(1,2) = (0.9, 0.1)$ indicated that the first expert assigned a strong relationship probability of 0.9 and a weak relationship probability of 0.1 between nodes A and B, with each expert providing their own assessments.

Next, we computed the Belief and Plausibility functions for the relationship between the two nodes using Equation (3). These functions were used to assess the

No	Node Name	Risk	States	Reference
1	Victim Portrait	Age	(1) Age 0–17 (2) Age 18–25 (3) Age 26–32 (4) Age 33–39 (5) Age 40	[13, 39, 40]
2		Educational level	(1) Low (2) Medium (3) High	[13]
3		Gender	(1) Female (2) Male	[11, 17]
4		Stable source of income	(1) Yes (2) No	[5]
5		Anti-fraud propaganda	(1) Yes (2) No	[18]
6		APP installation habit	(1) Yes (2) No	[7]
7		Deceived repeatedly	(1) Yes (2) No	[17]
8		Outsides	(1) Yes (2) No	[41]
9	Spatiotemporal distribution	Residential area	Area 1–10	[40, 42]
10		Deception time	(1) Day (2) Night	[40, 43]
11	Fraud process	Contact method	(1) Advertisement (2) App (3) Delivery (4) Phone (5) Text Message	[15, 44]
12		Inducing methods	(1) Acquaintance (2) Claims settlement (3) Help with matters (4) Make Friends (5) Make Money (6) Naked Chat (7) Others Shopping	
13	Fraud result	Types of fraud	(1) Extortion Fraud (2) Gambling fraud (3) Identity Fraud (4) Investment Fraud (5) Online Game Trade Fraud (6) Rebate Fraud (7) Shopping Service Fraud (8) Other Fraud	[45]
14		Money loss level	(1) Loss 0–5 000 (2) Loss 30 000–60 000 (3) Loss 5 000–30 000 (4) Loss 60 000–	[19, 45]

Table 3. BN node design in telecom fraud RIF model

strength of the causal relationship. A causal relationship was deemed to exist if the strength surpasses the threshold of 0.9. Conversely, if the strength fell below 0.10, the relationship was considered unlikely. In this case, the absence of a relationship equated to adding prohibited prior background knowledge between the two nodes, preventing the structure learning algorithms from associating the two risk node variables.

This paper established six mandatory relationships and 14 prohibited relationships based on expert judgments on the relationships between variables. The corresponding expert judgment data and node variables are displayed in Table 4.

Influencing Factors		$m_1(1,2)$	$m_2(1,2)$	$m_3(1,2)$	$m_4(1,2)$	Relationship Strength
Age	Fraud Type	$(0.671, 0.329)$	$(0.795, 0.205)$	$(0.701, 0.299)$	$(0.595, 0.405)$	$(0.96, 0.04)$
Culture	Fraud Type	$(0.661, 0.339)$	$(0.617, 0.373)$	$(0.328, 0.672)$	$(0.887, 0.113)$	$(0.92, 0.077)$
Stable Income	Fraud Type	$(0.776, 0.224)$	$(0.754, 0.246)$	$(0.538, 0.462)$	$(0.686, 0.314)$	$(0.96, 0.04)$
Administrative Division	Culture	$(0.184, 0.816)$	$(0.01, 0.90)$	$(0.40, 0.60)$	$(0.131, 0.869)$	$(0.01, 0.99)$
Culture	Gender	$(0.263, 0.737)$	$(0.298, 0.702)$	$(0.105, 0.895)$	$(0.609, 0.391)$	$(0.03, 0.97)$
Fixed Income	Gender	$(0.90, 0.1)$	$(0.25, 0.75)$	$(0.184, 0.816)$	$(0.13, 0.869$	$(0.09, 090)$

Table 4. Expert judgment on strength of relationship between RIFs

4.2.2 Data-Driven Modeling

Utilizing the mandatory relationships established by the D-S theory of evidence, we assessed three distinct structure learning algorithms to develop complete BN structures. We used GeNIe 4.0 to establish a complete BN model structure through various algorithms. GeNIe is an open-source software designed specifically for analyzing BN data. Our evaluation included three different algorithms, which comprised one constraint-based algorithm (i.e., PC) and two score-based algorithms (i.e., GTT and BS). If pseudocode for the algorithms is required, we will provide it in the appendix. Figure 2 displays the BN structures derived from the three corresponding structure learning algorithms, along with their simulation results subsequent to parameter learning.

From the subfigures on the left side of Figure 2, specifically a), c), and e), solid lines represent mandatory relationships between variables based on expert judgment, while dashed lines are derived from data-driven structure learning algorithms. The three structure learning algorithms identified various dependency relationships among RIFs, uncovering numerous potential connections between variables beyond expert experience. Notably, the score-based GTT and BS algorithms, both heuristic, showed similar learning capabilities and produced comparable structural relationships. For example, the BS algorithm indicated that the “age” factor not only influenced the “types of fraud” but also had a relationship with the “money loss level”. Additionally, the “types of fraud” also affected the “money loss level”. In

the GTT algorithm, a connection between the "age" factor and "fixed income" was also discovered. These factor relationships were captured through sample data, and the conditional probability tables for the nodes were obtained using the EM algorithm.

The right side of Figure 2 shows the simulation results of the BN models under the three algorithms, including complete forward and backward probability propagation. These results demonstrated how each algorithm processed information and the BNs' effectiveness in modeling and simulating complex dataset interactions.

4.3 Model Validation

4.3.1 Comparison of Results

Even when starting with identical background knowledge, different structure learning algorithms can produce different DAGs from the same dataset. In our comparison of three structure learning algorithms, both the BS and GTT algorithms identified 11 dependency relationships among variables, while the PC algorithm found 14. By integrating the viewpoints of four experts using the D-S theory of evidence, the additional potential relationships discovered by these algorithms were verified as reliable, thus negating the need for manual removal.

To further validate the robustness of the algorithms on datasets of varying sizes, we assessed the k value in k fold cross-validation and tracked the trends in accuracy and AUC for each algorithm. The results are shown in Figures 3 and 4.

As the k value rose, the training set's size gradually expanded, while the test set's proportion progressively diminished. Notably, the accuracy values of each algorithm showed a trend of steady improvement and tended to stabilize around $k = 10$. This indicated that increasing the amount of training data effectively improved the model's performance, but after $k > 10$, further increases in the k value had a limited impact on performance enhancement. Specifically, the GTT algorithm consistently outperformed others across all k values, with its accuracy enhancing from an initial approximate value of 0.78 to a stable value of about 0.875, showing strong robustness and stability; the BS algorithm ranked second, with performance close to GTT, achieving a final accuracy of around 0.833; the PC algorithm performed relatively poorly, with its accuracy rising from an initial approximate value of 0.70 to a final value of 0.790. Although the increase was modest, it also stabilized after $k = 10$.

In the AUC experimental results, the overall performance of the algorithm showed a gradual improvement as the amount of training data increased. While the calculation speed was faster with a lower k value (such as $k = 2$), the average AUC value for each algorithm was low, and the standard deviation was significantly higher, indicating instability in the model's performance at this stage. When the k value rose to around 10, the performance of each algorithm stabilized, and the AUC value fluctuations decreased notably. Specifically, the ROC values for both the GTT algorithm and the BS algorithm stabilized around 0.80, with the GTT algorithm achieving the best performance, nearly reaching 0.83. In contrast, the AUC value

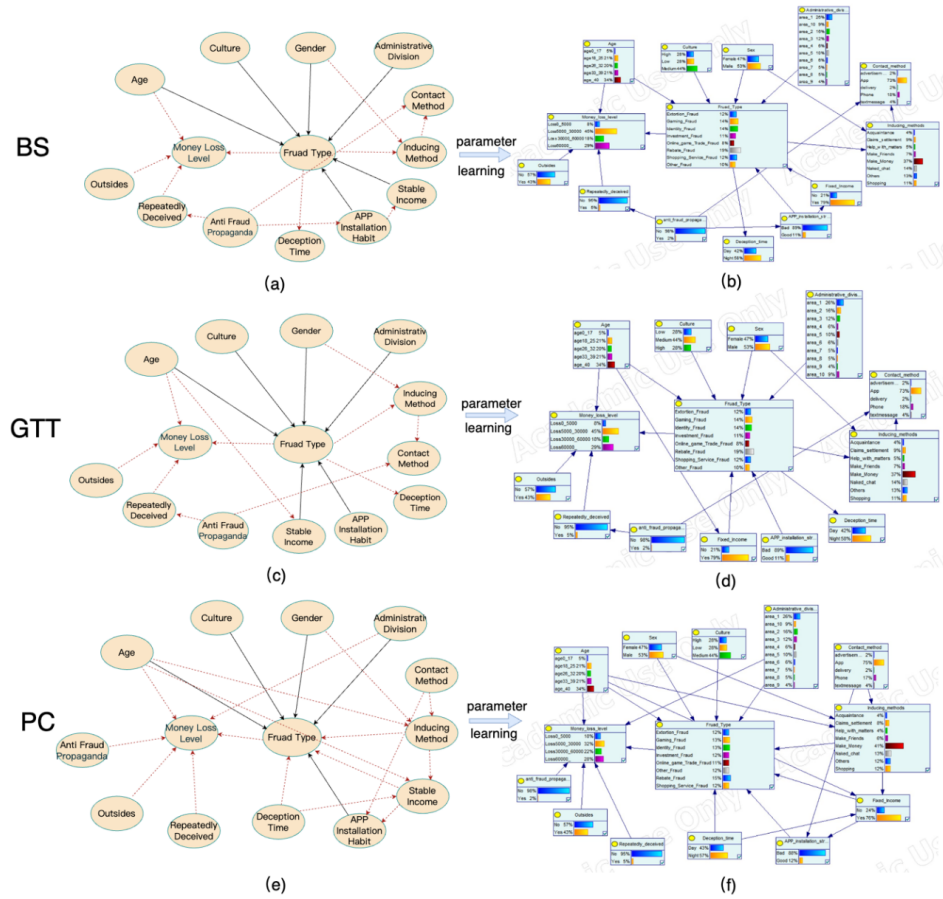


Figure 2. BN network and its parameter learning results using structure learning algorithm: a) BN structure derived from BS algorithm, b) BS-EM algorithm simulation results, c) BN structure derived from GTT algorithm, d) GTT-EM algorithm simulation results, e) BN structure derived from PC algorithm, f) PC-EM algorithm simulation results. The red dashed lines are derived from the data-driven algorithm, while the black solid lines are based on prior background knowledge in figures a), c), and e).

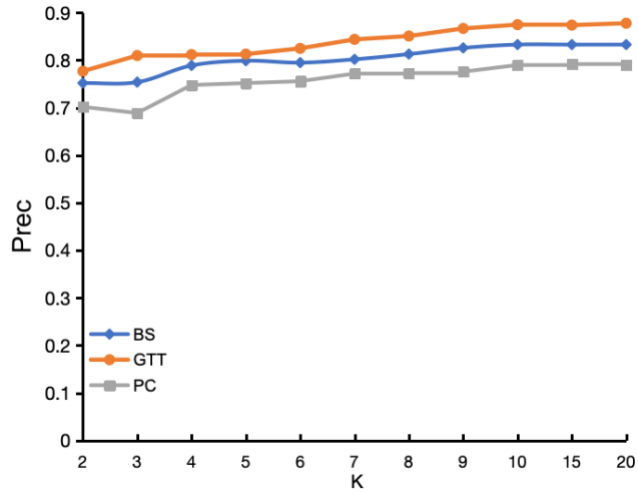


Figure 3. Effectiveness of three algorithms under different k values

of the PC algorithm remained low, leading to overall poor performance. Further increasing the k value (such as k = 20) might slightly improve performance, but this enhancement was often accompanied by a significant increase in computational time cost. Overall, the experimental results verified that the three structure learning algorithms maintained good robustness across different dataset sizes.

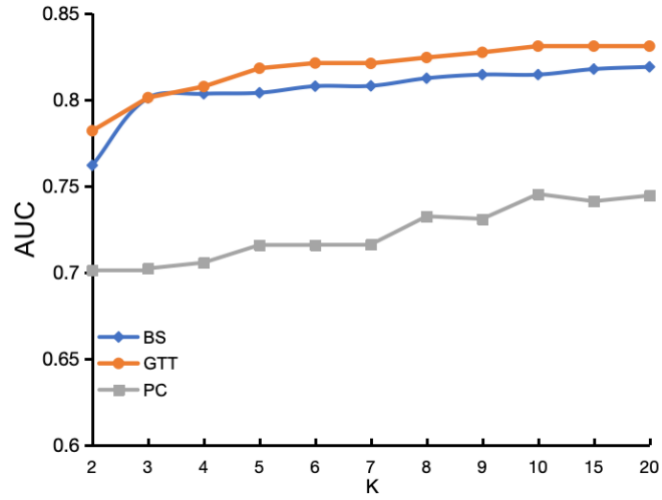


Figure 4. AUC values of three algorithms under different k values

To summarize, the GTT algorithm, as a structure learning tool, can be utilized for developing BN models aimed at analyzing RIF features in telecom fraud. The following discussions on scenario reasoning and sensitivity analysis presented in this paper are based on the BN structure derived from the GTT algorithm.

4.3.2 Sensitivity Analysis

Sensitivity analysis plays a vital role in the computational evaluation of BN models. Sensitivity analysis reveals how changes in local parameters can affect the target nodes, thereby identifying key nodal factors that influence variables such as "Type of fraud" and "Money loss level". This form of analysis also facilitates the implementation of effective measures to mitigate risks associated with these RIFs. Additionally, sensitivity analysis can, through simulation, uncover causal chains of accidents in certain accident analysis models.

We measured the mutual information and various metrics, such as the Percentage Reduction in Entropy (PRE), between the target node "type of fraud" and the RIFs. A higher mutual information value and PRE indicated a more substantial impact of the corresponding RIF on the "type of accident". Table 5 presents the sensitivity analysis results for the "type of fraud" node, including mutual information values, percentage reduction in entropy, and belief variance. The results revealed that age, gender, and fixed income were the three most influential RIFs affecting the type of fraud, with changes in these nodes significantly impacting the type of fraud experienced. Besides, factors such as residential area, educational level, and whether the individual is a migrant had a lesser influence on the type of fraud. Similarly, when targeting "money loss level", the most sensitive nodes were identified as age, whether the individual has been defrauded repeatedly, and whether they are a migrant.

Node	Mutual Information	Entropy Reduction Percent	Variance of Beliefs
Age	0.3246	24.84 %	$7.755^{*}10^{-3}$
Gender	0.1633	10.22 %	$3.763^{*}10^{-3}$
Fixed Income	0.15882	8.84 %	$2.811^{*}10^{-3}$
Contact method	0.14317	7.52 %	$2.691^{*}10^{-3}$
Inducing method	0.12375	7.99 %	$2.691^{*}10^{-3}$
Administrative division	0.11074	5.57 %	$2.022^{*}10^{-3}$
APP installation habit	0.09095	4.93 %	$2.071^{*}10^{-3}$
Culture	0.08870	2.89 %	$1.78^{*}10^{-3}$
Outsides	0.07952	2.64 %	$1.62^{*}10^{-3}$

Table 5. Sensitivity analysis of "type of fraud" nodes

Additionally, we needed to identify the most sensitive parent node states for each specific fraud type. This step was vital for implementing targeted anti-fraud tactics and strategies for risk reduction. Changes in posterior probabilities of nodes served as the basis for determining the most sensitive attributes of a particular fraud type.

From an overall perspective, the results of our model's analysis corroborated some of the existing speculative qualitative research [16, 17, 18]. For example, Rebate Fraud, the most prevalent case type, typically preys on people's greed by offering the lure of "easy money from home", attracting victims who often lack a stable income, such as young individuals and homemakers eager to earn money effortlessly, thus falling into the fraudsters' traps. Additionally, it has been observed that Gambling fraud and Extortion fraud are particularly sensitive to gender factors. Further analysis revealed that these fraud types mainly exploit male tendencies towards gambling and lascivious behavior. The sensitivity analysis results for all fraud types are shown in Table 6. It is essential to adopt distinct preventive strategies based on the sensitivity analysis results, which we will discuss in Section 5.2.

Fraud Type	Sensitive Node Status	Sensitivity Value
Extortion Fraud	Gender	9.89 %
Gambling Fraud	Gender	13.22 %
Identity Fraud	Area_8	8.25 %
Investment Fraud	APP installation habit	6.95 %
Online Game Trade Fraud	Age-0_17	22.60 %
Rebate Fraud	Stable Income	6.67 %
Shopping Service Fraud	Culture-Medium	4.45 %
Other Fraud	Area_1	7.74 %

Table 6. Sensitivity analysis results for all types of fraud

4.3.3 Scenario Analysis

BNs enable forward reasoning to predict potential scenarios based on various risk characteristics, allowing for the evaluation of victims' susceptibility to various types of fraud. Due to limitations in space, we chose a few specific multivariate features for combined analysis, which were utilized for BN reasoning and application in real-world situations; the results are shown in Table 7.

Based on the results of forward scenario reasoning in the BN, different risk coupling features led to changes in the probability distribution of the states for the target node "type of fraud". In Scenario 1, we selected an intermediate state for each variable, including an age range of 26 to 32 years, male, medium education level, stable income, and residing in Area_1. The BN forward reasoning results indicated that the most probable type of fraud to occur is Extortion fraud, representing 32 % of all fraud types. Given that age is the most sensitive factor in the sensitivity analysis, in Scenario 2, we adjusted the age range and compared it with Scenario 1. We noted a noticeable shift in the posterior probabilities of various fraud types: the likelihood of Extortion fraud dropped from 32 % to 28 %, while the probability of Investment fraud rose markedly, indirectly suggesting that older individuals might be more susceptible to Investment fraud. This finding is consistent

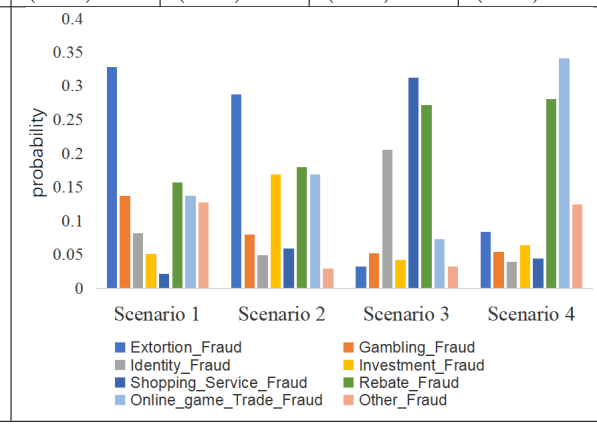
RIF	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Age	26–32	33–39	33–39	33–39
Gender	Male	Male	Female	Female
Culture	Medium	Medium	Medium	Medium
Stable Income	Yes	Yes	Yes	Yes
Administrative Division	Area_1	Area_1	Area_1	Area_2
App Installation Habit	Bad	Bad	Bad	Bad
Outsides	No	No	No	No
Anti-Fraud Propaganda	No	No	No	No
Loss Level	5 000–3 000 (46 %)	5 000–3 000 (44 %)	5 000–3 000 (42 %)	5 000–3 000 (48 %)
Fraud Type				

Table 7. Scenario reasoning results

with real-world observations, as extortion fraud often targets younger individuals who may be more impulsive or less adept at handling coercive situations, whereas investment fraud schemes typically appeal to older individuals who have accumulated more wealth and are seeking investment opportunities [39]. The changes in probabilities observed after adjusting for age support the hypothesis that fraudsters tailor their fraud methods based on factors like the age and economic status of their targets, underscoring the importance of considering victim profiles in fraud prevention strategies.

In Scenario 3, when the states of other RIFs remained unchanged except for altering the gender to female, there was a notable shift in the distribution of potential fraud types. The most likely fraud types were Rebate fraud and Shopping service fraud, with probabilities of 27 % and 31 %, respectively. This is due to fraudsters using tailored strategies, such as exploiting shopping preferences or of-

fering false rebates to attract consumer habits, which often target women. In Scenario 4, changing the administrative division also resulted in variations in the types of fraud experienced. This variability could stem from differences in the intensity of fraud awareness and intervention efforts across regions, leading to some areas being more affected by specific types of fraud. For example, regions with more effective awareness campaigns might see a reduction in certain fraud types, while others could experience an increase due to enforcement gaps or targeted strategies by fraudsters.

Furthermore, we employed reverse inference of the model to pinpoint high-risk variable attributes associated with specific fraud cases, demonstrating its practicality. For example, in the case of financial gambling fraud, Figure 5 shows the prediction results of the model. The group most susceptible to this type of fraud is individuals aged 18–25 (29 %) and males over 40 (27 %), typically with moderate educational backgrounds, primarily located in areas 1 and 2. Among them, 57 % are migrant workers who generally have a stable income and are most likely to incur losses ranging from 5 000 to 30 000 yuan. The main reason for being defrauded is the desire to make money. From the results, the model revealed how these risk variables interplayed to affect susceptibility to specific types of fraud, providing a solid foundation for future prevention and control strategies. For example, it suggested that measures such as strengthening financial security education for the mobile population and enhancing anti-fraud publicity at the community level should be taken.

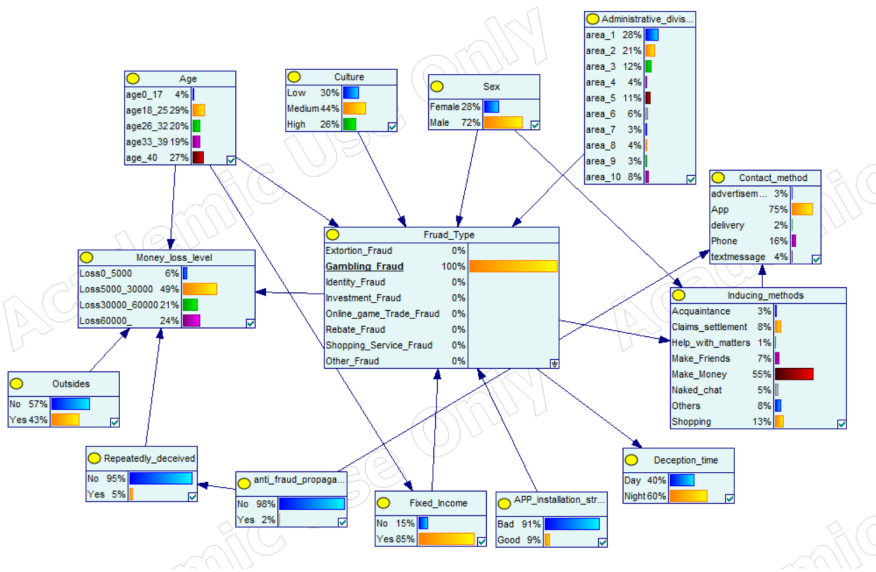


Figure 5. BN network and its parameter learning results under structure learning algorithm

4.3.4 Correctness Verification

To quantitatively validate the consequence model that utilizes the Bayesian Belief Network, this study also conducted partial validation analysis based on the axioms put forth by Jones et al. [46]:

1. Sensitivity of Posterior Probabilities: A slight increase or decrease in the prior probability of a parent node should lead to a corresponding change in the posterior probability of the related child node.
2. Consistent Impact: Changes in the probability distribution of a parent node should produce a consistent effect on the corresponding child node.
3. Cumulative Impact of Multiple Attributes: The influence of probability resulting from a combination of m attributes should consistently exceed the influence derived from a combination of $m - n$ ($n \in m$) attributes.

In the model, when the prior probability of the parent node "Gender" being "Female" was adjusted from 45 % to 50 %, the likelihood of "Identity fraud" declined from 12.9 % to 12.7 %, while the probability of the money loss level "5 000 to 30 000 yuan" also dropped from 43.4 % to 43.3 %. Subsequently, when the prior probability of another parent node "Educational level" in the "High" state was changed from 29 % to 40 %, the probability of "Identity Fraud" reduced from 12.8 % to 12.7 %, and the chances of losses in the range of "5 000–20 000 yuan" decreased from 43.38 % to 43.36 %. The changes observed in each RIF adhere to the triple axiom definition, thus achieving partial validation of the model. Similar evaluations can be applied to other RIFs.

5 DISCUSSION

5.1 Advantages of Data-Driven BN

Data-driven networks, as an alternative learning approach, can provide support in circumstances where prior knowledge or data is insufficient. These networks provide a more objective and effective solution than relying exclusively on the experience of experts. Furthermore, the integration of expert insights through the D-S theory of evidence improves the rationality of the BN structure. Importantly, data-driven methods reveal the potential relationships between RIFs, highlighting the necessity to establish connections among various factors for effective risk management [24]. This paper integrates background knowledge derived from expert experience using D-S evidence theory, followed by the incorporation of potential relationships identified by structure learning algorithms, which bolsters the rationality and robustness of the BN model for risk analysis. For example, our data-driven analysis revealed that "age" influences both "type of fraud" and "loss situation", and identified relationships among "method of deception", "type of fraud", and "method of contact", which have not been clearly defined in previous studies [19, 38].

Among the three algorithms for structure learning, the GTT algorithm outperformed the BS and the constraint-based PC algorithm. As an excellent search strategy, GTT features reduced time and space complexity and effectively prevents data overfitting, making it well-suited for telecom fraud datasets that involve small to medium sample sizes.

5.2 Fraud Prevention and Intervention Policies

In the realm of public security, a "one-size-fits-all" approach for fraud prevention often falls short, as many individuals, from a psychological perspective, naively assume they are not susceptible to deception. This results in anti-telecom fraud messages not being taken seriously [18]. By applying evidence reasoning and sensitivity analysis of BN, we can identify the most probable types of fraud and potential losses linked to various risk attribute characteristics, which can improve the fraud prevention initiatives undertaken by public security agencies.

Our sensitivity analysis identified age, gender, and whether an individual has a stable income as key factors influencing susceptibility to different types of fraud. Consequently, it is essential to implement targeted fraud prevention and intervention strategies tailored to different age groups, genders, and occupations. For example, as indicated by the results in Section 4.3.3, men aged 26–32 with a medium level of education should be especially vigilant against Extortion Fraud, while women aged 33–39 need to be more wary of Rebate Fraud and Shopping Service Fraud. Measures such as strengthening financial safety education for the migrant population, cautioning victims about false promises of "high returns", and exposing the tactics of gambling and investment fraud can help reduce the incidence of these fraud types. According to Section 4.3.2, minors aged 0–17 are the primary targets of Online Game Trade Fraud, highlighting the importance of increasing fraud awareness and prevention efforts for both children and their guardians.

By implementing model-based, targeted, and directed publicity strategies tailored to different groups, spatial distributions, and fraud methods, police resources can be efficiently allocated. For example, analyzing fraud cases by administrative regions and time frames enables the identification of high-risk fraud types and potential victim profiles in specific areas, creating region-specific anti-fraud operational models. This approach significantly mitigates the risk and occurrence of fraud, as observed in Area_8, there is a strong need to improve measures against Identity Fraud. For fraud types showing high incidence during certain periods, profiling and predicting susceptible demographics, especially in response to emerging fraud schemes, can achieve targeted swift effects to prevent individuals from being misled.

Even though the training data for the model comes from a southern Chinese city, it is still possible to achieve effective prediction results by dynamically adjusting the BN model according to local fraud incidents, demographic characteristics, regional economic development, and other societal factors. This method allows for a shift in strategic focus towards public awareness and prevention, transitioning from a generalized to a more targeted approach in combating fraud.

6 CONCLUSION

In this paper, we developed a data-driven BN model to evaluate telecom fraud risks, combining both expert knowledge and empirical data. We utilized three structural learning algorithms – PC, BS, and GTT – together with the EM algorithm to build the BN model. We employed K-fold cross-validation to assess the performance of these algorithms on a telecom fraud dataset, ultimately selecting the most effective model. By performing evidence reasoning and sensitivity analysis on the established BN, we gained a more profound understanding of telecom fraud. From the results and discussions, we come to the following conclusions:

1. The data-driven BN modeling approach is capable of identifying latent relationships among RIFs, thereby mitigating biases that are typically present in models that rely solely on expert judgment, and addressing issues related to the scalability of the model as data volumes increase.
2. Scenario reasoning can uncover quantitative relationships among various RIFs and the susceptibility to different types of fraud and losses. It quantifies issues from the victim's perspective, posing questions such as "What kind of people are more susceptible to specific types of fraud?" and "How do these fraudsters approach and deceive their victims?" This helps in formulating educational campaigns and making well-informed decisions.
3. According to the sensitivity analysis, "age", "gender", and "having a stable income" are the top three RIFs influencing susceptibility to various types of fraud. Additionally, each of the eight distinct types of fraud has its own set of highly sensitive characteristics.

While this study incorporates a diverse array of pertinent RIFs, the constantly evolving nature of telecom fraud tactics presents ongoing challenges. Fraud methods continue to adapt, which may introduce new RIFs that our current model fails to account for. Moreover, our dataset, extensive as it is, may not fully reflect the diversity and regional variability present in fraud cases.

To overcome these limitations, future work will focus on the following aspects:

1. Adapting to the quickly changing landscape of telecom fraud by incorporating real-time data streams.
2. Minimizing potential data biases and further improving risk analysis models by exploring datasets from various regions and incorporating additional RIFs, including psychological and behavioral dimensions.
3. Using dynamic Bayesian Networks to represent changes in fraud tactics and victim vulnerability; a digital twin technology can also be incorporated into the proposed model to capture real-time risk.

Acknowledgments

This research received financial support from the Basic Research Funds of the People's Public Security University of China, as well as from the Funding for Discipline Innovation and Talent Introduction Bases in Higher Education Institutions (Grant No. 2024JKF02, Grant No. B20087).

REFERENCES

[1] Geng, Y.: Research on How to Deal with the Dilemma of Global Cooperative Governance of Cross-Border Telecom Network Fraud in China. Chinese Studies, Vol. 6, 2017, No. 4, pp. 249–263, doi: 10.4236/chnstd.2017.64023.

[2] Wei, K.: Research on Multiple Cooperative Governance Mechanism of Telecom Fraud under the Background of Internet+. Social Security and Administration Management, Vol. 4, 2023, No. 5, pp. 13–18, doi: 10.23977/socsam.2023.040503.

[3] Huang, S. Y.—Lin, C. C.—Chiu, A. A.—Yen, D. C.: Fraud Detection Using Fraud Triangle Risk Factors. Information Systems Frontiers, Vol. 19, 2017, No. 6, pp. 1343–1356, doi: 10.1007/s10796-016-9647-9.

[4] Jiang, N.: Securing Large Cellular Networks via a Data Oriented Approach: Applications to SMS Spam and Voice Fraud Defenses. Ph.D. Thesis. University of Minnesota, Minneapolis, MN, USA, 2013.

[5] Guo, W.: Research on Semantic Analysis-Based Recognition of Telecommunication Fraud Discourse Patterns. Advances in Computer, Signals and Systems, Vol. 7, 2023, No. 8, pp. 71–77, doi: 10.23977/acss.2023.070808.

[6] Priya, G. J.—Saradha, S.: Fraud Detection and Prevention Using Machine Learning Algorithms: A Review. Proceedings of the 7th International Conference on Electrical Energy Systems (ICEES), IEEE, 2021, pp. 564–568, doi: 10.1109/ICEES51510.2021.9383631.

[7] Hu, X.—Chen, H.—Liu, S.—Jiang, H.—Chu, G.—Li, R.: BTG: A Bridge to Graph Machine Learning in Telecommunications Fraud Detection. Future Generation Computer Systems, Vol. 137, 2022, pp. 274–287, doi: 10.1016/j.future.2022.07.020.

[8] Wang, Y.—Chen, H.—Liu, S.—Li, X.—Hu, Y.: Feature Difference-Aware Graph Neural Network for Telecommunication Fraud Detection. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, Vol. 45, 2023, No. 5, pp. 8973–8988, doi: 10.3233/JIFS-221893.

[9] Raghavan, P.—Gayar, N. E.: Fraud Detection Using Machine Learning and Deep Learning. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, 2019, pp. 334–339, doi: 10.1109/ICCIKE47802.2019.9004231.

[10] Byun, J. E.—Song, J.: A General Framework of Bayesian Network for System Reliability Analysis Using Junction Tree. Reliability Engineering & System Safety, Vol. 216, 2021, Art. No. 107952, doi: 10.1016/j.ress.2021.107952.

[11] Liu, Z.—Li, D.: Research of Dempster-Shafer's Theory and Ensemble Classifier Financial Risk Early Warning Model Based on Benford's Law. Computational Economics, Vol. 65, 2025, No. 6, pp. 3361–3389, doi: 10.1007/s10614-024-10679-1.

[12] Yan, Q.—Sun, Y.—Cao, Y.—Yang, J.—Zhang, A.—Ju, J.—Shi, W.—Yang, X.—Yin, J.—Wang, Z.: An Adaptive Graph Neural Networks Based on Cost-Sensitive Learning for Fraud Detection. Proceedings of the 7th International Symposium on Autonomous Systems (ISAS), IEEE, 2024, pp. 1–6, doi: 10.1109/ISAS61044.2024.10552392.

[13] Hanoch, Y.—Wood, S.: The Scams Among Us: Who Falls Prey and Why. Current Directions in Psychological Science, Vol. 30, 2021, No. 3, pp. 260–266, doi: 10.1177/0963721421995489.

[14] Ye, N.—Cheng, L.—Zhao, Y.: Identity Construction of Suspects in Telecom and Internet Fraud Discourse: From a Sociosemiotic Perspective. Social Semiotics, Vol. 29, 2019, No. 3, pp. 319–335, doi: 10.1080/10350330.2019.1587847.

[15] Fischer, P.—Lea, S. E. G.—Evans, K. M.: Why Do Individuals Respond to Fraudulent Scam Communications and Lose Money? The Psychological Determinants of Scam Compliance. Journal of Applied Social Pyschology, Vol. 43, 2013, No. 10, pp. 2060–2072, doi: 10.1111/jasp.12158.

[16] Mao, F.: The Empirical Research on the Victims of Telecom Fraud in China – Based on the Record of Victims with Quantification Statistics. The Frontiers of Society, Science and Technology, Vol. 4, 2022, No. 4, pp. 62–68, doi: 10.25236/FSST.2022.040411.

[17] Button, M.—Lewis, C.—Tapley, J.: Fraud Typologies and the Victims of Fraud: Literature Review. National Fraud Authority, 2009.

[18] Chen, K.—Ma, C.: Data Analysis of the MBTI Personality Distribution of Telecommunication Fraud Victims. In: Kuang, Y., Zhu, L., Zhang, X., Khan, I. A. (Eds.): Proceedings of the 2024 5th International Conference on Education, Knowledge and Information Management (ICEKIM 2024). Atlantis Highlights in Computer Sciences (AHCS), Atlantis Press, Vol. 22, 2024, pp. 769–780, doi: 10.2991/978-94-6463-502-7_81.

[19] Ni, P.—Yu, W.: A Victim-Based Framework for Telecom Fraud Analysis: A Bayesian Network Model. Computational Intelligence and Neuroscience, Vol. 2022, 2022, Art. No. 7937355, doi: 10.1155/2022/7937355.

[20] Li, M.—Wang, H.—Wang, D.—Shao, Z.—He, S.: Risk Assessment of Gas Explosion in Coal Mines Based on Fuzzy AHP and Bayesian Network. Process Safety and Environmental Protection, Vol. 135, 2020, pp. 207–218, doi: 10.1016/j.psep.2020.01.003.

[21] Zhang, G.—Thai, V. V.: Expert Elicitation and Bayesian Network Modeling for Shipping Accidents: A Literature Review. Safety Science, Vol. 87, 2016, pp. 53–62, doi: 10.1016/j.ssci.2016.03.019.

[22] Nhat, D. M.—Venkatesan, R.—Khan, F.: Data-Driven Bayesian Network Model for Early Kick Detection in Industrial Drilling Process. Process Safety and Environmental Protection, Vol. 138, 2020, pp. 130–138, doi: 10.1016/j.psep.2020.03.017.

[23] Zhao, X.—Yuan, H.—Yu, Q.: Autonomous Vessels in the Yangtze River: A Study on the Maritime Accidents Using Data-Driven Bayesian Networks. Sustainability,

Vol. 13, 2021, No. 17, Art. No. 9985, doi: 10.3390/su13179985.

[24] Meng, H.—An, X.—Xing, J.: A Data-Driven Bayesian Network Model Integrating Physical Knowledge for Prioritization of Risk Influencing Factors. Process Safety and Environmental Protection, Vol. 160, 2022, pp. 434–449, doi: 10.1016/j.psep.2022.02.010.

[25] Barry, D. J.: Estimating Runway Veer-Off Risk Using a Bayesian Network with Flight Data. Transportation Research Part C: Emerging Technologies, Vol. 128, 2021, Art. No. 103180, doi: 10.1016/j.trc.2021.103180.

[26] Simsekler, M. C. E.—Qazi, A.: Adoption of a Data-Driven Bayesian Belief Network Investigating Organizational Factors That Influence Patient Safety. Risk Analysis, Vol. 42, 2022, No. 6, pp. 1277–1293, doi: 10.1111/risa.13610.

[27] Liu, K.—Yu, Q.—Yuan, Z.—Yang, Z.—Shu, Y.: A Systematic Analysis for Maritime Accidents Causation in Chinese Coastal Waters Using Machine Learning Approaches. Ocean & Coastal Management, Vol. 213, 2021, Art. No. 105859, doi: 10.1016/j.ocecoaman.2021.105859.

[28] Jiang, L.—Zhang, L.—Yu, L.—Wang, D.: Class-Specific Attribute Weighted Naive Bayes. Pattern Recognition, Vol. 88, 2019, pp. 321–330, doi: 10.1016/j.patcog.2018.11.032.

[29] Hu, Z.—Mahadevan, S.: Bayesian Network Learning for Data-Driven Design. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering, Vol. 4, 2018, No. 4, Art. No. 041002, doi: 10.1115/1.4039149.

[30] Li, H.—Ren, X.—Yang, Z.: Data-Driven Bayesian Network for Risk Analysis of Global Maritime Accidents. Reliability Engineering & System Safety, Vol. 230, 2023, Art. No. 108938, doi: 10.1016/j.ress.2022.108938.

[31] Liu, L.—Yager, R. R.: Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction. In: Yager, R. R., Liu, L. (Eds.): Classic Works of the Dempster-Shafer Theory of Belief Functions. Springer, Berlin, Heidelberg, Studies in Fuzziness and Soft Computing, Vol. 219, 2008, pp. 1–34, doi: 10.1007/978-3-540-44792-4_1.

[32] Li, S.—Zhang, J.: Review of Bayesian Networks Structure Learning. Computer Application Research, Vol. 32, 2015, No. 3, pp. 641–646, doi: 10.3969/j.issn.1001-3695.2015.03.001 (in Chinese).

[33] Kitson, N. K.—Constantinou, A. C.—Guo, Z.—Liu, Y.—Chobtham, K.: A Survey of Bayesian Network Structure Learning. Artificial Intelligence Review, Vol. 56, 2023, No. 8, pp. 8721–8814, doi: 10.1007/s10462-022-10351-w.

[34] Kjærulff, U. B.—Madsen, A. L.: Bayesian Networks and Dempster-Shafer Theory of Belief Functions: A Guide to Construction and Analysis. Springer, 2013, doi: 10.1007/978-1-4614-5104-4.

[35] Kelangath, S.—Das, P. K.—Quigley, J.—Hirdaris, S. E.: Risk Analysis of Damaged Ships – A Data-Driven Bayesian Approach. Ships and Offshore Structures, Vol. 7, 2012, No. 3, pp. 333–347, doi: 10.1080/17445302.2011.592358.

[36] Tonda, A.—Spritzer, A.—Lutton, E.: Balancing User Interaction and Control in BNSL. In: Legrand, P., Corsini, M. M., Hao, J. K., Monmarché, N., Lutton, E., Schoenauer, M. (Eds.): Artificial Evolution (EA 2013). Springer, Cham, Lecture

Notes in Computer Science, Vol. 8752, 2014, pp. 211–223, doi: 10.1007/978-3-319-11683-9_17.

[37] Dempster, A. P.—Laird, N. M.—Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B: Statistical Methodology, Vol. 39, 1977, No. 1, pp. 1–22, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[38] Hu, M.—Li, X.—Li, M.—Zhu, R.—Si, B.: A Framework for Analyzing Fraud Risk Warning and Interference Effects by Fusing Multivariate Heterogeneous Data: A Bayesian Belief Network. Entropy, Vol. 25, 2023, No. 6, Art. No. 892, doi: 10.3390/e25060892.

[39] Shang, Y.—Wu, Z.—Du, X.—Jiang, Y.—Ma, B.—Chi, M.: The Psychology of the Internet Fraud Victimization of Older Adults: A Systematic Review. Frontiers in Psychology, Vol. 13, 2022, Art. No. 912242, doi: 10.3389/fpsyg.2022.912242.

[40] Judges, R. A.—Gallant, S. N.—Yang, L.—Lee, K.: The Role of Cognition, Personality, and Trust in Fraud Victimization in Older Adults. Frontiers in Psychology, Vol. 8, 2017, Art. No. 588, doi: 10.3389/fpsyg.2017.00588.

[41] Wu, H.—Wang, Q.—Zheng, Z.: Spatial Characteristics and Influencing Factors of the Origin of Telecommunication Network Fraud Criminals. Geographical Research, Vol. 42, 2023, No. 12, pp. 3219–3234, doi: 10.11821/dlyj020230154 (in Chinese).

[42] Liu, L.—Zhang, C.—Feng, J.—Xiao, L.—He, Z.—Zhou, S.: The Spatial-Temporal Distribution and Influencing Factors of Fraud Crime in ZG City, China. Acta Geographica Sinica, Vol. 72, 2017, No. 2, pp. 315–328, doi: 10.11821/dlxb201702011 (in Chinese).

[43] Chen, R. Z.—Shih, C. H.: Applying Spatio-Temporal Analysis in Telecom Fraud Investigation. Procedia Computer Science, Vol. 192, 2021, pp. 3109–3113, doi: 10.1016/j.procs.2021.09.083.

[44] Workman, M.: Wisecrackers: A Theory-Grounded Investigation of Phishing and Pretext Social Engineering Threats to Information Security. Journal of the American Society for Information Science and Technology, Vol. 59, 2008, No. 4, pp. 662–674, doi: 10.1002/asi.20779.

[45] Marcot, B. G.—Hanea, A. M.: What Is an Optimal Value of K in K-Fold Cross-Validation in Discrete Bayesian Network Analysis? Computational Statistics, Vol. 36, 2021, No. 3, pp. 2009–2031, doi: 10.1007/s00180-020-00999-9.

[46] Jones, B.—Jenkinson, I.—Yang, Z.—Wang, J.: The Use of Bayesian Network Modelling for Maintenance Planning in a Manufacturing Industry. Reliability Engineering & System Safety, Vol. 95, 2010, No. 3, pp. 267–277, doi: 10.1016/j.ress.2009.10.007.

Binzhou Si is a postgraduate student of the People's Public Security University of China, whose research interests are machine learning and artificial intelligence. He is committed to applying models and algorithms to actual public security work. So far, he has participated in one national project and one provincial project.

Haichun Sun serves as Associate Professor and Master's supervisor in the People's Public Security University of China. She has a doctorate degree in computer science. Her professional fields are natural language processing and artificial intelligence. She served as the reviewer of many excellent journals, such as Computer Science and Exploration, and presided over a number of national, provincial and ministerial projects.

Mengyuan Shao is a postgraduate student of the People's Public Security University of China. Her research direction is large model security and artificial intelligence. She focuses on exploring the risk prevention and control mechanisms and technical optimization paths in the application of large models in the field of public security.