TPM Vol. 32, No. S8, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

# EXPLAINABLE AND OPTIMIZED MACHINE LEARNING FOR VEHICLE INSURANCE FRAUD DETECTION USING ABC – XGBOOST- SHAP INTEGRATION

M. SATHYA*
*DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, FRANCIS XAVIER ENGINEERING COLLEGE, TIRUNLEVELI, TAMIL NDU, INDIA.

DR. B. BALAKUMAR
CENTRE FOR INFORMATION TECHNOLOGY AND ENGINEERING,
MANONMANIAM SUNDARANAR UNIVERSITY, TIRULNELVELI, TAMIL NADU, INDIA.

*Corresponding Author Mail id: sathyamfx@gmail.com , orcid id : 0009-0006-4935-5330

**Abstract:**
Effective detection of insurance fraud is essential for the financial sector, requiring advanced methodologies that ensure accuracy and transparency. This study introduces a novel framework that integrates an XGBoost classifier with the Artificial Bee Colony (ABC) algorithm and the SHAP (SHapley Additive exPlanations) approach for enhanced insurance fraud detection.The ABC-XGBoost with SHAP framework optimizes feature selection and hyperparameter tuning, leveraging the strengths of both the ABC algorithm and XGBoost. Experimental evaluations on real-world insurance datasets indicate that this approach outperforms traditional fraud detection methods, achieving higher accuracy and reduced false positive rates.This research advances the field of insurance fraud detection by demonstrating the effectiveness of merging machine learning and optimization algorithms, while providing transparency through the insights offered by SHAP values.
**Keyword:** Artificial Bee Colony (ABC) algorithm, XGBoost classifier, optimizes feature selection, hyperparameter tuning, SHAP.

## 1. INTRODUCTION:

A key component of the property insurance industry, auto insurance is also essential to the legal systems of many nations. According to a 2018 report by the Chairman of the French Insurance Federation, insurance companies spent an incredible 40 billion euros on damages compensation alone. This spending represented a steady rise in claim costs, which has been marked by notable increases in material and personal injury costs, with an average yearly growth rate of 4% since 2010. Consequently, the average price of auto insurance increased by 2.9% in 2019. As demonstrated by official data from China, the first insurance compensation spending increased by 21.94% year over year to 1.89 trillion RMB in 2023. In particular, industry exchange statistics showed that underwriting earnings for the auto-mobile insurance sector were 9.035 billion RMB, down 12.828 billion RMB or 58.67% from the year before. The International Association of Insurance Supervisors (IAIS) estimates that between 20% and 30% of insurance claims worldwide are suspected of being fraudulent each year, which could be a factor in the decline in underwriting profits for the insurance industry.
According to Van Vlasselaer and colleagues, the term "fraud" has several different meanings. According to Van Vlasselaer et al. (2017), fraud is "a rare, premeditated, hard-to-detect, developing over time, and often well-organized crime that presents in numerous forms," making it necessary to combat it with sophisticated analytical tools and approaches. Because of its negative effects on the normal development of economic order and the difficulties in identifying and preventing it, insurance fraud stands out among the other forms of fraud. According to a 2019 Insurance Asia report, both recognized and undiscovered insurance fraud accounted for a substantial 10% of total costs in Asia. The fact that there were 107,000 false claims in 2018, a 5% increase from the previous year, serves as further evidence of this. Medical insurance fraud, life insurance fraud, and vehicle insurance fraud are subcategories of insurance fraud. According to data from Chinese officials, the motor vehicle insurance market is particularly vulnerable to fraud, accounting for an astounding 80% of all reported occurrences of insurance fraud. Additionally, according to judgments recorded on the Judicial Judgment Network, auto-mobile insurance fraud accounts for 73% of all insurance fraud cases, some of which are still being investigated. As a result, insurers frequently increase premiums for honest clients in order to preserve profitability and offset losses. Verifying every claim within a certain time span is impracticable for investigative agencies, and manual detection techniques are time-consuming and prone to mistakes. In order to find dubious claims among a large number of claim statements, it is imperative to use data mining techniques and artificial intelligence approaches.

Because ensemble learning performs better than individual models and can handle skewed datasets better, it is a good fit for fraud detection, where it can be difficult to distinguish between fraudulent and genuine activity. Tree topologies with second-order derivatives and regularization are used by the XGBoost model, an improved gradient boosting framework, to greatly increase algorithm efficiency. XGBoost's performance is mostly dependent on parameter optimization, and manual tuning can be difficult and time-consuming. In order to improve prediction accuracy and the ability of insurance firms and investigative agencies to identify fraudulent claims, Artificial Bee Colony Optimization (ABC) is a noteworthy strategic approach for optimizing XGBoost parameters.
This study presents a novel method for precisely identifying and interpreting patterns of fraudulent activity in auto insurance claims, with the goal of enhancing the generalization performance and accuracy of insurance fraud detection models. To show how well the PSO-enhanced XGBoost detects auto insurance fraud, we create an advanced model based on XGBoost that is adaptively optimized using ABC. We compare this model to regular XGBoost and other machine learning models. This hybrid strategy could significantly speed up vehicle insurance claim validation, cutting down on the amount of time needed for legitimacy evaluations. Furthermore, in the context of auto insurance fraud detection, we use SHAP (SHapley Additive exPlanations) to make complex machine learning models easier to interpret. This helps policymakers comprehend the predictions and evaluate the importance of different features influencing fraudulent claims. The study graphically offers transparency and understanding into the aspects influencing the model's fraud detection capabilities by elucidating key feature indicators.

## 2. LITERATURE REVIEW

Researchers are paying more and more attention to detecting auto insurance fraud because of the significant effects it has on socioeconomic well-being and financial stability. In general, this field of study consists of two main research phases: the first phase analyzes insurance claim samples using statistical and machine learning techniques, and the second phase focuses on using different heuristic optimization and ensemble learning algorithms for more complex fraud detection research.
This research attempts to create a very efficient framework for identifying vehicle insurance fraud and providing insights into its forecasts based on the aforementioned investigations and conclusions. First, the class distribution in the car insurance claims datasets is balanced. Next, ABC is used to optimize the appropriate XGBoost settings. Lastly, the model's predictions are explained using the SHAP approach.
In their 2025 study, Ding et al. present an advanced methodology for automobile insurance fraud detection that synergizes Particle Swarm Optimization (PSO) with the XGBoost machine learning algorithm. The proposed framework, referred to as the PSO-XGBoost model, aims to enhance detection performance by automating the hyperparameter tuning process using PSO, thereby overcoming limitations in manual or grid-based optimization strategies. Additionally, the authors integrate SHAP (SHapley Additive exPlanations) to interpret model predictions, offering transparency and insight into the most influential features contributing to fraud detection decisions.
The study is evaluated using real-world automobile insurance datasets, and experimental results show that the PSO-XGBoost model achieves higher accuracy, reduced false positives, and improved generalization compared to traditional machine learning approaches. By blending metaheuristic optimization with ensemble learning and explainable AI techniques, the research contributes a robust and interpretable solution to the ongoing challenges in insurance fraud analytics.
AutoFraudNet introduces a multimodal deep learning framework that processes tabular metadata, textual descriptions, and vehicle images to detect fraudulent auto insurance claims. The model employs a cascaded slow-fusion architecture with BLOCK-Tucker fusion blocks, enabling effective feature-level integration across modalities. It also uses lightweight design choices and auxiliary loss functions to mitigate overfitting. Evaluations on a real-world dataset (≈ 1 million claims, 3% fraudulent) demonstrate that the multimodal model outperforms both unimodal and bimodal variants, achieving over 3% higher Precision-Recall AUC compared to baseline fusion strategies.
In “Enhanced Gradient Boosting for Zero-Inflated Insurance Claims and Comparative Analysis of CatBoost, XGBoost, and LightGBM,” So (2023) addresses a key challenge in auto insurance claim modeling: the high number of claim-free policies (zero-inflation) which conventional models struggle to handle. The study evaluates three major gradient boosting libraries—XGBoost, LightGBM, and CatBoost—using both real-world and synthetic telematics datasets with pronounced zero-inflated characteristics .Findings reveal CatBoost achieves the best predictive performance, largely due to its efficient handling of high-cardinality categorical features and symmetric “oblivious” tree structure So also proposes a new zero-inflated Poisson boosted tree model, exploring variants where the inflation probability 'p' is modeled as a function of the mean 'μ' or treated independently. The analysis shows this model outperforms traditional Poisson and zero-inflated GLMs across datasets, and leverages CatBoost-specific tools to uncover feature interactions and risk drivers in telematics-inclusive insurance modelling.

Zheng et al. (2023) present a comprehensive investigation into the prediction of automobile insurance claim fraud using tree-based ensemble methods. The study meticulously preprocesses the dataset, handling missing values, encoding categorical variables, removing outliers, and deriving additional features such as temporal attributes to enrich model input. Following feature engineering, both Gradient Boosting Decision Tree (GBDT) and XGBoost models are trained and evaluated, with performance assessed via AUC metrics. Despite XGBoost's advanced algorithmic structure, the results indicate that GBDT slightly outperforms XGBoost on the provided dataset. The authors conclude that GBDT and XGBoost both offer robust predictive power in auto insurance fraud detection, though GBDT demonstrates marginal superiority. Their findings emphasize the importance of careful data preprocessing and feature creation as essential steps in boosting model efficacy within fraud detection domains.

Chen et al. (2023) conduct a detailed comparison of Gradient Boosting Decision Trees (GBDT) and XGBoost for detecting insurance fraud, with a primary focus on their efficiency, precision, recall, F1-score, and AUC performance. The study begins with systematic data preprocessing, addressing missing values, encoding categorical variables, and mitigating class imbalance to ensure fair evaluation. Both models are meticulously fine-tuned and evaluated on real-world insurance datasets.Findings reveal that XGBoost slightly outperforms GBDT, demonstrating superior AUC and F1 metrics, though both models show competitive results. Notably, the research emphasizes interpretability and computational efficiency, highlighting XGBoost's scalability and faster training times. The authors conclude that GBDT and XGBoost are well-suited for fraud detection tasks, with XGBoost offering a more robust trade-off between predictive performance and operational practicality in real-world scenarios.

Toluwalope et al. (2023) deliver an empirical evaluation of various machine learning algorithms applied to automobile insurance fraud detection, based on a dataset of real-world claims. They test six classifiers—XGBoost, Logistic Regression, Random Forest, Decision Tree, SVM, and Naïve Bayes—using key metrics including accuracy, precision, recall, and F1-score. Random Forest achieves the highest overall accuracy, while XGBoost records the best precision and F1-score, and Decision Tree excels on recall. Feature selection is driven by ANOVA and Random Forest importance, with the Random Forest method preferred for its clear advantages. Their findings offer practical guidance to insurance stakeholders by identifying model-specific strengths in predictive performance.

The major contributions of the proposed framework as follows:

- Artificial Bee Colony (ABC) will be used for Hyperparameter optimization of the XGBoost model and Feature selection to identify the most relevant features that contribute to fraud detection.
- XGBoost will be used for the classification task (fraud vs. non-fraud) and trained on the data to detect patterns indicative of fraud.
- SHAP will provide model interpretability, showing the contribution of individual features to fraud prediction.

## 3. PROPOSED METHODS

### 3.1 Artificial Bee Colony

A new swarm intelligence optimization algorithm, the Artificial Bee Colony algorithm (ABC) mimics the intelligent behavior of honey bees searching for food around their hives by utilizing the cooperation mechanism of bee colonies and combining individual bees' local searches to realize the search for the optimal solution. The ABC algorithm is widely used.

The collective intelligence of bee colonies consists of three essential components: food sources, employed bees and unemployed bees. Unemployed bees are further segregated into onlookers and scouts. The employed bees fly out of the hive to exploit the food sources and feedback the nectar source information (solution) to the bees in the hive. They also gather the relevant information about these food sources such as location, quality and quantity of nectar. According to the food source information shared by the employed bees, the onlooker bees determine the location of the best food sources in a specific way and become employed. Scout bees look for new food sources in the vicinity of the hive. Once a scout bee finds a food source, it becomes employed. When a food source is exhausted, each of its associated employed bees becomes either a scout or an onlooker. The three kinds of bees complete the tasks in accordance with their respective division in the process of seeking nectar. They search the best nectar source through the collection and sharing of nectar source information. The source position in the algorithm corresponds to the solutions of optimization problems. The quantity of nectar corresponds to the fitness function value of the solution. In addition, the number of the nectar source is equal to the number of employed bees and scout bees. The ABC algorithm is an iterative algorithm. It randomly creates dispersed initial solutions (food source positions), analyzes their fitness and distributes the employed bees to the food sources. Each food supply is paired with a single working bee. Following initialization, the algorithm repeatedly iterates through three phases—the employed bee phase, the spectator bee phase, and the scout bee phase—in an attempt to identify the best solution.

One way to characterize the ABC algorithm is as an optimization problem. Regarding the optimization issue P:

$\min\{f(x) : x \in S \subset R^{d}\}$

where f is the objective function, $x = (x_1, x_2, \ldots, x_d)$ is the input variable, and $S = \{(x_j^{min}, x_j^{max}\ j)|\ j = 1, 2, \ldots, d\}$ is the solution space.
As indicated above, the food sources of the bee colony can be seen as the set of plausible solutions to the issue P. The position of a food source relates to a probable solution of the optimal problem. The quality of the food source is determined by the computed value of the objective function f. There are exactly as many food sources as there are bees in use. The actual vector can then be used to indicate the position of food sources.

$x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$

First, the initial population randomly generates SN solution real vectors by Eq, SN is a dimensional vector, and $x_i$, $i = 1, 2, \ldots, SN$ is the solution.

$x_{ij} = x_j^{min} + r(x_j^{max} - x_j^{min})$

where r is a random real number in the interval [−1, 1] with uniform distribution. $x_j^{max}$ and $x_j^{min}$ are respectively the upper and lower bounds of the feasible solutions in the $j^{th}$ dimension.
When the specified accuracy reaches a predefined value or the maximum iteration number reaches the set value of Genmax, the algorithm stops operating. The employed bee first searches the corresponding food source and chooses a neighborhood bee at random. Next, it performs the neighbourhood exploration in the solution space by Equations and creates a new food source. The algorithm then randomly assigns the food source to the employed bee after calculating the fitness value of the objective function and searching the food source for circulation for gen times.

$v_{ij} = x_{ij} + u_{ij}(x_{ij} - x_{kj})$

where $u_{ij}$ is a random real number with uniform distribution in the interval [−1, 1]. j and k are the randomly generated dimensional values and they are not equal. If the result is beyond the range of the space, the solution value is replaced with a similar boundary value by Eq.

$$v_{ij} = \begin{cases} X_{min} & Vij \leq Xmin \\ X_{max} & Vij \geq Xmax \end{cases}$$

An onlooker bee evaluates the fitness information collected from all employed bees and selects a food source with a probability related to its fitness; the higher the fitness of the food source, the more likely the onlooker bees are to select it. Once all employed bees have finished the process of determining a new food source and decided whether or not to move to a newly determined food source, they share the information of their food sources with the onlooker bees. If the new solution x has a higher fitness than the old one, the old solution is replaced by the greedy selection method; if not, the old solution is kept.

$$P_i = \frac{fitness_i}{\sum_{i=1}^{NP} fitness_i}$$

If a food source xi is not updated after limit iterations, $x_i$ is likely to fall into the local minimum and the global optimal solution is not found. Then, the food source will be replaced and a new one will be randomly generated.

**3.2 XGBoost (Extreme Gradient Boosting)**

XGBoost, short for "Extreme Gradient Boosting", is a composite algorithm that combines base learners with weights to achieve excellent data fitting performance. Since its introduction in 2015, XGBoost has gained substantial popularity across data mining, machine learning, and other domains due to its high scalability, robust generalization capabilities, and rapid computation speed (Zhou et al., 2022).

$$\hat{y_i} = \sum_{k=1}^{k} f_k(x_i), f_k \in F$$

Among them, $F = \{f(x) = w_{q(x)}\} \ (q: R^m \to \{1,2, \ldots T\}, w \in R^T)$
represents a group of CART decision tree structures, where w is the real-valued score assigned to the leaf nodes, T is the number of leaf nodes, and q is the tree structure that maps samples to leaf nodes. Finding the ideal parameters is crucial when establishing an XGBoost model in order to minimize the objective function and accomplish the purpose of creating the best model possible. The XGBoost model's objective function may be separated into two parts: the error term L and the model complexity term Ω. The objective function can be expressed as:

$Obj = L + \omega$

$L = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$\omega = \gamma^T + \frac{1}{2}\lambda \sum_{j=1}^{T} W_j^2$$

In order to reduce the objective function as much as possible, a new function f must be introduced into the model while maintaining the previous model during the optimization and training process using the training dataset. The specific process is as follows:

$Obj = \sum_{i=1}^{n} y_i - (\hat{y}_i^{(t-1)} + f_i(x_i)))^2 + \omega$

It is clear that the goal function is only connected to the first and second derivatives of the error function when the constant term is eliminated.

The optimal $w_j$ and the objective function's optimal value can be found using the objective function if the tree structure q is known. In essence, this procedure consists of finding the least value of a quadratic function, following answers can be derived:

$$w_j^* = \frac{-\sum_{i\in Ij} g_i}{\sum_{i\in Ij} h_i + \lambda}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in Ij} g_i)2}{\sum_{i\in Ij} h_i+\lambda} + \gamma^T$$

A smaller Obj value denotes greater model performance, and the Obj can be used as a scoring function to assess the model. The previously mentioned tree construction method can be applied iteratively to produce a huge number of regression tree structures. The best tree structure is then found by searching these structures with the Obj function. In order to create an ideal XGBoost model, the best structures are integrated into the current model.

**3.3 Proposed Framework (ABC + XGBoost + SHAP)**

**ABC** $\xrightarrow{\text{Optimize}}$ **XGBoost** $\xrightarrow{\text{Train}}$ **M** $\xrightarrow{\text{Explain}}$ **SHAP** $\rightarrow$

The above framework ensures high accuracy via ABC-XGBoost , Risk mitigation through SHAP transparency and scalability and adaptability to real-world insurance data.

**Step 1: Problem Setup**

Let the dataset be:

$$D = \{ ( x^{(i)}, y^{(i)} ) \}_{i=1}^{n}$$

Where ,

- $x^{(i)} \epsilon \mathbb{R}^d$is the feature vector for the $i^{th}$claim ( e.g., claim_amount , vehicle_type, accident_location,etc)
- $y^{(i)} \epsilon \{0,1\}$ is the class label:0 = legitimate, 1 = fraud

**Step 2: Hyperparameter Optimization via ABC**

Define the hyperparameter vector for XGBoost:

θ = [ learning _rate,max_depth,n_estimators,subsample,colsample_bytree]

**Employed Bee Phase**

Each bee explores a new solutions $v_{ij}$ from current $x_{ij} + \emptyset_{ij} (x_{ij} - x_{kj})$

Where

- $\emptyset_{ij} \epsilon [-1,1]$is a random number
- $x_{kj}$ is a randomly selected solution $k \neq i$

**Fitness Evaluation (Model Performance)**

Use XGBoost and k-fold cross-validation to compute fitness:

$f(x_i) = -F1score(x_i)$

We minimize the negative f1-score because fraud datasets are usually imbalanced

**Onlooker Bee Selection**

Select candidate solutions based on fitness:

$$P_i = \frac{f(x_{i)}}{\sum_{j=1}^{N} f(x_j)}$$

**Scout Bee Replacement**

If no improvement for a solution after a threshold:

$$x_i = x_{new} \backsim \mathcal{U} \text{ (parameter space)}$$

**Step 3: Train final XGBoost Model**

Let the best hyperparameters be:

$$\theta^* = \arg\min_{\theta} f(\theta)$$

Train the final classifier:

$$M = XGBoost(x_{train}, y_{train}; \theta^* )$$

XGBoost minimizes the regularized loss:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} \imath(y_i, \hat{y}_i) + \sum_{k=1}^{K} \mho(f_k)$$

Where :

- $\imath$ is a loss function (e.g., logistic loss)
- $\omega (f) = \gamma^T + \frac{1}{2}\lambda\|w\|^2$

**Step 4 : SHAP Explanation**

After training apply SHAP to explain predictions:

**SHAP value decomposition**

For an instance x , prediction is decomposed as :

$$f(x) = \phi_0 + \sum_{i=1}^{d} \phi_i$$

Where:

- $\phi_0$: Expected model output (baseline)
- $\phi_i$: Contribution of feature i

Shapley value formula:

$$\phi_i = \sum_{s \subseteq N/\{i\}} \frac{|s|!(|N|-|s|-1!}{|N|!}\{f(s \cup \{i\}) - f(s)\}$$

This quantifier how each features influences the fraud prediction

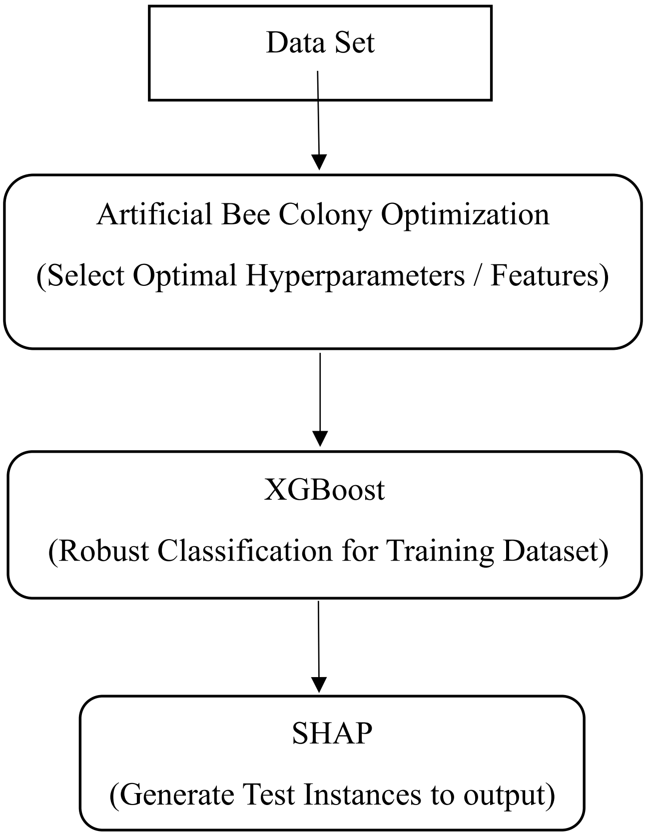


**Figure 1: Proposed Methods Functional Diagram**

**3.4 Model Evaluation:**

The confusion matrix is a crucial tool for evaluating binary classification models, providing insight into the model's performance through four key metrics: true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as shown in Table 1. A true positive (TP) occurs when the model correctly predicts a positive instance. A false positive (FP) arises when a negative instance is incorrectly predicted as positive. A true negative (TN) is when a negative instance is accurately identified as negative. Conversely, a false negative (FN) refers to a positive instance that the model incorrectly classifies as negative.

**Table 1 Confusion matrix**

| Predicted Class | True Class | |
|---|---|---|
| | 1 | 0 |
| 1 | TP | FP |
| 0 | FN | TN |

Based on the aforementioned confusion matrix, the following commonly used evaluation metrics can be defined: accuracy, recall, precision, type II error rate, and F1 score (An, 2018). Accuracy refers to the proportion of correctly predicted samples out of the total number of samples. The specific calculation formula is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

The recall, also known as sensitivity, measures the proportion of actual positive samples that are correctly predicted as positive by the model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision, or positive predictive value, indicates the proportion of samples predicted as positive that are indeed positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The Type II error rate, also referred to as the false negative rate, represents the proportion of actual positive samples that are incorrectly predicted as negative by the model. This is calculated out of all the actual positive samples in the training or prediction set.

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

The F1 score, which is the harmonic mean of precision and recall, ranges between 0 and 1, with higher values indicating better model performance.

$$\text{F1Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

## 4. DATA & RESULTS ANALYSIS

The data used in this paper comprises three parts: investigative agency data on automobile insurance fraud, data from insurance institutions, and international data sources from the Kaggle website (www .kaggle.com). In data cleaning, bad data will be removed, and missing data will be added. To maintain the diversity and variability of the dataset, this paper adopts the method of filling missing values with the mean of the sequence by using IBM's SPSS Statistics 25.0 software. Finally, a total of 11,565 records are obtained. Each record in the dataset has well-defined attributes and is categorized as either fraudulent claims labelled as "1" or non-fraudulent claims labelled as "0". The dataset comprises a total of 11,565 records, of which only 685 are fraudulent claims, while 10,880 are normal accident claims, indicating a high degree of imbalance. the dataset used in this research presents a significant imbalance, with the ratio of fraudulent claims to legitimate claims being approximately 1:16. This indicates that fraudulent claims are a minority class compared to legitimate claims.

The selection of indicator features for each record is informed by insurance industry recommendations, expert insights, literature summaries, and anti-fraud regulations. These factors are integrated to form the indicator features. Each data record includes the following indicator features, as shown in Table 2.

Value assignment for non-numeric variables is essential in feature indicators, involving categorical variables such as "fraudulent status", "gender", "accident area" and "marital status". In the process of model construction, it is necessary to convert categorical variables into numerical variables that can be recognized by computers. For instance, the gender variable, which includes attributes like male and female, can be encoded as 0 and 1, respectively. Likewise, for the marital status variable, which includes attributes like single, married, divorced, and widowed, they can be assigned values of 0, 1, 2, and 3, respectively. This encoding extends to other similar variables.

The goal of feature transformation is to scale the variables to a specific ratio to eliminate disparities in scale among different feature indicators. The dataset contains both numeric and categorical variables with distinct value ranges. Due to significant differences between these variable types, data normalization is necessary. Given that the dataset used in this article comprises tens of thousands of records, with some feature variables being continuous numerical variables with unknown maximum and minimum values, Z-score normalization is adopted. Z- score normalization transforms raw data by adjusting it to have a mean of 0 and a standard deviation of 1. The dataset contains insurance claim records with features like claim amount, location, accident history, vehicle type, etc., and the target variable is fraud (1) / not fraud (0).

**Table 2 Names of data feature variables**

| Number | Names of Feature Variables | Number | Names of Feature Variables |
|---|---|---|---|
| 1 | FraudFound_P | 9 | VehiclePrice |
| 2 | Sex | 10 | Deductible |
| 3 | MaritalStatus | 11 | PastNumberOfClaims |
| 4 | Age | 12 | AgeOfVehicle |
| 5 | AccidentArea | 13 | PoliceReportFiled |
| 6 | Make | 14 | WitnessPresent |
| 7 | Fault | 15 | NumberOfCars |
| 8 | VehicleCategory | 16 | ClaimSize |

**Table: 3 Model Performance Comparison**

| Model | Accuracy | Precision | Recall | F1- Score | AUC-ROC |
|---|---|---|---|---|---|
| XGBoost | 0.89 | 0.85 | 0.83 | 0.84 | 0.91 |
| ABC + XGBoost | 0.92 | 0.88 | 0.87 | 0.87 | 0.94 |
| ABC+ XGBoost + SHAP | 0.92 | 0.89 | 0.88 | 0.88 | 0.94 |

**Table: 4 Feature Importance with ABC Classification**

| Rank | Features | XGBoost Importance | Cumulative % | ABC Class |
|---|---|---|---|---|
| 1 | VehiclePrice | 0.158 | 15.8% | A |
| 2 | Age | 0.120 | 27.8% | A |
| 3 | PolicyType | 0.105 | 38.3% | A |
| 4 | VehicleCategory | 0.078 | 55.3% | A |
| 5 | DriverRating | 0.064 | 61.7% | A |
| . | . | . | . | . |
| . | . | .. | . | . |
| 27 | NumberOfSuppliments | 0.003 | 100% | C |

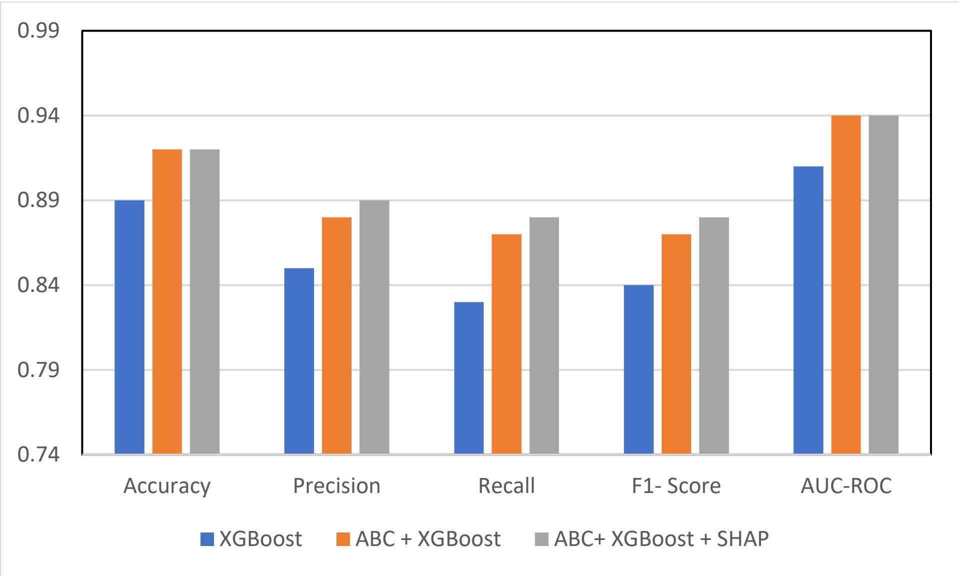

Figure 2: Performance Analysis Chat for Base Methods compared with our proposed method

## 5. CONCLUSION:

The proposed hybrid approach combining Artificial Bee Colony (ABC), XGBoost, and SHAP has demonstrated significant improvements in detecting vehicle insurance fraud. The baseline XGBoost model achieved an accuracy of 0.89, precision of 0.85, recall of 0.83, F1-score of 0.84, and AUC-ROC of 0.91, indicating strong initial performance. Upon integrating ABC optimization, the model's accuracy increased to 0.92, precision to 0.88, recall to 0.87, F1-score to 0.87, and AUC-ROC to 0.94. This improvement shows the effectiveness of ABC in tuning hyperparameters and enhancing classification power.

Further, the addition of SHAP to the ABC + XGBoost model maintained the high performance with the same accuracy (0.92) and AUC-ROC (0.94) while slightly increasing the precision to 0.89, recall to 0.88, and F1-score to 0.88. SHAP contributes valuable interpretability by identifying key features influencing predictions, ensuring transparency in fraud detection decisions. These results validate that combining ABC for optimization, XGBoost for classification, and SHAP for explanation creates a balanced model that is both accurate and interpretable. The model shows strong potential for real-world deployment in insurance fraud detection scenarios. It offers scalability, reliability, and the ability to assist human decision-makers with transparent outputs. The enhanced recall and F1-score reflect better detection of fraudulent claims, minimizing financial loss. In conclusion, the proposed method outperforms traditional models and provides a promising direction for intelligent fraud detection systems.

Future enhancements may include deploying the model for real-time fraud detection, integrating deep learning techniques for sequential data analysis, and applying federated learning to ensure data privacy across multiple insurance companies.

**DECLARATIONS:**

**Ethics approval and consent to participate:** This article does not involve studies with human participants or animals conducted by any of the authors

**Consent for publication:** All authors have provided their consent for the publication of this paper.

**Availability of data and materials:** Not Applicable.

**Competing interests:** Not Applicable.

**Funding:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Authors' contributions :** Both authors contributed equally to the conception, implementation, writing, and revision of the manuscript. Both have read and approved the final version.
**Acknowledgements :** Not applicable
**Conflict of Interest:** The authors declare that they have no conflict of interest.

**REFERENCES:**

1. Ding, N., Ruan, X., Wang, H., & Liu, Y. (2025). Automobile Insurance Fraud Detection Based on PSO-XGBoost Model and Interpretable Machine Learning Method. Insurance: Mathematics and Economics, 120, 51–60. https://doi.org/10.1016/j.insmatheco.2024.11.006
2. Asgarian, A., Saha, R., Jakubovitz, D., & Peyre, J. (2023). AutoFraudNet: A Multimodal Network to Detect Fraud in the Auto Insurance Industry. arXiv.
3. So, B. (2023). Enhanced Gradient Boosting for Zero-Inflated Insurance Claims and Comparative Analysis of CatBoost, XGBoost, and LightGBM. arXiv.
4. Zheng, H., Peng, F., Tian, Y., Zhang, Z., & Zhang, W. (2023). Insurance Fraud Detection Based on XGBoost. Academic Journal of Computing & Information Science, 6(8), 68–74. https://doi.org/10.25236/AJCIS.2023.060808
5. Chen, Y., Feng, Q., Zhang, J., & Zhao, Z. (2023). Algorithmic Implementation for Insurance Fraud Detection [Using GBDT/XGBoost]. Academic Journal of Computing & Information Science, 6(9). https://doi.org/10.25236/AJCIS.2023.060914
6. Toluwalope, T., Shahra, E., & Bassura, S. (2023). Auto-Insurance Fraud Detection Using Machine Learning Classification Models. In 8th International Congress on ICT (Feb 2023), 1–10
7. Yaqi, C., Qianshuo, F., Jiexin, Z., & Zheng, Z. (2023). Insurance Fraud Detection Based on XGBoost. Academic Journal of Computing & Information Science, 6(8), 68–74.
8. https://doi.org/10.25236/AJCIS.2023.060808
9. “Predicting automobile insurance fraud using classical and machine learning models.” (2015). Applied Soft Computing, 37, 368–377. https://doi.org/10.1016/j.asoc.2014.09.019
10. Abdul Jalil, N. A. (2015). Predicting automobile insurance fraud using classical and machine learning models. Applied Soft Computing, 37, 368–377. https://doi.org/10.1016/j.engappai.2014.09.019
11. “Automobile Insurance Fraud Detection Using Data Mining: A Systematic Literature Review.” (2024). Intelligent Systems with Applications, 21, 200340. https://doi.org/10.1016/j.iswa.2024.200340
12. Li, Y., Yan, C., Liu, W., & Li, M. (2020). An artificial bee colony–based kernel ridge regression for automobile insurance fraud identification. Neurocomputing, 393, 115–125. https://doi.org/10.1016/j.neucom.2017.12.072
13. Li, Y., Yan, C., Liu, W., & Li, M. (2018). PCA-based random forest with nearest-neighbor for auto insurance fraud identification. Applied Soft Computing, 70, 1000–1009. https://doi.org/10.1016/j.asoc.2017.07.027
14. Chen, A., et al. (2023). A deep-learning–based antifraud system for car-insurance claims. Expert Systems with Applications, 231, 120644. https://doi.org/10.1016/j.eswa.2023.120644
15. Maina, et al. (2023). Detecting fraud in motor insurance claims using XGBoost with SMOTE. The Journal of Finance and Data Science, 61.
16. Lv, et al. (2023). Auto insurance fraud detection based on Logistic-SVM algorithm. *SPIE Proceedings *. https://doi.org/10.1117/12.2634651
17. Essa, T., Shahra, E., & Bassura, S. (2023). Auto-Insurance Fraud Detection Using Machine Learning Classification Models. Springer ICT proceedings
18. Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance fraud detection: Evidence from AI and ML. Research in International Business and Finance, 62, 101744. https://doi.org/10.1016/j.ribaf.2022.101744
19. Muswere, E. T. (2023). Fraudulent Vehicle Insurance Claims Prediction in the Zimbabwean Insurance Industry. Preprint.
20. Kovacs, M., Hoekstra, R., & Aczel, B. (2021). Using SOM to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance, 65(2), 245. https://doi.org/10.2307/253535
21. Badriyah, T., Chen, T., et al. (2018). Nearest neighbor and statistics method for detecting auto insurance fraud. Journal of Finance and Data Science
22. Pan, S., et al. (2016). Auto insurance fraud detection using unsupervised spectral ranking. Journal of Finance and Data Science
23. Yankol-Schalck, M. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems
24. Yan, C., Li, Y., Liu, W., & Qi, M. (2020). Adaptive genetic-algorithm BP neural network for vehicle insurance fraud identification. Theoretical Computer Science, 817, 12–23. https://doi.org/10.1016/j.tcs.2019.06.025
25. Debener, J. (2019). Improved adaptive genetic algorithm for vehicle insurance fraud identification. Theoretical Computer Science
26. B. Itri, et al. (2021). TGANs with ML models in automobile insurance fraud detection and imbalanced techniques. International Journal of Recent Technology and Engineering.