

Article type:
Original Research

Article history:
Received 11 August 2026
Revised 04 October 2026
Accepted 13 October 2026
Initial Publish 24 November 2026
Published online 01 July 2027

Samira. Bahr Kazemi¹, Reza. Ehtesham Rasi^{2*}, Sadegh. Abedi³

- 1 Ph.D of Student, Industrial Management, Qa.c., Islamic Azad University, Qazvin, Iran
- 2 Associate Professor, Department of Industrial Management, Qa.c., Islamic Azad University, Qazvin, Iran
- 3 Assistant Professor, Department of Industrial Management, Qa.c., Islamic Azad University, Qazvin, Iran

Corresponding author email address:
reza.ehteshamrasi@iau.ac.ir

How to cite this article:

Bahr Kazemi, S., Ehtesham Rasi, R., & Abedi, S. (2027). Designing an Innovative Risk Management Framework for Fraud Detection in Automobile Insurance: A Case Study of Saman Insurance. *Future of Work and Digital Management Journal*, 5(4), 1-22. <https://doi.org/10.61838/fwdmj.379>



© 2027 the authors. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Designing an Innovative Risk Management Framework for Fraud Detection in Automobile Insurance: A Case Study of Saman Insurance

ABSTRACT

This study aimed to design and validate an innovative fraud detection risk management framework for automobile insurance at Saman Insurance by integrating fuzzy Delphi-based expert consensus with an XGBoost machine-learning model using real claims data. This applied mixed-method study was conducted in two sequential phases. In the first phase, dimensions, components, and indicators of automobile insurance fraud detection were extracted through literature review, claims-file analysis, and expert consultation and were subsequently screened and validated through a three-round fuzzy Delphi process involving 50 insurance, fraud-detection, machine-learning, and data-mining specialists. In the second phase, 4,860 automobile insurance claims files were analyzed, including 772 fraudulent and 4,088 non-fraudulent claims. Of 47 expert-validated indicators, 24 measurable predictors were retained after preprocessing and entered into an XGBoost binary classification model. Model development incorporated an 80:20 training-test split, five-fold cross-validation, hyperparameter optimization, and early stopping. Performance was evaluated using accuracy, precision, recall, specificity, F1 score, AUC-ROC, precision-recall AUC, and balanced accuracy. The fuzzy Delphi process resulted in seven validated dimensions and 47 indicators, with technical and expert damage indicators receiving the highest fuzzy mean (0.892), followed by policyholder behavioral indicators (0.874) and accident characteristics (0.853). The final XGBoost model achieved a test accuracy of 0.927, precision of 0.826, recall of 0.792, specificity of 0.952, F1 score of 0.809, AUC-ROC of 0.946, precision-recall AUC of 0.838, and balanced accuracy of 0.872. The most influential predictors were inconsistency between reported accident and observed damage, contradiction among submitted documents, previous claims frequency, suspicion of staged accidents, and delay in claim reporting. The integration of expert-derived fraud indicators with XGBoost produced a robust and interpretable framework capable of supporting probability-based fraud risk scoring, claims prioritization, and more efficient allocation of investigative resources in automobile insurance.

Keywords: Automobile insurance; Fraud detection; Risk management; XGBoost; Fuzzy Delphi; Machine learning; Insurance claims; Predictive analytics

Introduction

Insurance fraud constitutes one of the most persistent and complex sources of operational, financial, and reputational risk in contemporary insurance markets. Unlike ordinary underwriting uncertainty, fraud is associated with intentional misrepresentation, concealment, fabrication, exaggeration, or manipulation of information for the purpose of obtaining illegitimate financial benefits from an insurer. Its occurrence increases claim costs, distorts actuarial calculations, weakens the accuracy of pricing models, raises administrative expenses, and may ultimately transfer part of the financial burden to honest policyholders through higher premiums. Fraudulent behavior is particularly difficult to control because it is adaptive, heterogeneous, and often embedded within otherwise legitimate claims processes. Fraudsters may alter their methods in response to existing control mechanisms, while suspicious claims frequently resemble genuine losses in many observable

respects. Consequently, fraud management requires more than traditional auditing or rule-based screening and increasingly depends on integrated analytical systems capable of combining expert judgment, organizational control, and intelligent predictive modeling [1, 2].

The automobile insurance sector is especially vulnerable to fraud because of its large claims volume, the frequency of relatively small and medium-sized losses, the diversity of accident circumstances, the participation of multiple actors, and the considerable amount of information that must be evaluated during claims processing. Fraud may involve staged collisions, fabricated accidents, deliberate vehicle damage, inflated repair costs, manipulated documents, false statements concerning accident circumstances, duplicate claims, concealment of prior damage, or collusion among policyholders, repair centers, assessors, or other parties. Detecting such behavior is difficult because a fraudulent claim may contain both legitimate and manipulated elements and because the observable characteristics of fraud vary across products, regions, claimant profiles, and organizational processes. A systematic review of automobile insurance fraud detection research has shown that data mining and machine-learning methods have become increasingly important because of their ability to detect nonlinear patterns and interactions that conventional manual inspection may overlook [3]. Recent research focusing specifically on automobile insurance further indicates that machine learning and deep learning can substantially strengthen fraud screening, although their practical value depends on data quality, class imbalance, interpretability, model validation, and organizational implementation [4].

Traditional insurance fraud detection has generally relied on claims assessors, predefined red flags, expert rules, audits, and post-claim investigation. These approaches remain valuable because domain experts can recognize contextual inconsistencies that may not be easily represented by simple numerical variables. Nevertheless, exclusively manual systems are constrained by processing capacity, subjectivity, inconsistency among assessors, and an inability to evaluate large numbers of multidimensional interactions simultaneously. Purely statistical approaches also have limitations when fraud mechanisms are nonlinear or when important relationships emerge only through combinations of claim characteristics. This has encouraged a transition toward machine-learning systems that can process large claims databases and assign risk scores to individual cases. Comparative evidence on supervised and unsupervised learning suggests that machine-learning techniques can improve fraud identification when models are appropriately matched to the structure of insurance data and when prediction is integrated with practical screening procedures [2]. Practical insurance research has likewise emphasized that the effectiveness of machine learning depends not only on selecting a high-performing algorithm but also on constructing an operational workflow that connects model outputs to fraud investigation and claim-management decisions [5].

Among ensemble-learning methods, boosting algorithms have attracted substantial attention in fraud detection because they combine multiple weak decision trees into a stronger predictive system. Comparative research in insurance fraud risk analysis indicates that boosting methods can provide substantial predictive advantages because successive learners focus on observations that previous learners did not classify adequately [6]. XGBoost, or Extreme Gradient Boosting, has become particularly prominent because it integrates gradient boosting with regularization, efficient tree construction, handling of nonlinear interactions, and mechanisms for controlling model complexity. These properties make it well suited to structured insurance claims databases that include numerical, categorical, behavioral, contractual, and accident-related information. In automobile insurance, an optimized PSO-XGBoost model combined with interpretable machine learning has demonstrated the potential of boosted-tree approaches to achieve high fraud-detection performance while also identifying the contribution

of specific variables [7]. Other work in vehicle insurance has similarly demonstrated that optimized XGBoost architectures can be combined with explainability techniques to generate both accurate classifications and interpretable fraud-risk signals [8].

The performance advantages of advanced machine-learning models, however, introduce an important challenge concerning transparency. Insurance fraud detection is not merely a technical classification task; model predictions may affect claim investigation, payment delays, referral to specialized units, and the treatment of policyholders. For this reason, insurers need to understand why a claim receives a high fraud-risk score. Explainable artificial intelligence has emerged as a response to this requirement by providing methods that clarify how specific variables contribute to individual or global predictions. SHAP-based explanations have become especially influential because they provide theoretically grounded estimates of feature contributions and can be applied across different model classes. Studies of financial fraud detection have shown that XGBoost combined with SHAP can provide strong predictive performance while generating interpretable representations of fraud risk [9]. SHAP-weighted and anchor-based approaches have further demonstrated that explanation mechanisms can support the interpretation of XGBoost fraud predictions at both global and instance levels [10]. Research on vehicle insurance claims has also shown that Shapley-based explanations can increase the interpretability of intelligent fraud-detection models by identifying the variables responsible for individual predictions [11].

The current movement toward explainable fraud analytics is part of a broader methodological shift from black-box classification toward accountable risk intelligence. Recent work has proposed stacking and explainable AI frameworks in financial fraud contexts, demonstrating that ensemble methods can achieve high predictive performance while preserving an interpretable basis for decision support [12]. Interpretable ensemble approaches in credit card fraud detection likewise indicate that combining several learners may strengthen predictive robustness while maintaining sufficient transparency for practical application [13]. Statistical machine-learning research has also emphasized the value of SHAP-based explanation for understanding the contribution of predictors to fraud classifications and translating model outputs into actionable financial-risk information [14]. These developments are important for insurance organizations because fraud-control decisions are rarely made solely on the basis of a binary label; rather, they require prioritization, escalation, and the allocation of investigative resources according to differing levels and sources of risk.

The need for explainability is accompanied by concerns regarding fairness, reproducibility, and regulatory defensibility. Fraud models can inherit biases from historical data, organizational practices, or imbalanced sampling, and their apparent performance may therefore conceal systematic errors affecting certain groups or claim types. Recent research on explainable and fair anti-money-laundering systems has emphasized that transparent model development, reproducible SHAP analysis, and explicit fairness evaluation are essential for responsible use of artificial intelligence in financial institutions [15]. Similarly, Shapley value-guided adaptive ensemble learning has been proposed as a way to combine predictive accuracy with explanation and regulatory validation in financial fraud applications [16]. These concerns are particularly relevant to insurance, where model outputs may influence the intensity of claim investigation and therefore require a defensible analytical basis. A systematic review of explainable AI in fraud detection has highlighted recurring methodological challenges involving explanation stability, inconsistent validation procedures, limited reproducibility, unclear evaluation standards, and the potential gap between technical explanations and the information actually needed by decision-makers [17].

A further development in intelligent fraud detection is the integration of multiple data structures and model families rather than relying on a single predictive architecture. Graph-augmented approaches, for example, can capture relationships among entities that conventional tabular models may miss. An explainable graph-augmented XGBoost framework for health insurance fraud has illustrated how relational information and boosted-tree prediction can be combined to improve detection while preserving interpretability [18]. Other recent studies have examined explainable combinations of CatBoost, recurrent neural networks, attention mechanisms, and transformer-based tabular architectures, illustrating the increasing diversity of models available for complex fraud and risk classification tasks [19]. Integrated machine-learning and statistical procedures have also been proposed to improve insurance fraud detection accuracy by combining complementary analytical strengths rather than treating statistical and machine-learning approaches as mutually exclusive [20]. These developments suggest that the central methodological issue is no longer simply whether machine learning should be used, but how predictive models should be embedded within a broader fraud-management architecture.

This distinction is particularly important because fraud detection and fraud risk management are not identical concepts. Fraud detection focuses on identifying suspicious or fraudulent cases, whereas fraud risk management encompasses the wider process of identifying risk factors, evaluating their importance, continuously scoring claims, prioritizing cases, defining escalation procedures, investigating high-risk claims, and updating controls based on new patterns. An effective framework must therefore integrate technical indicators from claim files with behavioral, contractual, organizational, and infrastructural factors. Recent agentic AI approaches have moved in this direction by proposing automated frameworks that combine fraud detection with predictive risk intelligence and decision support [21]. Such models represent a broader conception of intelligent fraud management in which analytics become part of a dynamic organizational control system rather than a stand-alone classification mechanism.

This broader approach is also consistent with theoretical perspectives on fraudulent behavior. Fraud Triangle Theory, for example, emphasizes the interaction of pressure, opportunity, and rationalization and has been applied empirically to explain fraudulent conduct in financial contexts [1]. Although claim datasets may not directly measure these psychological dimensions, their operational manifestations can appear in behavioral inconsistencies, unusual timing, repeated claims, documentary contradictions, or exploitation of weak internal controls. Therefore, a fraud-detection framework that considers only demographic or transaction-level variables may fail to represent the organizational and behavioral context in which fraudulent claims occur. Incorporating expert knowledge can partly address this limitation by allowing experienced claims professionals and fraud specialists to identify indicators that are meaningful in practice but may be inadequately represented in generic machine-learning datasets.

Expert-based screening is especially useful when there is a need to distinguish theoretically plausible indicators from operationally meaningful ones. In insurance companies, many variables may be technically available, but not all are equally informative, reliable, or relevant to fraud risk. Conversely, some indicators recognized by experienced assessors may not be systematically recorded in existing databases. A structured expert-validation procedure can therefore serve two purposes: it can define the conceptual dimensions of fraud risk and identify the variables that should be incorporated into future intelligent systems. Fuzzy Delphi techniques are particularly suitable for this purpose because expert judgments about the relevance of fraud indicators are inherently uncertain and linguistic. By converting qualitative assessments into fuzzy numerical representations and iteratively measuring consensus, fuzzy Delphi can produce a more robust indicator set than a

single-round questionnaire or unstructured consultation. This expert-based approach complements machine learning rather than competing with it, because the expert phase establishes the conceptual architecture while the predictive phase tests measurable indicators against real claims outcomes.

The need to combine domain knowledge with intelligent analytics becomes even stronger when the practical weaknesses of existing fraud-detection studies are considered. Many studies emphasize algorithm comparison while giving less attention to how variables were conceptually selected, how model outputs should be translated into risk-management decisions, or how organizational infrastructure affects implementation. Practical guidance from insurance fraud research has warned that high predictive performance alone does not ensure operational effectiveness unless the model is aligned with data availability, investigation resources, and claims workflows [5]. Similarly, methodological reviews indicate that fraud research frequently encounters class imbalance, concept drift, limited access to verified fraud labels, changing fraud strategies, data-quality limitations, and inconsistencies in evaluation practices [3, 17]. These issues create a need for studies that connect expert validation, actual organizational data, predictive modeling, and actionable risk-management structure within a single framework.

Within automobile insurance specifically, recent XGBoost-based studies confirm the algorithm's usefulness but also reinforce the need for interpretable and optimized modeling. The PSO-XGBoost approach has demonstrated that optimization can improve fraud classification while interpretable machine learning clarifies how predictors contribute to results [7]. ABC-XGBoost-SHAP integration has similarly shown that optimization and explainability can be combined effectively for vehicle insurance fraud detection [8]. More broadly, interpretable ensemble fraud models suggest that predictive accuracy should be evaluated together with precision, recall, F1 score, AUC, and other metrics rather than relying solely on overall accuracy, particularly when fraud cases are substantially less frequent than non-fraudulent cases [12, 13]. Such multidimensional evaluation is essential because an insurer faces two different errors: failing to detect fraudulent claims and incorrectly flagging legitimate policyholders. A practical model must balance these competing risks.

At the same time, increasingly sophisticated models create an opportunity to move from static fraud classification toward continuous predictive intelligence. Recent research in automated and explainable insurance fraud systems proposes architectures in which intelligent agents, risk scores, explanation modules, and adaptive monitoring operate together [21]. Advanced ensemble and explainable modeling also indicates that fraud systems can become more responsive to new patterns when they are periodically updated and validated [16, 20]. However, such systems require a clearly defined organizational foundation, including standardized data collection, reliable claims coding, integrated databases, technical infrastructure, expert oversight, and escalation procedures. In this respect, intelligent fraud detection should be understood as a socio-technical system combining algorithms with human judgment and organizational controls.

Despite significant advances, there remains a methodological and practical gap in the integration of expert-derived fraud dimensions with real automobile claims data and machine-learning prediction within a unified risk-management framework. Existing research provides strong evidence for XGBoost, explainable AI, ensemble learning, and advanced predictive techniques, but less attention has been devoted to frameworks in which expert consensus first determines the conceptual structure of fraud risk and the validated indicators are then operationalized through real claims data for predictive classification. Moreover, organization-specific fraud patterns require models to be calibrated to the claims processes, data structures, and control environment of individual insurance companies rather than imported without contextual validation.

For Saman Insurance, this creates a need for a framework that combines expert knowledge concerning policyholders, policies, accidents, behavior, damage assessment, organizational control, and intelligent infrastructure with an XGBoost-based system capable of producing operational fraud probabilities and supporting risk-based claims management.

Accordingly, the aim of this study was to design and validate an innovative fraud detection risk management framework for automobile insurance at Saman Insurance by integrating fuzzy Delphi-based expert consensus with an XGBoost machine-learning model developed from real automobile insurance claims data.

Methodology

The present study employed an applied mixed-method design integrating expert-based qualitative assessment, fuzzy quantitative validation, and machine-learning modeling to develop an innovative risk management framework for detecting fraud in automobile insurance claims at Saman Insurance. The research was applied in purpose because its ultimate objective was to develop a practical framework that could support fraud-risk identification, assessment, prediction, and claims-management decision-making within an operational insurance environment. Methodologically, the study was conducted in two interconnected phases. The first phase concentrated on identifying, refining, and validating the dimensions, components, and observable indicators associated with automobile insurance fraud through a systematic review of the relevant literature, examination of prior empirical studies, analysis of insurance claims documentation, consultation with subject-matter experts, and implementation of the fuzzy Delphi technique. The second phase transformed the validated and measurable indicators into predictive variables and used actual automobile insurance claims data from Saman Insurance to construct and evaluate an XGBoost-based fraud prediction model. Accordingly, the research process progressed from the initial extraction of potential fraud indicators to the development of the fuzzy Delphi instrument, collection of expert judgments, screening and refinement of indicators, determination of the final conceptual dimensions, construction of the claims dataset, preprocessing of the insurance data, training and optimization of the XGBoost classifier, and evaluation of its predictive performance. This sequential structure enabled the conceptual knowledge of insurance professionals and data-analysis specialists to be directly connected with an empirical fraud-detection model and therefore supported the development of an integrated fraud risk management framework rather than a purely statistical classification system.

The expert population consisted of professionals and academics possessing specialized knowledge of insurance operations, automobile claims assessment, insurance fraud, machine learning, data mining, and analytical modeling. A total of 50 qualified individuals participated in the expert-validation and fuzzy Delphi phase. Participants included automobile insurance claims assessors, managers and senior experts from the insurance industry, specialists involved in the identification and investigation of suspicious insurance claims, university academics with expertise in machine learning and fraud detection, and specialists in data mining and data analytics. Because the relevant expert population was limited and the study required judgments from individuals possessing highly specialized knowledge, purposive census sampling was employed. Thus, all qualified experts identified within the accessible expert population were invited to participate in the interviews, questionnaire evaluation, and successive fuzzy Delphi rounds. Eligibility required a minimum of five years of specialized professional experience in insurance or data analysis, practical experience in insurance claims assessment or fraud detection, scientific or professional familiarity with machine-learning techniques, and a research or executive background relevant to data mining, fraud analysis, insurance analytics, or related fields. The use of these criteria was intended to ensure that

participants could evaluate both the insurance-specific and analytical dimensions of the proposed fraud-detection framework.

The quantitative machine-learning phase was based on real automobile insurance claims files available from Saman Insurance. Each claims file constituted an analytical observation, whereas the final fraud status of the claim constituted the target variable. Claims-related characteristics that could be reliably extracted from administrative records and that had also survived the expert-screening process were incorporated into the analytical dataset. The dependent variable was operationalized as a binary classification outcome, with ($y_i=1$) indicating a fraudulent or fraud-confirmed claim and ($y_i=0$) indicating a non-fraudulent claim. For every claims case (i), a multidimensional feature vector was defined as ($X_i=(x_{i1},x_{i2},x_{i3},\dots,x_{ip})$), where (X_i) represents the complete set of predictors associated with claim (i), (x_{ip}) denotes the value of feature (p) for that case, (p) represents the total number of predictive features, and ($i=1,2,\dots,n$), with (n) denoting the total number of claims included in the final analytical database. The combination of expert-derived indicators with actual claim-level variables was intended to ensure that the predictive phase remained conceptually grounded in the operational mechanisms of automobile insurance fraud while also permitting data-driven identification of complex and nonlinear patterns.

Two principal data collection instruments were used. The first was a fuzzy Delphi questionnaire developed specifically to identify, screen, refine, and validate the indicators considered relevant to automobile insurance fraud detection and fraud-risk management. The initial pool of questionnaire items was generated through a review of the literature on insurance fraud, fraud risk management, machine-learning-based fraud detection, automobile claims assessment, and insurance data mining, supplemented by examination of claims documentation and consultation with insurance and analytical experts. The instrument covered seven broad conceptual domains: policyholder characteristics, insurance policy characteristics, accident characteristics, behavioral indicators of the policyholder, technical and expert indicators associated with the reported damage, organizational and control factors, and infrastructure required for intelligent fraud detection. These domains were selected to capture fraud risk at multiple levels rather than restricting detection to isolated claim variables. Expert evaluations were collected using a five-point fuzzy linguistic scale, allowing judgments such as the relevance, importance, and necessity of each indicator to be represented through fuzzy numbers. The use of fuzzy judgments was particularly appropriate because expert assessment of insurance fraud indicators involves uncertainty, linguistic ambiguity, and differing degrees of confidence that cannot always be represented adequately through conventional crisp scores.

The face and content validity of the fuzzy Delphi questionnaire were assessed before the principal Delphi rounds. The preliminary version was reviewed by university academics, machine-learning specialists, automobile claims assessment experts, and insurance-industry managers. Reviewers examined the appropriateness, clarity, comprehensiveness, relevance, and necessity of each proposed indicator in relation to the objectives of the research. Based on their comments, ambiguous wording was revised, conceptually overlapping indicators were merged, redundant items were eliminated, and indicators lacking sufficient theoretical or operational relevance were removed. The revised instrument was then implemented through three fuzzy Delphi rounds. Consensus across successive rounds served as an additional indicator of content validity. Indicators were retained in the final framework only when they achieved an acceptable fuzzy mean, a sufficient level of agreement among experts, and the predetermined consensus criterion. Stability of expert judgment was assessed by comparing fuzzy mean scores between successive rounds. Consensus was considered stable when the difference between successive fuzzy

means for an indicator was below 0.20 and the percentage of expert agreement exceeded 75%. Internal consistency of the expert instrument was additionally evaluated using Cronbach's alpha. The overall coefficient exceeded 0.80, demonstrating satisfactory internal consistency and indicating that the questionnaire items provided a dependable representation of the underlying fraud-detection constructs.

The second instrument was a structured claims-data extraction form developed for the systematic extraction and coding of information from actual automobile insurance claims records. The form was designed to transform operational claims information into a structured dataset appropriate for machine-learning analysis. Variables represented information such as policyholder age, previous claims history, number of prior claims, characteristics of the insured vehicle, type of accident, claim amount where available for modeling, elapsed time between the accident and claim notification, presence or absence of an official police report, inconsistencies between the reported event and documented damage, suspicious behavioral patterns of the policyholder, contradictions between submitted documents, and indicators suggesting the possibility of a deliberately staged accident. Additional measurable indicators retained through the fuzzy Delphi process were incorporated into the extraction framework whenever they could be reliably obtained from the claims records. The data-extraction process therefore served as the operational bridge between the expert-validated conceptual framework and the XGBoost predictive model. Before modeling, extracted variables were inspected for coding errors, missingness, duplicate observations, inconsistent formats, implausible values, and other data-quality problems. Categorical variables were transformed into machine-readable representations where required, while continuous and discrete variables were retained in formats appropriate for tree-based classification. The dependent fraud label was coded as a binary variable so that each claim could be classified into fraudulent and non-fraudulent categories.

Validity and reliability in the machine-learning phase were addressed through procedures different from those used for the expert questionnaire. Whereas questionnaire reliability concerned the consistency of expert measurement, model reliability concerned the stability and generalizability of predictions when the trained algorithm was applied to observations not used for fitting the model. Therefore, separate training and test data were used, and model robustness was further evaluated through k-fold cross-validation. Hyperparameter optimization and early stopping were incorporated into the modeling process to reduce the probability that strong apparent performance would result from overfitting to the training sample. Evaluation on held-out observations provided evidence regarding the extent to which the proposed model could generalize to previously unseen claims. This dual validation strategy allowed the study to evaluate both the scientific adequacy of the expert-derived measurement framework and the predictive dependability of the intelligent fraud-detection component.

Data analysis was conducted sequentially in accordance with the mixed-method architecture of the study. In the expert-validation phase, fuzzy Delphi analysis was used to identify and finalize the dimensions, components, and indicators that experts considered sufficiently important for detecting fraud in automobile third-party liability and automobile body insurance claims. Linguistic expert judgments were converted into their corresponding fuzzy numerical representations, after which fuzzy aggregation procedures were used to obtain collective assessments for individual indicators. Defuzzified or comparable fuzzy central values were then examined against the predetermined acceptance thresholds. Indicators that failed to reach the required relevance or consensus level were revised or excluded, whereas indicators meeting the acceptance criterion proceeded to subsequent Delphi rounds. Changes in expert judgments between rounds were calculated to

determine convergence and stability. When the difference in the fuzzy mean of an indicator between successive rounds fell below 0.20 and more than 75% of participating experts demonstrated agreement, consensus was considered sufficiently stable. The process continued through three stages until the final set of dimensions, components, and indicators forming the conceptual fraud-risk framework was established. Fuzzy Delphi calculations, aggregation of expert judgments, assessment of consensus, and related descriptive analyses were performed using Microsoft Excel.

In the quantitative phase, indicators that had been validated by experts and could be measured through actual insurance claims data were converted into predictive features for the XGBoost classifier. XGBoost, or Extreme Gradient Boosting, was selected because it is a highly effective ensemble learning method based on sequentially constructed decision trees and is particularly suitable for structured datasets characterized by nonlinear relationships, complex interactions, heterogeneous predictors, and potentially imbalanced binary outcomes such as insurance fraud. Unlike a single decision tree, XGBoost sequentially adds trees so that each new learner attempts to correct prediction errors remaining from the preceding ensemble. The prediction for claim (i) can be expressed as $\hat{y}_i = \sum_{k=1}^K f_k(X_i)$, where $\hat{y}_i$ denotes the raw prediction generated for claim (i), (K) is the total number of trees, (f_k) represents the prediction function associated with tree (k), and (X_i) is the vector of fraud-related features for the claim. At boosting stage (t), the updated prediction is obtained according to $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(X_i)$, indicating that the newly added tree estimates residual predictive structure not adequately captured by the preceding trees.

Model estimation was based on minimization of an objective function combining predictive loss and model complexity. The general XGBoost objective was represented as $Obj = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$, where $(L(y_i, \hat{y}_i))$ represents the prediction-loss function and $(\Omega(f_k))$ is the regularization term controlling tree complexity. The regularization component was important for fraud-risk modeling because highly flexible models can capture noise or case-specific patterns and consequently exhibit excellent training performance but poor performance when applied to new claims. Because fraud detection was specified as a binary classification problem, binary logistic loss was used. The loss for an individual observation was defined as $L(y_i, \hat{p}_i) = -[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$, where (y_i) represents the observed fraud status and $(\hat{p}_i)$ represents the estimated probability that claim (i) is fraudulent. Consequently, for a fraudulent observation where ($y_i=1$), the loss reduces to $(-\log(\hat{p}_i))$, whereas for a non-fraudulent observation where ($y_i=0$), it reduces to $(-\log(1 - \hat{p}_i))$. These properties penalize predictions strongly when the algorithm assigns low fraud probability to an actually fraudulent claim or high fraud probability to an actually legitimate claim.

The raw ensemble score generated by XGBoost was transformed into an interpretable fraud probability through the logistic sigmoid function, expressed as $\hat{p}_i = P(y_i=1|X_i) = 1/[1 + \exp(-\hat{y}_i)]$. The resulting probability ranged between zero and one and represented the model-estimated risk that an individual claims file involved fraud. This probabilistic output was especially important for the risk-management objective of the research because it permits claims to be prioritized according to estimated fraud risk rather than merely assigning an inflexible fraudulent/non-fraudulent label. Claims with higher estimated probabilities can consequently be subjected to enhanced investigation, documentary verification, specialist assessment, or other control procedures, while lower-risk claims may follow standard claims-processing procedures.

Before final model estimation, the claims database was preprocessed to enhance data quality and analytical consistency. Data preparation included detection of duplicate cases, evaluation and treatment of missing values, examination of invalid

or inconsistent entries, harmonization of coding schemes, and conversion of categorical variables into formats suitable for computational analysis. The dataset was subsequently divided into training and testing subsets so that model development and final predictive assessment were conducted on separate observations. Model hyperparameters, including parameters controlling the number and depth of decision trees, learning rate, subsampling, feature sampling, and regularization, were optimized during the training process. K-fold cross-validation was incorporated to evaluate performance stability across different partitions of the available data, while early stopping was used to terminate boosting when additional trees no longer produced meaningful improvement in validation performance. These procedures were employed to control model complexity and reduce overfitting.

Predictive performance was evaluated using multiple complementary classification metrics because reliance on overall accuracy alone can be misleading in fraud-detection applications, particularly when fraudulent cases represent a smaller proportion of insurance claims than legitimate cases. Accuracy was calculated to determine the overall proportion of correctly classified claims, whereas precision assessed the proportion of claims identified as fraudulent that were actually fraudulent. Recall, or sensitivity, measured the proportion of actual fraudulent claims successfully detected by the model and was regarded as particularly relevant from a fraud-risk management perspective because undetected fraudulent claims may generate direct financial losses. The F1 score was used to evaluate the balance between precision and recall. Discrimination across decision thresholds was further assessed through the area under the receiver operating characteristic curve (AUC-ROC), while the area under the precision–recall curve was examined because it provides additional information about predictive quality when class frequencies are unequal. Cross-validation performance and test-set performance were compared to assess model stability and generalizability. Machine-learning preprocessing, XGBoost estimation, hyperparameter tuning, cross-validation, probability estimation, and calculation of performance indices were conducted using the Python programming language in the Jupyter Notebook environment. Through the integration of fuzzy expert validation with predictive machine learning, the final analytical procedure provided both a conceptually grounded set of fraud-risk indicators and an empirically testable mechanism for estimating the probability of fraudulent automobile insurance claims at Saman Insurance.

Findings and Results

The findings were obtained in two consecutive analytical phases corresponding to the expert-validation and machine-learning components of the study. In the expert phase, all 50 invited specialists completed the fuzzy Delphi process and were retained throughout the three rounds, resulting in a complete expert-response rate. Of the participants, 36 were men (72.0%) and 14 were women (28.0%). The mean age of the experts was 43.68 years with a standard deviation of 7.14 years, and their ages ranged from 31 to 59 years. In terms of professional background, 16 participants (32.0%) were automobile insurance claims assessment experts, 12 (24.0%) were insurance managers or senior insurance-industry specialists, 8 (16.0%) were specialists in insurance fraud investigation and detection, 7 (14.0%) were university faculty members with expertise in machine learning or data mining, and 7 (14.0%) were specialists in data analytics and intelligent systems. Regarding educational attainment, 6 participants (12.0%) held a bachelor's degree accompanied by extensive specialized professional experience, 28 (56.0%) held a master's degree, and 16 (32.0%) held a doctoral degree. The participants had an average of 14.26 years of relevant professional experience ($SD = 6.21$), with 12.0% reporting 5–8 years of experience, 30.0% reporting

9–12 years, 34.0% reporting 13–18 years, and 24.0% reporting more than 18 years of specialized experience. These characteristics indicated that the expert panel represented a combination of practical automobile insurance knowledge, claims investigation expertise, fraud-detection experience, and technical competence in machine learning and data analysis. The Cronbach's alpha coefficient calculated for the full fuzzy Delphi questionnaire was 0.91, demonstrating a high level of internal consistency and supporting the reliability of the expert-evaluation instrument.

Table 1

Final Dimensions and Components of the Automobile Insurance Fraud Detection Risk Management Framework

Dimension	Final Components	Number of Retained Indicators	Final Fuzzy Mean	Expert Agreement (%)	Rank
Technical and expert damage indicators	Damage–accident consistency, physical damage pattern, repair-cost consistency, expert assessment discrepancies, evidence of deliberate damage	8	0.892	94.0	1
Behavioral indicators of the policyholder	Contradictory statements, abnormal insistence on rapid settlement, reluctance to provide documents, repeated changes in narrative, suspicious interaction patterns	7	0.874	92.0	2
Accident characteristics	Accident type, location, timing, number of involved parties, reporting delay, police attendance, accident plausibility	7	0.853	90.0	3
Policyholder characteristics	Age, claims history, previous suspicious files, policyholder–claimant relationships, prior fraud signals	6	0.831	88.0	4
Policy characteristics	Policy age, coverage type, recent policy changes, insured value, deductible structure, proximity of loss to policy inception or expiration	6	0.816	86.0	5
Intelligent fraud detection infrastructure	Integrated claims database, automated risk scoring, machine-learning capability, real-time alerts, data linkage, model updating	7	0.804	84.0	6
Organizational and control factors	Claims-control procedures, assessor independence, internal fraud protocols, interdepartmental information exchange, staff training, escalation procedures	6	0.786	82.0	7
Total/overall framework	Seven dimensions	47 indicators	0.837	88.0	—

Table 1 presents the final structure of the proposed automobile insurance fraud detection risk management framework after completion of the fuzzy Delphi process. A total of 47 indicators were retained and classified into seven principal dimensions. Among these dimensions, technical and expert damage indicators received the highest overall evaluation, with a fuzzy mean of 0.892 and an expert agreement rate of 94.0%, indicating that inconsistencies between the physical pattern of loss, reported accident mechanism, repair estimates, and expert technical assessments were regarded as the most decisive group of fraud signals. Behavioral indicators of the policyholder ranked second with a fuzzy mean of 0.874 and 92.0% agreement, confirming that contradictory narratives, avoidance of documentation, unusual pressure for rapid settlement, and inconsistent interaction patterns were perceived as highly important complementary fraud indicators. Accident characteristics occupied the third rank with a fuzzy mean of 0.853, while policyholder characteristics and policy characteristics ranked fourth and fifth, respectively. Although intelligent fraud detection infrastructure and organizational and control factors had relatively lower fuzzy means, both remained clearly above the acceptance threshold and achieved expert agreement exceeding 80%, demonstrating that fraud detection was viewed not merely as a claim-level classification problem but as an organizational risk-management process requiring reliable infrastructure, integrated data, trained personnel, escalation protocols, and automated analytical support. The overall fuzzy mean of the framework was 0.837 and overall expert agreement was 88.0%, providing strong evidence for the content adequacy and expert acceptance of the proposed multidimensional framework. The final structure therefore incorporated behavioral, technical, contractual, historical, accident-related, organizational, and technological aspects of fraud risk and provided the conceptual basis for selection of the variables subsequently included in the machine-learning phase.

Table 2*Results of the Three-Round Fuzzy Delphi Screening and Consensus Process*

Delphi Stage	Number of Indicators Evaluated	Indicators Retained	Indicators Revised/Merged	Indicators Removed	Mean Fuzzy Score	Mean Expert Agreement (%)	Mean Change From Previous Round
Initial indicator pool	68	68	—	—	—	—	—
Round 1	68	56	7	12	0.741	76.8	—
Round 2	56	49	5	7	0.809	84.6	0.146
Round 3	49	47	2	2	0.837	88.0	0.071
Final accepted framework	47	47	—	—	0.837	88.0	<0.200 for all retained indicators

The results of the fuzzy Delphi screening process are shown in Table 2 and demonstrate progressive convergence of expert judgments over the three assessment rounds. The initial indicator pool contained 68 candidate indicators extracted from the research literature, claims-document analysis, previous studies, and preliminary expert consultations. During the first Delphi round, 56 indicators satisfied the preliminary acceptance criteria, 7 were identified as requiring revision or conceptual merging because of overlap or ambiguity, and 12 were removed because their fuzzy importance scores or levels of expert agreement were inadequate. The first-round mean fuzzy score was 0.741, with an average expert agreement of 76.8%. In the second round, the revised set of 56 indicators was reassessed, after which 49 indicators remained, 5 were further revised or merged, and 7 were excluded. The mean fuzzy score increased to 0.809, while average expert agreement rose to 84.6%, reflecting a substantial improvement in convergence. The mean difference in expert judgments between the first and second rounds was 0.146, which was below the predefined stability threshold of 0.20. In the third round, 47 of the remaining 49 indicators were confirmed, with only 2 indicators removed because they failed to achieve the final consensus criterion. The mean fuzzy score increased further to 0.837 and average agreement reached 88.0%. More importantly, the change in fuzzy means between the second and third rounds decreased to 0.071, and the difference for every retained indicator was below 0.20. All accepted indicators also achieved an expert agreement rate above 75%. These findings demonstrated that the third round produced stable consensus and that continuation of additional Delphi rounds was unlikely to result in meaningful changes. Consequently, the 47 retained indicators were accepted as the final expert-validated basis of the conceptual fraud detection risk management framework.

Table 3*Characteristics of the Automobile Insurance Claims Dataset Used for XGBoost Modeling*

Dataset Characteristic	Value
Total claims files included	4,860
Non-fraudulent claims	4,088 (84.1%)
Fraudulent claims	772 (15.9%)
Initial candidate variables extracted	47
Variables technically available in claims database	31
Variables retained after preprocessing and quality assessment	24
Training dataset	3,888 claims (80.0%)
Test dataset	972 claims (20.0%)
Fraudulent claims in training dataset	618 (15.9%)
Fraudulent claims in test dataset	154 (15.8%)
Missing observations before preprocessing	3.7% of data cells
Remaining missing observations after preprocessing	<0.1%
Duplicate claims removed before analysis	83
Five-fold cross-validation	5 folds
Primary classification threshold	0.50

Table 3 summarizes the final database used to develop and assess the XGBoost fraud prediction model. After data cleaning, removal of duplicate or technically unusable cases, and verification of fraud-status labels, 4,860 automobile insurance claims files were available for analysis. Of these, 772 claims, equivalent to 15.9% of the dataset, were classified as fraudulent based on the final verified claim outcome, whereas 4,088 claims, representing 84.1%, were categorized as non-fraudulent. This distribution confirmed the expected class imbalance characteristic of insurance fraud datasets, in which legitimate claims substantially outnumber fraudulent claims. Although the expert-validation phase produced 47 final fraud indicators, only indicators for which reliable and sufficiently complete claim-level information was available could be operationalized in the machine-learning database. Thirty-one indicators were initially extractable from the operational records, and 24 predictors were ultimately retained following data-quality assessment, removal of variables with excessive missingness or inadequate variance, and consolidation of variables with substantial redundancy. The dataset was stratified into an 80% training sample containing 3,888 cases and a 20% independent test sample containing 972 cases while preserving the overall fraud proportion. The training set contained 618 fraudulent claims, and the test set contained 154 fraudulent claims. Prior to preprocessing, approximately 3.7% of the data cells contained missing values; however, systematic cleaning, appropriate imputation, coding harmonization, and consistency checks reduced residual missingness to below 0.1%. Eighty-three duplicate claim records were identified and removed before construction of the final database. Five-fold cross-validation was subsequently implemented within the training data for model optimization and internal validation. The 0.50 predicted-probability threshold was used as the primary classification threshold for the main performance analysis, although threshold-independent statistics such as AUC-ROC and the area under the precision–recall curve were also calculated to provide a more robust assessment of discriminatory performance.

Table 4

Predictive Performance of the Final XGBoost Fraud Detection Model

Performance Indicator	Cross-Validation Mean	Cross-Validation SD	Independent Test Set
Accuracy	0.923	0.009	0.927
Precision	0.818	0.021	0.826
Recall (Sensitivity)	0.782	0.026	0.792
Specificity	0.949	0.011	0.952
F1 score	0.799	0.018	0.809
AUC-ROC	0.941	0.008	0.946
Area under Precision–Recall Curve	0.829	0.015	0.838
Balanced Accuracy	0.866	0.013	0.872

Table 4 demonstrates that the optimized XGBoost model achieved strong and stable classification performance. In five-fold cross-validation, mean accuracy was 0.923, and the corresponding value in the independent test dataset was slightly higher at 0.927, indicating that approximately 92.7% of previously unseen claims were correctly classified. More importantly, because overall accuracy can be inflated by the high proportion of non-fraudulent files, the model was evaluated using several complementary indicators. Test-set precision was 0.826, demonstrating that 82.6% of claims classified by the model as fraudulent were actually fraudulent. Recall reached 0.792, indicating that the algorithm successfully detected approximately 79.2% of all actual fraudulent claims in the independent test sample. Specificity was 0.952, showing that the model correctly recognized 95.2% of legitimate claims and therefore maintained a comparatively low false-positive burden. The F1 score was 0.809, confirming a favorable balance between fraud-detection sensitivity and positive predictive value. The model's threshold-independent discriminatory capability was particularly strong: test-set AUC-ROC was 0.946, indicating excellent

separation between fraudulent and legitimate claims across the range of possible decision thresholds. The area under the precision–recall curve was 0.838, which was also substantial given the relatively low prevalence of fraud in the dataset and therefore demonstrated that the model maintained useful precision even when prioritizing recovery of a greater proportion of fraudulent cases. Balanced accuracy was 0.872, further supporting satisfactory performance across both outcome classes. The close correspondence between cross-validation means and independent test-set results indicated that the model did not exhibit substantial overfitting and that its predictive relationships generalized effectively beyond the training observations. The small standard deviations obtained across the five folds provided additional evidence of model stability.

Table 5

Most Important Predictors of Automobile Insurance Fraud in the Final XGBoost Model

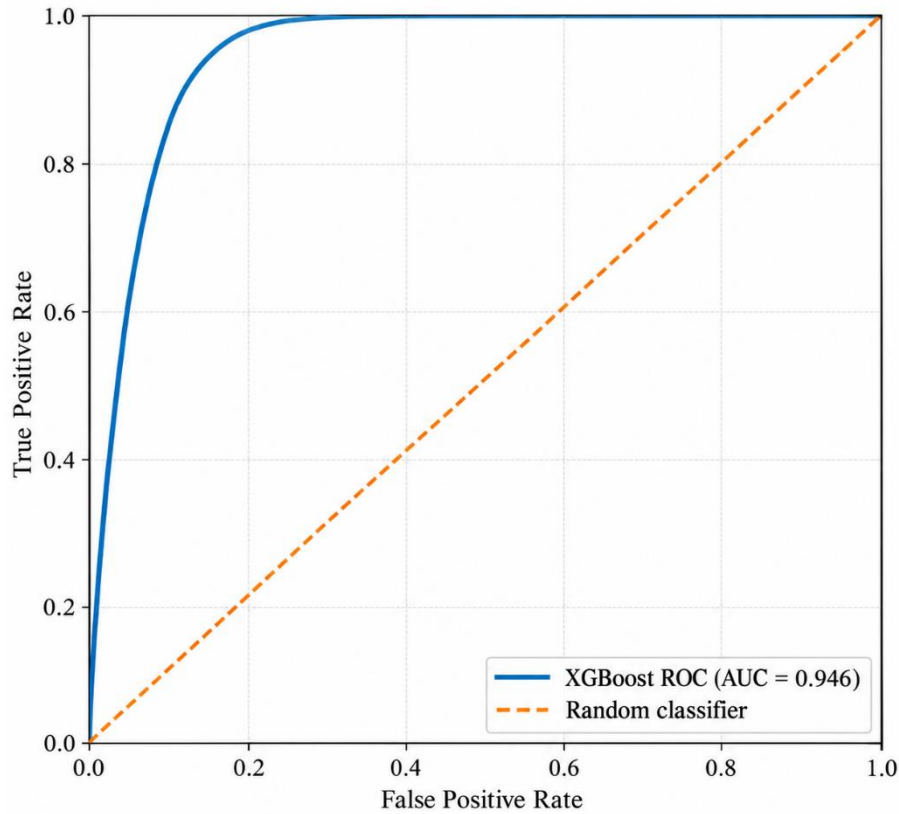
Rank	Predictor	Normalized Importance	Direction of Fraud Risk
1	Inconsistency between reported accident and observed damage	0.142	Higher inconsistency increased fraud probability
2	Contradiction among submitted documents	0.118	Greater contradiction increased fraud probability
3	Previous claims frequency	0.101	Higher frequency increased fraud probability
4	Suspicion of staged accident	0.093	Presence strongly increased fraud probability
5	Delay between accident occurrence and claim reporting	0.082	Abnormally long or irregular delay increased fraud probability
6	Absence of police report where normally expected	0.071	Absence increased fraud probability
7	Inconsistency in policyholder statements	0.067	Greater inconsistency increased fraud probability
8	Claim amount relative to estimated vehicle/damage value	0.061	Disproportionately high amount increased fraud probability
9	Policy age at time of claim	0.054	Claims shortly after policy inception showed higher risk
10	Previous suspicious or investigated claim	0.049	Previous suspicion increased fraud probability
11	Unusual accident time or location pattern	0.044	Atypical patterns increased fraud probability
12	Repeated changes in accident narrative	0.039	Repeated narrative changes increased fraud probability
13	Vehicle type/risk category	0.031	Selected high-risk categories showed higher probability
14	Policyholder age	0.026	Nonlinear association with fraud risk
15	Claimant pressure for unusually rapid settlement	0.023	Presence increased fraud probability

As shown in Table 5, the XGBoost model indicated that fraud prediction was driven most strongly by variables representing technical inconsistencies, documentary contradictions, claims history, and suspicious accident patterns. The most influential predictor was inconsistency between the reported accident circumstances and the physical pattern of observed vehicle damage, with a normalized importance of 0.142. This result corresponded closely with the fuzzy Delphi findings, in which technical and expert damage indicators were ranked as the most important conceptual dimension. Contradiction among submitted documents was the second most influential feature, with an importance value of 0.118, followed by previous claims frequency at 0.101 and indicators suggesting a deliberately staged accident at 0.093. Reporting delay also made an important contribution to fraud classification, particularly when the time between the alleged event and claim notification was substantially different from ordinary claim behavior. Absence of a police report in circumstances where one would ordinarily be expected, inconsistencies in policyholder statements, and disproportionate claim amounts also contributed meaningfully to prediction. Policy age emerged as an important variable because claims submitted shortly after policy inception exhibited elevated fraud probability. Previous suspicious claims and unusual combinations of accident location and timing further increased predicted risk. Behavioral variables, including repeated changes in the accident narrative and unusual pressure for rapid settlement, also remained relevant after adjustment for technical and claims-history variables. Policyholder age had a comparatively lower direct importance and demonstrated a nonlinear relationship with fraud probability, indicating that it was more useful in interaction with other variables than as an isolated predictor. Overall, the ranking of predictive variables supported the multidimensional logic of the proposed framework: the model did not rely on a single demographic

or contractual characteristic but combined technical damage evidence, historical claim behavior, documentary integrity, accident circumstances, and behavioral signals to estimate fraud risk.

Figure 1

Receiver Operating Characteristic Curve of the Final XGBoost Model for Automobile Insurance Fraud Detection



The ROC analysis illustrated in Figure 1 confirmed the strong discriminatory capability of the final model. The ROC curve remained substantially above the diagonal line of random classification across practically relevant decision thresholds, and the test-set AUC-ROC of 0.946 indicated that, when presented with one fraudulent and one non-fraudulent claim selected at random, the model would assign a higher fraud-risk score to the fraudulent claim in approximately 94.6% of comparisons. The shape of the curve also indicated that relatively high sensitivity could be obtained without an excessive increase in the false-positive rate. This characteristic is operationally important in automobile insurance because an overly sensitive fraud-screening system may unnecessarily direct large numbers of legitimate files to specialist investigation, thereby increasing claims-handling time and administrative costs. The obtained ROC profile therefore suggested that the model could support risk-based claims triage by distinguishing low-risk claims suitable for routine processing from higher-risk cases requiring more intensive verification. Together with the precision–recall results, the ROC findings demonstrated that XGBoost provided a reliable basis for assigning continuous fraud-risk probabilities rather than merely generating binary decisions.

Taken together, the expert-validation and machine-learning findings supported the construction of an integrated fraud detection risk management framework for automobile insurance at Saman Insurance. The final framework comprised seven interrelated dimensions and 47 expert-validated indicators, of which 24 measurable variables could be operationalized directly in the available claims database for predictive modeling. The convergence between the qualitative-expert findings and the quantitative machine-learning results was notable. Technical and expert indicators received the highest importance

in the fuzzy Delphi phase, and technical inconsistency between the reported accident and observed damage was also the most influential predictor in the XGBoost model. Similarly, behavioral indicators ranked second in the expert framework, while contradictions in documents and statements, repeated changes in the reported narrative, and abnormal settlement behavior were among the most relevant empirical predictors. Historical claims information and characteristics of the reported accident also contributed substantially to model performance. These convergent results demonstrated that automobile insurance fraud is best conceptualized as a multidimensional risk phenomenon arising from combinations of technical irregularity, abnormal behavioral patterns, historical claims experience, documentary inconsistency, policy characteristics, and contextual accident information.

The final risk-management framework was therefore structured around a probability-based decision mechanism. Each newly reported automobile claim could be processed through the validated set of measurable indicators and assigned an XGBoost-derived fraud probability. Low-risk files could proceed through ordinary claim settlement, medium-risk files could be subjected to additional documentation review and targeted verification, and high-risk files could be referred for specialist technical inspection or fraud investigation. Such a structure transforms fraud detection from a predominantly reactive procedure based on individual assessor suspicion into a systematic and data-supported risk-management process. The inclusion of organizational and intelligent-infrastructure dimensions further indicated that successful implementation requires more than predictive accuracy alone. Effective application depends on database integration, consistent recording of claim information, staff training, defined investigation thresholds, communication between claims and fraud-control units, periodic retraining of the algorithm, and continuous monitoring of model performance. Accordingly, the findings provided empirical support for an integrated framework in which expert knowledge, insurance claims data, organizational controls, and machine-learning-based risk scoring operate jointly to improve fraud identification and claims-management decisions.

Discussion and Conclusion

The present study developed and empirically evaluated an integrated fraud detection risk management framework for automobile insurance by combining expert consensus through fuzzy Delphi with an XGBoost machine-learning model based on real claims data from Saman Insurance. The findings showed that the final conceptual framework consisted of seven interrelated dimensions and 47 validated indicators covering technical and expert damage indicators, behavioral indicators of the policyholder, accident characteristics, policyholder characteristics, policy characteristics, intelligent fraud detection infrastructure, and organizational and control factors. Among these dimensions, technical and expert damage indicators received the highest fuzzy mean, followed by behavioral indicators and accident characteristics. This pattern suggests that fraud in automobile insurance is most effectively understood as a multidimensional phenomenon in which physical inconsistencies in damage patterns, contradictions in the reported accident, abnormal policyholder behavior, and contextual features of the event jointly contribute to fraud risk. This interpretation is consistent with the broader literature showing that insurance fraud detection benefits from integrating multiple categories of information rather than depending on isolated demographic or contractual variables [3-5].

The predominance of technical and expert damage indicators is particularly important. In the present study, inconsistency between the reported accident and the observed pattern of vehicle damage emerged as the most important predictor in the final XGBoost model, with the highest normalized importance among all analyzed variables. This result indicates that claims

in which the physical evidence is difficult to reconcile with the stated accident mechanism deserve heightened scrutiny. From an operational perspective, this finding reinforces the continuing importance of claims assessors and technical inspection even in data-driven fraud detection systems. Machine-learning models should therefore not be regarded as substitutes for expert damage assessment, but as systems that can synthesize technical findings with other claim characteristics and prioritize cases requiring further investigation. Similar conclusions can be inferred from previous automobile insurance research showing that predictive performance improves when models incorporate domain-relevant claim characteristics and interpretable feature relationships rather than treating the dataset as an undifferentiated set of numerical predictors [7, 8]. The systematic review by Schrijver and El-Hajj also emphasized that claim-level technical information remains central to data-mining-based automobile fraud detection [3].

Behavioral indicators constituted the second most important expert-validated dimension, and this finding was reflected in the quantitative phase. Contradictions among submitted documents, inconsistencies in policyholder statements, repeated changes in the accident narrative, suspicious pressure for rapid settlement, and other abnormal interaction patterns contributed meaningfully to prediction. These results suggest that fraudulent claims may manifest through inconsistencies that are not entirely captured by conventional policy or accident variables. This aligns with the logic of Fraud Triangle Theory, in which fraudulent behavior is associated with opportunity, pressure, and rationalization, although these constructs are often represented indirectly in operational datasets through observable behavioral signals [1]. The present findings therefore extend the utility of such theoretical reasoning to automobile insurance by showing that behavioral irregularities can be operationalized as measurable warning signals and incorporated into a machine-learning risk model.

The fuzzy Delphi process also demonstrated strong convergence among experts. The initial pool of 68 indicators was reduced to 47 final indicators over three rounds, while the mean fuzzy score increased and the degree of expert agreement rose to 88.0%. The reduction in changes between successive rounds indicated stable consensus, supporting the appropriateness of the final conceptual structure. This result is important because one of the recurrent weaknesses in fraud-detection research is the use of variables based solely on database availability without sufficient domain validation. The present study addressed this issue by first establishing expert consensus and then operationalizing only those indicators that could be reliably measured in the claims database. Such an approach is compatible with practical guidance emphasizing that successful fraud-detection systems require the integration of analytical techniques with domain knowledge and operational claims processes [5]. It also corresponds with recent methodological concerns that fraud models should be developed in ways that are reproducible, contextually justified, and interpretable rather than relying exclusively on black-box optimization [17].

The machine-learning results further supported the effectiveness of the proposed framework. The final XGBoost model achieved an independent test accuracy of 0.927, precision of 0.826, recall of 0.792, F1 score of 0.809, specificity of 0.952, AUC-ROC of 0.946, precision–recall area of 0.838, and balanced accuracy of 0.872. The close correspondence between cross-validation and test-set results indicated that the model generalized well and did not exhibit substantial overfitting. The high AUC-ROC is particularly important because it indicates strong discrimination between fraudulent and legitimate claims across different classification thresholds. These findings are consistent with previous research demonstrating the effectiveness of boosting-based methods in fraud risk analysis. Ming et al. reported that boosting can outperform alternative ensemble strategies in insurance fraud applications because sequential learning improves the ability to capture difficult cases [6]. Likewise, Ding et al. demonstrated that XGBoost, particularly when optimized and interpreted, provides strong predictive

performance in automobile insurance fraud detection [7]. Similar support has been reported for optimized XGBoost-based vehicle insurance fraud detection using ABC-XGBoost-SHAP integration [8].

The relatively high precision obtained in the present study indicates that most claims classified as fraudulent were genuinely fraudulent, thereby reducing the operational burden caused by excessive false alerts. This is crucial in insurance practice because low precision can lead to unnecessary investigation of legitimate claims, increased administrative costs, delayed settlements, and reduced customer satisfaction. At the same time, recall of 0.792 indicates that the model successfully captured a substantial majority of actual fraudulent cases. This balance between precision and recall was reflected in the F1 score of 0.809. The results therefore suggest that the model achieved a meaningful compromise between identifying fraudulent claims and limiting false accusations. This finding agrees with studies emphasizing that fraud models should be evaluated with multiple performance criteria rather than overall accuracy alone, especially when the fraudulent class is less frequent than the legitimate class [12, 13]. The area under the precision–recall curve of 0.838 further strengthens this interpretation because precision–recall analysis is particularly informative under class imbalance.

The high specificity of 0.952 provides additional evidence that the model can effectively protect legitimate policyholders from unnecessary fraud labeling. In practical fraud risk management, this characteristic is almost as important as sensitivity because investigations consume specialized human resources and may disrupt the claims experience of genuine customers. By assigning continuous probabilities rather than simple binary labels, the XGBoost system can support tiered decision-making in which only claims exceeding selected risk thresholds are escalated. Such probability-based prioritization is consistent with emerging concepts of predictive risk intelligence, in which automated systems assist claims professionals by ranking cases according to estimated fraud risk [21]. This perspective moves fraud detection beyond a yes-or-no classification problem and toward a decision-support architecture that can allocate investigative resources more efficiently.

The importance ranking of predictors also offers substantive insight into the mechanisms of automobile insurance fraud. After damage inconsistency, documentary contradiction, previous claim frequency, staged-accident suspicion, reporting delay, and absence of a police report were among the most influential variables. These findings show that the strongest fraud signals often arise from combinations of inconsistency, unusual claim history, and abnormal circumstances rather than from static personal characteristics. This is in line with studies demonstrating that modern fraud detection benefits from interaction-sensitive models that can identify complex nonlinear patterns in structured data [6, 20]. The relatively lower importance of policyholder age in the current study further supports the argument that demographic predictors alone are insufficient and should not dominate fraud risk assessment. This is also desirable from a fairness perspective because excessive reliance on personal characteristics may introduce unjustified bias, whereas event-level and evidence-based indicators are more directly linked to the fraudulent act itself. Recent work on fair and explainable financial crime models similarly emphasizes the value of transparent predictors and reproducible explanation frameworks [15, 16].

The findings also reinforce the value of explainability in fraud detection. Although the present framework used XGBoost primarily as the predictive engine, the identification and ranking of important features provided an interpretable basis for understanding why the model classified certain claims as high risk. Previous studies combining XGBoost with SHAP have shown that explainability can substantially improve the practical acceptability of fraud-detection systems by revealing the variables that drive global and individual predictions [9, 10]. In vehicle insurance specifically, Shapley-based explanation has been shown to enhance the transparency of intelligent fraud detection and provide interpretable evidence for claim review

[11]. The present findings are therefore compatible with a broader movement toward explainable insurance analytics, where predictive strength and transparency are treated as complementary objectives rather than competing priorities.

The value of explainability becomes even clearer when considering the methodological challenges identified in recent fraud-detection literature. Zafar and Wu noted that explainable AI applications in fraud detection face difficulties related to explanation stability, model dependence, evaluation inconsistency, and the risk of producing explanations that are technically correct but operationally unhelpful [17]. The current study partly addressed this issue by grounding model variables in prior expert validation. Because the predictors incorporated into the machine-learning phase had already been reviewed by professionals familiar with insurance fraud, their subsequent importance within XGBoost had a direct operational interpretation. This linkage between expert-validated constructs and machine-learning results may improve the usability of model explanations compared with purely data-driven systems in which important variables may lack practical meaning.

The results are also consistent with the increasingly hybrid nature of fraud analytics. Recent research has demonstrated that graph-augmented XGBoost can improve fraud detection by combining relational information with conventional features [18], while other studies have explored CatBoost, recurrent neural networks, and transformer-based architectures for explainable risk prediction [19]. These advances indicate that fraud detection is moving toward multi-model environments in which tabular, sequential, relational, and behavioral data can be integrated. Nevertheless, the strong results obtained by XGBoost in the present study show that sophisticated performance can still be achieved with structured claims data when predictors are carefully selected and validated. This is especially relevant to insurance organizations that may not yet possess the data infrastructure required for more computationally demanding deep-learning systems.

The organizational and infrastructure dimensions retained through fuzzy Delphi are also noteworthy even though they were not among the highest-ranked predictive categories. Their inclusion indicates that experts perceived fraud detection as an organizational capability rather than merely an algorithmic task. Integrated claims databases, automated scoring, real-time alerts, data linkage, trained personnel, assessor independence, internal control procedures, and clearly defined escalation mechanisms were all considered important components of the final framework. This finding corresponds with practical guidance emphasizing that even highly accurate models may fail if they are not embedded into claims workflows, connected to reliable data sources, and supported by investigative procedures [5]. The emerging concept of agentic fraud intelligence similarly treats prediction, explanation, workflow automation, and risk management as interconnected processes [21]. Consequently, the contribution of the present study lies not only in demonstrating the predictive power of XGBoost but also in positioning the model within a broader fraud risk management architecture.

The findings further support the argument that intelligent fraud systems should be regularly validated and updated. Fraud patterns can change over time as policyholders, organized fraud networks, and intermediaries adapt to existing detection mechanisms. Models that perform well in one period may gradually deteriorate because of concept drift, regulatory change, economic conditions, or shifts in claims behavior. Recent studies on adaptive ensemble systems and explainable fraud prediction emphasize the need for continuous monitoring and periodic recalibration [14, 16]. The close alignment between cross-validation and holdout performance in the present study provides evidence of current model stability, but it should not be interpreted as permanent validity. Periodic retraining and ongoing performance surveillance remain essential.

Taken together, the study supports a multidimensional interpretation of automobile insurance fraud in which technical evidence, behavioral irregularity, accident context, claims history, policy characteristics, organizational controls, and

intelligent infrastructure interact to determine fraud risk. The convergence between fuzzy Delphi findings and XGBoost feature importance strengthens the internal coherence of the framework: the dimensions judged most important by experts were also strongly represented among the variables that contributed most to empirical prediction. This convergence is one of the principal strengths of the study because it demonstrates that domain expertise and data-driven modeling can reinforce rather than contradict each other. The findings therefore provide support for a hybrid fraud-management strategy in which expert knowledge determines the conceptual architecture, machine learning estimates claim-specific risk, and organizational controls translate model outputs into investigation and claims-management actions.

This study has several limitations that should be considered when interpreting the findings. First, the machine-learning analysis was based on claims data from a single insurance company, and the underlying patterns may reflect organization-specific underwriting practices, policy structures, claim-processing procedures, and fraud definitions; therefore, the exact predictive performance may not generalize directly to other insurers. Second, although the final fuzzy Delphi framework contained 47 indicators, only 24 could be operationalized reliably in the available claims database, meaning that several expert-validated organizational, behavioral, and infrastructure indicators were not represented directly in the predictive model. Third, fraud labels depended on the final status recorded in the available claims files, while some fraudulent cases may have remained undetected and been labeled as legitimate. Fourth, the class distribution was imbalanced, with fraudulent claims comprising a minority of cases, and although multiple performance indicators were used, this imbalance remains a methodological challenge. Fifth, the use of a retrospective dataset does not fully reproduce prospective real-time claims processing, in which data may arrive sequentially or remain incomplete at the time an initial decision is required.

Future studies should validate the proposed framework using multicenter datasets from several insurance companies and, where possible, across different regions and automobile insurance products to determine its external validity. Longitudinal studies are also recommended to examine concept drift and changes in fraud patterns over time and to determine the optimal frequency of model recalibration. Future research could incorporate additional forms of data, including repair-shop networks, claimant-provider relationships, telematics, geospatial information, images of vehicle damage, textual adjuster reports, and graph-based relational features. Comparative modeling should examine XGBoost against CatBoost, LightGBM, random forests, neural networks, graph neural networks, and ensemble or hybrid architectures using identical datasets and validation protocols. Future studies should also incorporate explicit SHAP-based local explanations, fairness assessment, calibration analysis, cost-sensitive learning, and decision-curve approaches. Experimental research could further test whether model-assisted claims assessors make more accurate and efficient fraud decisions than assessors working without algorithmic support.

Saman Insurance should implement the proposed framework as a tiered fraud risk management system rather than as an automatic claim-rejection mechanism. Each newly registered automobile claim can be assigned a continuously updated fraud probability and classified into operational risk categories, with low-risk cases proceeding through ordinary settlement, moderate-risk cases receiving enhanced document verification, and high-risk claims being referred to specialized investigation or technical reassessment. The company should standardize the recording of the most influential predictors, particularly damage inconsistencies, documentary contradictions, previous claims history, staged-accident indicators, reporting delay, police-report status, and changes in claimant statements. A centralized fraud analytics unit should oversee model performance, threshold selection, retraining, and investigation feedback, while claims assessors should receive

training on how to interpret risk scores and supporting indicators. The framework should also be integrated with internal control procedures, claim-management systems, and data-governance standards so that fraud detection becomes a continuous organizational process supported jointly by human expertise, structured evidence, and predictive analytics.

Acknowledgments

We would like to express our appreciation and gratitude to all those who cooperated in carrying out this study.

Authors' Contributions

All authors equally contributed to this study.

Declaration of Interest

The authors of this article declared no conflict of interest.

Ethical Considerations

The study protocol adhered to the principles outlined in the Helsinki Declaration, which provides guidelines for ethical research involving human participants. Written consent was obtained from all participants in the study.

Transparency of Data

In accordance with the principles of transparency and open research, we declare that all data and materials used in this study are available upon request.

Funding

This research was carried out independently with personal funding and without the financial support of any governmental or private institution or organization.

References

- [1] M. J. Rahman and X. Jie, "Fraud Detection Using Fraud Triangle Theory: Evidence from China," *Journal of Financial Crime*, vol. 31, no. 1, pp. 101-118, 2024, doi: 10.1108/JFC-09-2022-0219.
- [2] J. Debener, V. Heinke, and J. Kriebel, "Detecting insurance fraud using supervised and unsupervised machine learning," *Journal of Risk and Insurance*, vol. 90, no. 3, pp. 743-768, 2023, doi: 10.1111/jori.12427.
- [3] D. K. G. Schrijver and S. M. El-Hajj, "Automobile insurance fraud detection using data mining: A systematic literature review," *Intelligent Systems with Applications*, vol. 21, p. 200340, 2024, doi: 10.1016/j.iswa.2024.200340.
- [4] M. Yankol-Schalck, "Auto insurance fraud detection: Machine learning and deep learning applications," *Journal of Risk and Insurance*, vol. 93, no. 2, pp. 534-568, 2026, doi: 10.1111/jori.70046.
- [5] D. Banulescu-Radu and M. Yankol-Schalck, "Practical guideline to efficiently detect insurance fraud in the era of machine learning: A household insurance case," *Journal of Risk and Insurance*, vol. 91, no. 4, pp. 867-913, 2024, doi: 10.1111/jori.12452.
- [6] R. Ming, O. Mohamed, N. Innab, and M. Hanafy, "Bagging vs. Boosting in Ensemble Machine Learning? An Integrated Application to Fraud Risk Analysis in the Insurance Sector," *Applied Artificial Intelligence*, 2024, doi: 10.1080/08839514.2024.2355024.
- [7] N. Ding, X. Ruan, H. Wang, and Y. Liu, "Automobile insurance fraud detection based on PSO-XGBoost model and interpretable machine learning method," *Insurance: Mathematics and Economics*, vol. 120, pp. 51-60, 2025, doi: 10.1016/j.insmatheco.2024.11.006.

- [8] M. Sathya and B. Balakumar, "Explainable and optimized machine learning for vehicle insurance fraud detection using ABC–XGBoost–SHAP integration," *TPM – Testing, Psychometrics, Methodology in Applied Psychology*, vol. 32, no. S8, pp. 1205-1213, 2025. [Online]. Available: <https://tpmap.org/submission/index.php/tpm/article/view/2838>.
- [9] S. Selvam and M. Sughasiny, "Smart and Explainable Credit Card Fraud Detection Using XGBoost and SHAP," *Journal of ISMAC*, vol. 7, no. 2, pp. 155-169, 2025, doi: 10.36548/jismac.2025.2.004.
- [10] P. Thanathamathsee, S. Sawangarereerak, S. Chantamunee, and D. N. Mohd Nizam, "SHAP-Instance Weighted and Anchor Explainable AI: Enhancing XGBoost for Financial Fraud Detection," *Emerging Science Journal*, vol. 8, no. 6, pp. 2404-2430, 2024, doi: 10.28991/ESJ-2024-08-06-016.
- [11] A. Wilson, K. Xu, Z. Zhang, and Y. Qiao, "The Interpretable Artificial Neural Network in Vehicle Insurance Claim Fraud Detection Based on Shapley Additive Explanations," *Journal of Information, Technology and Policy*, vol. 2, no. 1, 2024, doi: 10.62836/jitp.2024.367.
- [12] F. Almalki and M. Masud, "Financial fraud detection using explainable AI and stacking ensemble methods," 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2505.10050>
- [13] S. Iqbal, K. M. Awan, S. Kamal, and Z. U. Rehman, "Interpretable ensemble learning models for credit card fraud detection," *Applied Sciences*, vol. 15, no. 22, p. 12073, 2025, doi: 10.3390/app152212073.
- [14] Y. Lu, "Statistical machine learning for economic fraud prevention: An interpretable ensemble approach to credit card transaction analysis with SHAP-based explanations," presented at the Proceedings of the 2025 International Conference on Computer Science and Management Technology, 2026. [Online]. Available: <https://doi.org/10.1145/3795154.3795221>.
- [15] P. T. Mazumder, "Explainable and fair anti-money laundering models using a reproducible SHAP framework for financial institutions," *Discover Artificial Intelligence*, 2026, doi: 10.1007/s44163-026-00944-7.
- [16] M. N. Uddin and M. M. Aziz, "Shapley value-guided adaptive ensemble learning for explainable financial fraud detection with U.S. regulatory compliance validation," in "arXiv," 2026. [Online]. Available: <https://arxiv.org/abs/2604.14231>
- [17] U. Zafar and F. Wu, "Methodological challenges in explainable AI for fraud detection: A systematic literature review," *Artificial Intelligence Review*, vol. 59, p. 115, 2026, doi: 10.1007/s10462-026-11516-7.
- [18] M. Aruna, H. Sairaj, V. Sahil, and T. Harshini, "An explainable graph-augmented XGBoost framework for health insurance fraud detection," *International Journal for Research in Applied Science and Engineering Technology*, 2026, doi: 10.22214/ijraset.2026.79131.
- [19] M. J. D. Ebinezer and B. C. Krishna, "Explainable artificial intelligence models using SHAP enhanced CatBoost, Bi-GRU with attention, and TabTransformer," *Scientific Reports*, vol. 16, p. 20357, 2026, doi: 10.1038/s41598-026-49961-7.
- [20] A. A. Khalil, "Enhancing insurance fraud detection accuracy with integrated machine learning and statistical methods," *Computational Economics*, 2026, doi: 10.1007/s10614-025-11074-0.
- [21] A. R. Aitha, "Explainable Agentic AI Framework for Automated Insurance Fraud Detection and Predictive Risk Intelligence," *International Journal of Advanced Research in Computer Science & Technology*, 2026, doi: 10.15662/IJARCST.2026.0903009.