Computational Economics
https://doi.org/10.1007/s10614-025-11074-0

Enhancing Insurance Fraud Detection Accuracy with Integrated Machine Learning and Statistical Methods

Ahmed Abdelreheem Khalil[1]

Accepted: 24 July 2025

Abstract

The insurance industry plays a critical role in managing risks and providing financial security globally. However, the industry faces challenges, particularly with the increasing complexity of fraudulent activities. To address these challenges, this work seeks to construct suitable decision models by integrating methods such as feature discretization, feature selection, data resampling, and binary classification, in order to create a prediction system for identifying insurance fraud. The research investigates various scenarios, including different combinations of classifiers, feature selection methods, feature discretization techniques, and data resampling strategies, and the performance of the predictive system is evaluated using established metrics. The experimental results revealed that integrating multiple methodologies during data preprocessing significantly enhances the performance of classification models. The model that utilizes the KBD+RFE+Over +RF scenario achieves the highest AUC and F1-score, indicating exceptional performance in detecting insurance fraud. Our research demonstrates that the proposed models' ability to predict insurance fraud has been significantly enhanced by utilizing resampling methods and highlights the importance of these techniques in improving the efficiency of the utilized integrated artificial intelligence techniques. In addition, the article concludes that the insurance industry can greatly benefit from modern predictive methods to make sound decisions.

Keywords Classification · Data mining · Feature discretization · Insurance fraud · Imbalance data · Machine learning · Prediction system

✉ Ahmed Abdelreheem Khalil
Ahmedrohim91@aun.edu.eg

[1] Present address: Statistics, Mathematics, and Insurance Department, Faculty of Commerce, Assuit University, Assuit, Egypt

Published online: 28 August 2025

1 Introduction

The insurance industry is an essential element of the worldwide financial environment, serving as a crucial player in the management of risks and the provision of financial security. The insurance industry operates on the fundamental principle of mitigating the likelihood of financial loss or risk. The insurance sector is composed of several important players, including insurance firms, insureds, brokers, and regulatory authorities (Khalil et al., 2022a).

Insurance companies face significant challenges in a complex environment shaped by dynamic economic, technological, and regulatory factors. One of the most critical challenges is the rise in fraudulent activities, driven by advanced technology and global communication networks, leading to annual financial losses totaling billions of dollars worldwide (Akhtar et al., 2023). These fraudulent activities not only impact the profitability of insurance companies but also affect their pricing strategies and the overall socio-economic benefits they provide (Wang & Xu, 2018).

To address these challenges, insurance companies must implement strong measures for fraud detection and prevention, as insurance fraud represents a significant portion of their operational costs. In addition to fraud-related challenges, insurance firms also encounter operational difficulties due to the growing complexity of internal procedures and systems, which can hinder efficiency improvements and obstruct the integration of data analytics and artificial intelligence (AI) into risk evaluation and claims handling (Hassan & Abraham, 2013; Singh & Chivukula, 2020). Therefore, adopting a proactive and flexible approach is crucial for ensuring resilience, financial stability, and fostering innovation in the face of uncertainty and change (Khalil et al., 2024b). The Oxford English Dictionary (Pearsall, 1999) defines fraud as "the act of intentionally deceiving others to obtain financial or personal benefits,". Policyholder fraud is one of the four distinct categories of insurance fraud which will be the focus of this article due to the limited available data on other types of fraud.

Data mining is extensively employed in the insurance industry for various reasons such as fraud detection, analysis of claims, processing of underwriting, assessment of risks, and sales prediction. since it is frequently utilized to extract and reveal concealed insights from vast amounts of data (Turban, 2011). Data mining involves the discovery of statistically reliable, previously unknown, and actionable insights from data. The data must possess the qualities of accessibility, pertinence, sufficiency, and integrity. Utilizing integrated algorithms in claim analysis assists insurers in enhancing their comprehension of claims filing and identifying instances of fraud (Prasasti et al., 2020).

Ensemble learning approaches are widely recognized as a prominent field of research that is adaptable and applicable to a range of machine learning (ML) applications, such as classification, regression, and even unsupervised learning (Alsuwailem et al., 2023). Their remarkable performance stems from their capacity to boost model generalization, mitigate overfitting, and improve performance in situations where individual models may struggle. Ensemble learning approaches can greatly enhance predicted accuracy, but they also present difficulties in terms of computing complexity and model interpretability (Khalil et al., 2022b; Piovezan et al., 2023). Ensemble learning is a fundamental technique that aims to improve the performance of ML

models. It provides a strong and adaptable method to solving complicated issues in various fields (Das et al., 2021).

Prediction systems, which are typically supported by different Data mining approaches such as resampling the data, feature discretization, and feature selection play a crucial role in identifying risks. By selecting a relevant subset of features, the computational cost is decreased, and the efficiency and comprehensibility of the model are greatly enhanced (Gupta et al., 2022). Additionally, the precision of prediction algorithms can be influenced by dataset imbalance, which refers to an uneven distribution of positive and negative cases. In such cases, the overall performance of the models can be enhanced through data resampling (Baesens et al., 2021; Subudhi & Panigrahi, 2018). Reviewing the literature on insurance fraud reveals a dearth of research that constructs classification models by integrating the aforementioned data mining strategies with AI classifiers such as (Ensemble and Classic ML approaches) into a unified processing procedure to develop a classification model for insurance fraud.

This research aims to develop a resilient predictive system by creating diverse detection scenarios using a combination of approaches such as resampling the data, feature discretization, and feature selection, as well as different classifiers. The focus of this predictive system is the identification of insurance fraud, with validation conducted on a genuine dataset obtained from insurance companies. The research assesses predictive accuracy through the utilization of two distinct datasets: the insurance fraud dataset and the insurance claims dataset. There is a significant gap in the insurance area that has not been filled by current research, according to a critical review of the literature.

Consequently, the principal contributions of this study are delineated as follows: First, this study presents primary advancements characterized by the introduction of distinct scenarios utilizing different Classifiers, specifically employing two different Methods for feature selection, feature discretization, and three diverse data resampling strategies. The overarching objective is the development of a predictive system characterized by precision and robustness in insurance fraud detection. Second, Investigated the impact of applying the discretization followed by feature selection on binary classification results. Third, Assessing the impact of resampling techniques on binary classification results. Fourth, the practical implementation of the proposed prediction systems on two disparate datasets to affirm their validity, with the provided code made openly accessible. Finally, a thorough evaluation and comparison of performance differences among various detection scenarios is carried out, using four well-recognized metrics: Accuracy, sensitivity (Recall), F1 score, precision, and AUC. Statistical analysis is used alongside these evaluation measures to determine the most favourable situation for the proposed datasets.

The remainder of this work is outlined as follows: A brief introduction to prior studies is given in Sect. 2. Section 3 provides a detailed account of the research methodology, including the research design, data collection methods, describing the specific techniques and approaches that are employed to detect insurance fraud in the study, and evaluation metrics. The findings are presented and analyzed in Sect. 4. Finally, we present the conclusion in Sect.5.

2 Literature Review

Throughout time, insurance companies have found strong reasons to adopt a proactive and flexible approach into their operations to achieve both their long and short-term objectives and effectively navigate the complex challenges that face such risk prediction, fraud detection, claims analysis, and pricing strategies to maintain financial stability and succeed in an ever-changing environment (Barry & Charpentier, 2020). Turban (2011) defines data mining as a method of extracting valuable insights from large databases by applying mathematical, statistical, artificial intelligence, and ML techniques.

Insurance fraud remains a pervasive challenge, costing the industry billions annually and necessitating robust regulatory frameworks, cross-market collaboration, and advanced analytical tools. Scholarly and industry research highlights the National Association of Insurance Commissioners (NAIC) as a pivotal body in standardizing anti-fraud measures through model laws, such as its Special Investigative Unit (SIU) Guidelines, which mandate insurer compliance and data-sharing protocols (NAIC, 2022; Saylor, 2023). For instance, Hoyt et al. (2006) analyzed auto insurance fraud data, finding antifraud laws had mixed results. Mandatory SIUs and felony classification reduced fraud, but mandatory reporting to law enforcement increased it, suggesting inefficiency when replacing private efforts. The $85 billion/year fraud problem was also significantly influenced by market-specific factors beyond legislation. Findings highlight the need for targeted anti-fraud measures.

Moreover, Saylor (2023) seminar paper examines opportunistic auto insurance fraud, emphasizing its prevalence, economic impact, and detection challenges. The study highlights that opportunistic fraud accounts for a significant portion of insurance fraud but is often overlooked in favor of high-profile "hard" fraud cases. Using Neutralization Theory, Saylor analyzes how perpetrators justify fraud (e.g., denying responsibility or victimhood) and proposes deterrence strategies, including public awareness campaigns, enhanced claims-handling procedures, and inter-agency cooperation. Key recommendations focus on activating internal ethical controls and improving industry tools like ISO databases. (Aivaz et al., 2024) analyze economic fraud research trends through bibliometrics, identifying the U.S. and China as top contributors. Key findings highlight growing focus on digital detection tools (AI, blockchain) and socioeconomic impacts. The study reveals increasing publications but declining citation impact, suggesting need for more impactful research.

Insurance companies and data mining researchers encounter various obstacles in insurance's data including issues of data availability, data quality, and missing values. In addition, they also struggle with imbalanced datasets, and the interpretability of model choices (Cappiello, 2020). Consequently, numerous research have been conducted in the field of insurance utilizing diverse methodologies. For instance, Bhowmik (2011) presents a strategy for detecting fraud in auto insurance utilizing algorithms based on decision trees (DT) and naïve Bayesian classification and uses procedures such as Rule-Based Classification, decision tree visualization, and Bayesian naïve visualization to analyse predictions. The results show how well these methods work in identifying car insurance fraud. Dhieb et al. (2019, 2020) utilized ML

techniques to autonomously detect and categorize motor insurance fraudulent claims and also include alert mechanisms for identifying suspicious claims.

Further, Kowshalya and Nandhini (2018) used data mining techniques to predict insurance premiums and fraudulent claims, reducing time spent on claims analysis. They generated a synthetic dataset based on automobile insurance fraud research to develop classification algorithms for detecting false claims. Itri et al. (2019) developed a novel method to improve the accuracy of fraud prediction by testing (10) ML algorithms to determine which were most effective and reliable. Using automobile insurance claims data, the study revealed that Random Forest outperformed all other algorithms in predicting fraud. Subudhi and Panigrahi (2020) introduced a GA-based Fuzzy C-Means (FCM) clustering with supervised classifiers to detect fraud in auto insurance claims, proving its efficiency on real data. Nordin et al. (2024) compared traditional and ML models for predicting automobile insurance fraud, finding that the tree-augmented naïve Bayes (TAN) model outperformed others in accuracy and sensitivity. The study emphasizes the effectiveness of ML in detecting fraud and recommends improving data preparation and model settings for better results.

On the other side, researchers work to improve fraud prediction in different fields by addressing data quality issues like imbalanced data and missing values, as well as optimizing machine learning model parameters for better performance. Various methodologies have been suggested and utilized in the literature to address imbalanced and missing values classification challenges when it comes to acquiring high-quality data for insurance fraud detection modelling. For instance, Sundarkumar et al. (2015) employed the Random under-sampler resampling approach together with Probabilistic Neural Network (PNN), DT, SVM, Logistic Regression (LR), and Group approach of Data Handling (GMDH) models. The study's findings demonstrated that the DT model had the highest efficacy in fraud detection. In a similar manner Hassan and Abraham (2016b) employed a Random under-sampler in conjunction with DT, NN, and SVM models. Their findings indicated that the DT model exhibited the highest performance. Wang and Chen (2020) presents a three-way ensemble method for addressing missing data by grouping objects without missing values and filling gaps with average attributes from each group. Though experiments from the UCI ML repository demonstrate its effectiveness, the approach lacks a comprehensive strategy for managing missing values.

Hanafy and Ming (2021) studied nine SMOTE family methods to address imbalanced data in predicting insurance premium defaults. They evaluated these techniques using 13 machine learning classifiers, and the results showed a notable improvement in classifier performance with the application of SMOTE techniques. Jovanovic et al. (2022) improve credit card fraud detection with ML and the group search firefly algorithm. A real-world, unbalanced dataset of European credit card transactions is used to tune SVMs with extreme gradient boosting. The study found that tuned models outperform other leading approaches in accuracy, recall, precision, and area under the curve after synthetic minority over-sampling expanded the dataset. Tayebi and Kafhali (2024) optimize hyperparameters in ML models for credit card fraud detection using metaheuristic algorithms such genetic algorithms, particle swarm optimization, and artificial bee colonies. These methods improve detection accuracy, recall, and computing economy over grid search, especially for imbalanced datasets.

Future research involves using deep learning models and advanced data balancing approaches to boost fraud detection accuracy.

Based on a comprehensive analysis of available research, this research aims to fill a gap in existing literature by doing a thorough analysis a unified framework for processing datasets and developing a classification model for detecting insurance fraud. Given this study gap, it's unclear if integrating recommended methodologies and techniques during dataset processing could improve classification models. The research also prioritizes positive outcomes and develops methods to explain specific predictions to improve model interpretability.

3 Proposed Computational Methodology

In this section, we present the proposed computational methodology for building a robust detecting system for insurance fraud. Our approach aims to present different systems for detecting fraud by integrating different AI classifiers with data mining techniques such as feature discretization, Feature selection, and resampling for addressing challenges such as Feature importance, and imbalance data. By outlining the specific steps and techniques to be employed, we aim to achieve robust and reproducible results that contribute to the advancement of insurance field.

The proposed methodology is structured to ensure systematic and efficient handling of the computational tasks involved. We begin with preprocessing stages, such as data collection or preprocessing. Subsequently, we detail feature discretization, SelectKbest (Kbset) and Recursive feature elimination (RFE) techniques for feature selection, three diverse data resampling strategies, and different Classifiers. Then, we evaluate the performance of proposed systems by using four well-recognized evaluation metrics, namely, Accuracy, sensitivity (Recall), F1 score, precision, and AUC. Each of these phases is deemed essential to the overall effectiveness of the proposed systems.

Furthermore, we emphasize the adaptability of our methodology, which allows for scalability and applicability to a range of datasets or scenarios within the scope of insurance field. This flexibility enables us to effectively address variations in data characteristics and research requirements, thereby enhancing the generalizability of our findings. A visual representation of the procedures involved in the system is depicted in Fig. 1, which serves as a framework for the detection of insurance fraud within the context of this study.

3.1 Data Collection

In this analysis, two different datasets are used to assess the accuracy of proposed systems. The datasets for this study were obtained from Kaggle.com. The target variable in the datasets is different. In the first dataset[1], we implement the proposed system to detect the insurance fraud for Automobile insurance sector, so the target feature is

[1] https://www.kaggle.com/code/jwilda3/classifying-fraud-by-decision-trees.

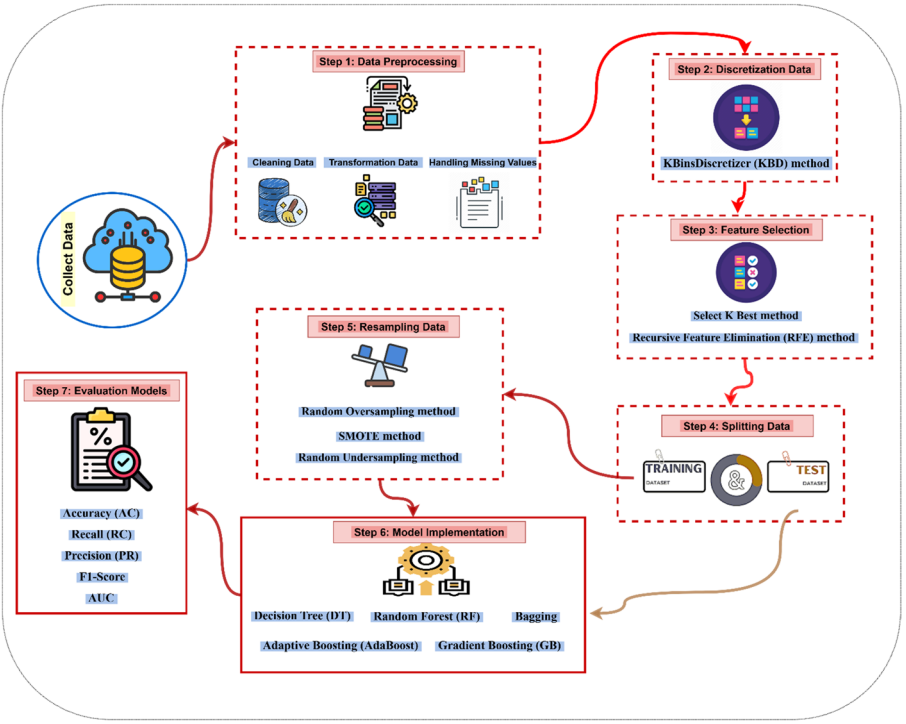

Fig. 1 The block diagram of the proposed system

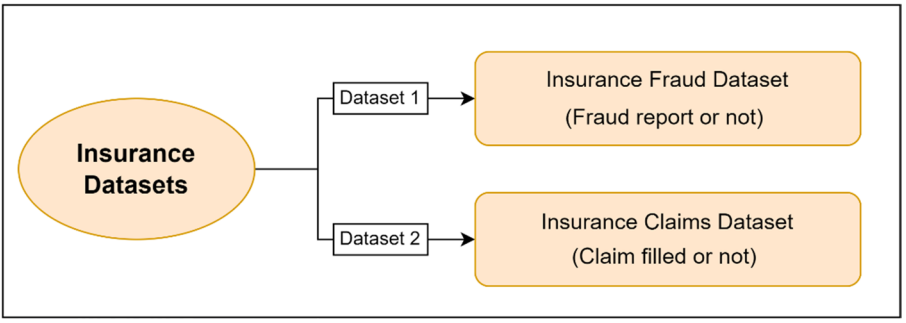

Fig. 2 Category of the target feature in the dataset

"fraud report" column. In the second dataset[2], we implement the proposed system for claim analysis for Automobile insurance sector, so the target feature is "claim flag" column which Indicates whether a claim was filed or not. Figure 2 shows the target variable in both datasets is categorical. Hence, classification systems are used to do the analyses. Figure 3 shows the distribution of the target variables in datasets. In fraud dataset, the ratio between the non-fraud and fraud claims is 94–6%. In the claim

[2] https://www.kaggle.com/datasets/xiaomengsun/car-insurance-claim-data/data.

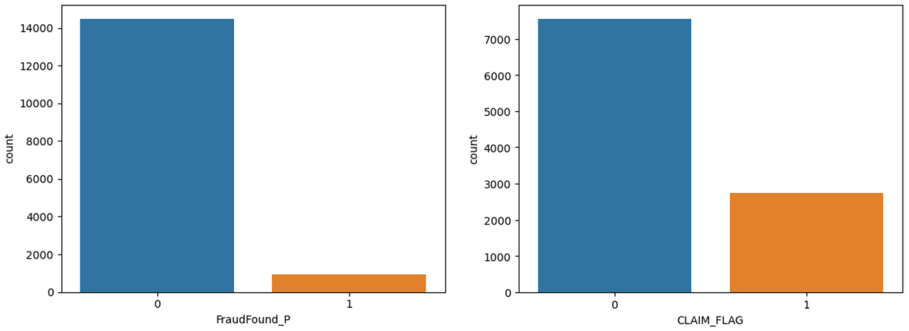

Fig. 3 The distribution of the binary target variable

dataset, the ratio between the non-filled and filled claims is 73.3–26.7%. This refers to the datasets suffering from imbalanced data problems.

The dataset on fraudulent car insurance claims has 15,419 instances, with 923 cases classified as fraudulent, showing a notable class imbalance in the data distribution. Each claim in this dataset is defined by 32 distinct attributes, as listed in Table 1. The insurance claims dataset contains 10,302 car insurance claims, with 2,746 classed as claim filled, highlighting a significant class imbalance. Each claim in this dataset is characterized by 26 unique attributes, as outlined in Table 2.

The two datasets used in this study differ notably in size, label distribution, and the extent of missing data. The first dataset, which focuses on fraudulent claims in the automobile insurance sector is larger with 15,419 entries and exhibits a high imbalance where only 6% of claims are labeled as fraudulent, leaving 94% non-fraudulent. In contrast, the second dataset which analyzes whether a claim was filed contains 10,302 entries and a less severe imbalance with 26.7% of claims marked as filed and 73.3% non-filed. In addition, missing data varies significantly between the two datasets. In the first dataset, the highest percentage of missing values in a single feature reaches 45.7% (Number of Supplements) with a minimum of 0.35% (Days Policy Accident). Conversely, the second dataset has a lower range of missing data with a maximum of 6.45% in the Occupation feature and a minimum of 0.07% in the Age feature. These differences highlight the need for tailored data handling approaches to address the unique class distribution and data completeness issues in each dataset.

3.2 Data Preprocessing

One of the most important steps in the application of classification approaches is data preprocessing, which is also illustrated in the initial phase of Fig. 1. Data must be processed before any future operations because the data may include multiple errors. Thus, this phase involves essential data processing tasks, such as imputing missing values, scoring data, encoding features, discretization data, and dividing the data into training and testing datasets.

Table 1 The fraud dataset description

No.	Features
X1	Unique identifier for each insured (not use)
X2	The month which the accident occurred
X3	The week in the month the accident occurred
X4	The days of the week the accident occurred on
X5	The vehicle manufacturer brand
X6	The area of the accident occurred
X7	The day of the week the claim was filled
X8	The month of the year the claim was filled
X9	The week of the month the claim was filled
X10	The insured's gender
X11	The insured's marital status
X12	The age of the insured
X13	The person responsible for the accident
X14	Type of vehicle insurance policy
X15	The categorization of the vehicle
X16	The price category for vehicles
X17	The deductible amount of the insured.
X18	The driver rating
X19	The days between policy purchase and accident
X20	The days between policy purchase and claim filed
X21	The previous number of claims filed by policy holder
X22	The age of vehicles at time of the accident
X23	The intervals of insured age
X24	If there is a police report or not
X25	If there is a witness or not
X26	The agent who is handling the claim
X27	The number of supplements
X28	If the address of insured change or not
X29	The number of vehicles involved in the accident
X30	The year of accident occurred
X31	The type of insurance coverage
X32	Fraud class (The target feature)

3.2.1 Data Cleaning and Encoding

To improve the efficiency and quality of datasets, data cleaning is used to find and fix errors, corruption, and missing information. This makes analysis and Classification models more effective (Cerda & Varoquaux, 2022; Li et al., 2021). Firstly, features with missing values are either completely removed or their missing values are altered, depending on the data set. In our study, we removed two features (X21, X27) from the first dataset because they have a high percentage of missing values as shown in Table 3. For the remaining features across the two datasets, to fill in missing values in binary and category variables, the mode of the column values is utilized. In contrast, the mean of the column values is used to fill in missing values in all continuous variables.

For the target variable, to find Features with high multicollinearity in the data with it the correlation matrix and Variance Inflation Factor (VIF) analysis were examined.

Table 2 The claims dataset description

No.	Features
X1	Unique identifier for each customer (not use)
X2	The date of birth of the insured (not use)
X3	The age of the insured
X4	The insured's gender
X5	The insured's education level.
X6	The insured's work.
X7	The employment years of the insured
X8	The daily travel time to work.
X9	The insured's marital status
X10	The insured's single parent status (Y/N)
X11	The home value of the insured
X12	The income of the insured
X13	Number of insured's kids
X14	The car usage purpose
X15	The value of the insured's car
X16	The time in force (Years).
X17	The vehicle manufacturer brand
X18	If the customer has a red car or not
X19	The total amount of previous claims
X20	The number of claims report during past 5 years
X21	If the insured's driver license was revoked or not
X22	The number of motor vehicle record points
X23	The Possible future claims
X24	The car age
X25	The urbanicity type
X26	The claim filled (The target variable)

Table 3 The missing values in the datasets

Dataset	Feature name	Feature number	Dataset size	Number of missing values	%
Dataset (1)	Days Policy Accident	X19	15,419	55	0.35%
	Age	X12		316	2.06%
	Past Number of Claims	X21		4351	28.2%
	Number of Supplements	X27		7046	45.7%
Dataset (2)	Age	X3	10,302	7	0.07%
	Years of Job	X7		548	5.32%
	Home Value	X11		575	5.58%
	Income	X12		570	5.53%
	Car Age	X24		639	6.20%
	Occupation	X6		665	6.45%

This strategy reduced multicollinearity. In dataset (2), X23 feature was excluded due to its significant correlation with the target variable. However, this strategy was important to improve the model's predictive accuracy and guarantee that the remaining variables provided clear and interpretable insights into their relationship with target variable.

Data encoding is an important part of preprocessing, which is taking raw data and transforming it into a numerical representation that can be used by algorithms, and statistics models. As an example, we converted category feature to a numerical format, such as assigning the insured's gender to the integers "1" for "male" and "0" for "female".

3.2.2 Discretization Method

Discretization algorithms are fundamental to data preprocessing in machine learning, data mining, and statistical analysis. Converting continuous variables to discrete ones simplifies analysis and algorithm application. Several reasons necessitate discretization. First, discrete inputs are needed for many machine learning methods, therefore continuous data must be transformed. By grouping comparable values, discretization simplifies interpretation. Different discretization methods for continuous variables have pros and cons. Unsupervised and supervised methods are used. Unsupervised approaches like equal width and frequency binning separate data by distribution. However, supervised approaches like decision tree-based discretization and entropy-based binning use class labels or target variables to guide discretization.

In this study, we employed KBinsDiscretizer (KBD) method. KBD method uses binning techniques to convert continuous variables into discrete bins, enabling the use of continuous data in models that need categorical inputs. This method can improve clarity, decrease computational intricacy, and potentially enhance the efficiency of machine learning algorithms, especially those influenced by the characteristics of the input data.

3.3 Feature selection techniques

Feature selection (FS), also known as attribute selection or variable subset selection, is a widely used technique for reducing the dimensionality of feature space while maintaining the performance of a given methodology. Feature selection presents a complex challenge due to the need for complementary features to address interactions and redundancies. The objective of FS is to identify and eliminate irrelevant or redundant features that do not contribute significantly to the learning process. The primary goal of FS is to improve learning performance by enhancing the accuracy and comprehensibility of the methodology being used (A. Singh & Jain, 2019). Therefore, A more efficient global search technique is necessary to tackle feature selection effectively. In this study, we employed two different techniques for Feature Selection.

3.3.1 Select K Best

SelectKBest with ANOVA F-value is a univariate selection method. In this method, features are ranked to eliminate irrelevant ones in this method. The ranking is determined by statistical scores calculated from the association between features and the target variable (Visalakshi & Radha, 2014). It selects the K best features by computing the ANOVA F-value between each feature and the target variable. Features with higher F-values are considered more relevant to the target variable. This method is particularly useful when dealing with many features and is computationally efficient. However, it does not consider interactions between features, and the choice of K needs to be carefully determined to balance model performance and dimensionality reduction. K is the number of top features to be used for feature selection (Srivatsan & Santhanam, 2021).

3.3.2 Recursive Feature Elimination (RFE) with Random Forest Classifier

Recursive Feature Elimination (RFE) is a method of selecting features that involves iteratively removing features from the dataset while repeatedly fitting the model. RFE is utilized with a Random Forest Classifier as the estimator in this scenario. The technique initially trains the model using all features and then assesses the significance of each feature (Lakshmanarao et al., 2022; Visalakshi & Radha, 2014). The algorithm eliminates the least significant feature and iterates this procedure until the desired number of features is achieved. Utilizing Random Forest Classifier with Recursive Feature Elimination (RFE) is efficient for pinpointing the most significant features through the importance scores generated by the random forest model. The classifier used can influence the selected features, and the quantity of features to choose must be carefully adjusted to enhance model performance (Visalakshi & Radha, 2014).

3.4 Resampling Methods

The problem of imbalanced data is pervasive in many datasets, leading to biased classifier models that cannot make accurate predictions for minority classes (Kotsiantis et al., 2006). Consequently, addressing the issue of imbalanced data is imperative. Various methods have been developed to resolve this issue, with one of the most successful approaches involving the use of sampling-based techniques, such as random oversampling and random undersampling (Basit et al., 2022; Zhang et al., 2024). Table 4 displays the fundamental properties of each resampling approach.

3.5 Classifier models

In the proposed work, we employed different classical ML and ensemble learning classifier models, namely, Decision Tree (DT), Random Forest (RF), AdaBoost, Gradient Boosting (GB), and Bagging, Boosting. Recognizing that the performance of any ML model is contingent upon the specific values assigned to its parameters.

Table 4 Resampling method descriptive

Method	Description	Advantage	Disadvantage
Random Oversampling (Xiaolong et al., 2019)	Random oversampling increases the weight of the minority class. Bootstrapping creates artificial instances based on the conditional density estimations of the two groups. Creating duplicates of samples increases the size of the dataset. This method maintains the variety of samples for both categorical and continuous data.	• Helps to balance imbalanced classes effectively. • Does not introduce new samples, preserving sample variability.	• May result in overfitting by duplicating current samples. • Raises computational burden by expanding the dataset.
SMOTE (Amirruddin et al., 2022)	SMOTE generates new minority samples by synthesizing data from two minority samples and their K nearest neighbours. It creates additional instances of the minority class by using existing instances and their closest neighbours to increase the sample size. New instances do not duplicate minority samples.	• Enhances minority class representation without duplicating current samples. • Utilizes attributes of current examples and their neighbouring examples to generate novel examples.	• May introduce synthetic samples that do not accurately represent the true distribution of the minority class. • Requires careful tuning of parameters, such as the number of nearest neighbours (K), which can impact results.
Random undersampling (Liu et al., 2020)	Random undersampling aims to balance imbalanced data by randomly removing examples from the majority class in the training dataset. It reduces the dataset size by discarding examples from the majority class.	• Simple and straightforward method to address class imbalance. • Reduces computational overhead by decreasing dataset size.	• Discarding important information from the majority class can reduce predictive power. • Removing too many samples can worsen class imbalance.

3.5.1 Decision Tree (DT)

DT is a flexible and interpretable classification technique that recursively divides data by selecting the feature that best separates it into homogeneous subsets, maximizing class label purity until a stopping criterion is reached, that forming a model for new predictions (Bansal et al., 2022). Its strength lies in its ability to capture com-

plex, non-linear relationships and handle various data types that make it useful across many analytical scenarios. However, decision trees are prone to overfitting, as they can memorize training data, which leads to poor generalization of new data without regularization. They are also sensitive to slight changes in the training set, which can affect their predictions, so careful parameter tuning, and ensemble approaches are essential for optimal performance (Bansal et al., 2022).

3.5.2 Random Forest (RF)

RF is an ensemble learning technique that trains multiple decision trees and combines their predictions to improve accuracy and generalization in machine learning tasks. For regression tasks, it takes the average prediction of the individual trees, while for classification tasks, it uses the mode of class predictions (Roy & George, 2017). RF is highly accurate and robust, often avoiding overfitting and performing well even with missing data, which makes it reliable across a range of applications. However, RF requires considerable processing power for large datasets and lacks the interpretability of single decision trees, though its high performance across diverse tasks makes it a valuable tool in machine learning (Roy & George, 2017).

3.5.3 Adaptive Boosting (AdaBoost)

AdaBoost builds a strong classifier by combining multiple weak learners and adjusting weights on misclassified instances in each iteration, which focusing more on difficult cases and allowing later learners to correct previous errors resulting in an accurate model (Hassan & Abraham, 2016a). AdaBoost's strength lies in its ability to improve upon weak learners by integrating base classifiers, making it flexible and adaptable across various data types and problem domains. However, AdaBoost's performance can suffer with noisy data, as it may overfit if the base classifiers are overly complex or unstable which can impact its generalization. Thus, its effectiveness depends on data quality and the simplicity of the classifiers used (Ben Jabeur et al., 2023).

3.5.4 Gradient Boosting (GB)

GB is an ensemble learning technique that builds a powerful predictive model by sequentially adding weak learners, and usually decision trees build to reduce errors from previous models. This process optimizes a loss function through gradient descent and creating a highly accurate model (Dhieb et al., 2019). GB models are known for their strong performance, robustness to outliers, and ability to handle both numerical and categorical data, which make them highly versatile across various classification and regression tasks. However, despite these strengths, GB can overfit particularly if regularization is inadequate or the learning rate is too high, and its iterative, and complex model-building process can be computationally intensive on large datasets (Liu et al., 2019).

3.5.5 Bagging

Bagging is an ensemble learning method that improves model performance by reducing variance through training multiple models independently on random subsets of the training data. By capturing data variability across models and combining their predictions through averaging (for regression) or voting (for classification), Bagging enhances both accuracy and generalization and effectively minimizing overfitting and increasing model stability (Park & Kwon, 2024). Its strengths lie in its ability to leverage diverse model predictions, which making it especially useful for complex models and large datasets by optimizing performance and scalability. However, Bagging may not improve results for stable models with low variance and could introduce bias if the base models or the dataset itself is biased. Thus, assessing model stability and dataset characteristics is essential to achieve optimal results with Bagging.

3.6 Classification accuracy evaluation metrics

A key component of finding and comparing the best model is evaluation metrics, which assess the efficiency of classifiers. A popular indicator that shows the percentage of correct predictions is accuracy. A greater accuracy value shows that the classifier is performing better overall. While accuracy is important, it might not be enough to solve classification difficulties, particularly when working with data that is imbalanced (Hossin & Sulaiman, 2015; Khalil et al., 2024a). In response to this challenge, Various classification evaluation criteria are used to assess the classifier's performance.

$$Accuarcy \ \ (AC) = \frac{TP + TN}{TP + FP + TN + FN}, \tag{1}$$

$$Recall \ \ (RC) = \frac{TP}{TP + FN}, \tag{2}$$

$$Precision \ \ (PR) = \frac{TP}{TP + FP}, \tag{3}$$

$$F1 - Score \ F = \frac{2 \times \ TP}{2 \times \ TP + FP + FN}, \tag{4}$$

where TP represents true positives, TN indicates true negatives, FP is the false positives, and FN is the false negatives.

4 Predictive Analysis and Model Interpretability

In this section, we present the experiments, and results conducted to address the research questions outlined in the preceding sections which if integrating recommended methodologies and techniques during dataset processing could improve clas-

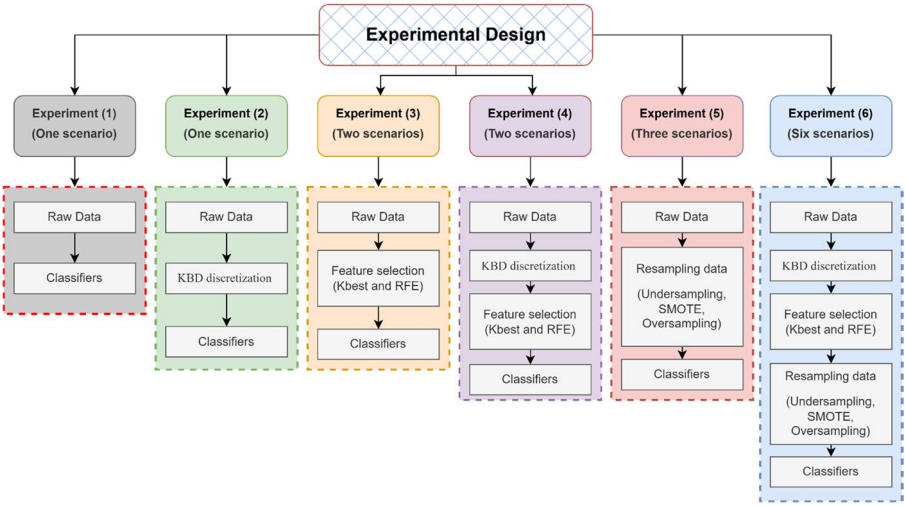

Fig. 4 The experimental design for the study

sification models. Our study aims to build a robust classification model to detect insurance fraud. To achieve this goal, we designed and implemented a series of experiments by creating diverse detection scenarios using a combination of approaches such as feature discretization, feature selection and resampling the data, as well as different classifiers.

4.1 Experimental Setup

Experiments were run on a machine with a 2.60 GHz Intel(R) Core (TM) i7- 12,700 F CPU and 32 GB RAM. We use 64-bit Windows 11. Python is used to implement the Framework. The Pandas data frame loads the dataset. The Scikit Learn (Pedregosa et al., 2011) library implements ML and ensemble models. To ensure reproducibility of experimental models, parameter configurations, and reported results, we made the proposed work's source code, visualizations, and data openly available on the author's GitHub website[3].

4.2 Experimental Design

In our study, we aim to explore how combining various technologies in a unified framework might improve the building of predictive models, leading to the development of a reliable system for detecting insurance fraud accurately. We want to determine the effectiveness of this integrated strategy in enhancing prediction models and enhancing the accuracy and dependability of fraud detection mechanisms in the insurance sector through thorough study and experimentation.

Thus, our study comprised six experiments and each of the experiments employed specific combinations of approaches with different scenarios as shown in Fig. 4. We

[3] https://github.com/AhmedKhalil91/classification-model.git.

analyzed various conditions within each experiment to understand their individual effects. The foundation of the six experiments can be summarized as follows:

(a) In the first experiment, the data is used directly without processing discretization, feature selection or imbalance issues to fed directly into classification models, and then the performance of the models is evaluated using metrics like accuracy, F1-score, and AUC-ROC. This baseline evaluation serves as a reference point to compare the impact of various preprocessing techniques in subsequent experiments.
(b) The second experiment examines the effect of applying KBD discretization method on continuous features in the dataset, and then the modified data fed directly into classification models, and then the performance of the models is evaluated to determine how the discretization impacts model learning especially in handling continuous data more effectively compared to the baseline.
(c) The third experiment explores the use of feature selection (Kbest and RFE) techniques to reduce the feature space and potentially improve model performance. After applying the feature selection, the data is fed directly into classification models, and then the models' performance is evaluated using standard metrics to understand whether focusing on relevant features can enhance accuracy and reduce overfitting.
(d) The fourth experiment assesses the combined effect of KBD discretization followed by feature selection using KBest and RFE on classification performance. The data after applying the discretization and feature selection is fed directly into classification models, and then the performance of the models is evaluated. This experiment investigates the potential synergy between discretization and feature selection in enhancing predictive accuracy and model robustness
(e) In the fifth experiment, the impact of resampling (Under, Over, and SMOTE) techniques is analyzed to handle class imbalance in the dataset. Each resampling method is applied separately to balance the training set, and the models are trained using the resampled data and then the performance of the models is evaluated with a focus on identifying which resampling method best improves results in cases of class imbalance.
(f) The final experiment involves a comprehensive preprocessing approach that applies KBD discretization, followed by feature selection, and resampling (under-sampling, over-sampling, and SMOTE) to address class imbalance. The processed data is then used to train classification models, and their performance is evaluated with metrics. This experiment aims to demonstrate the cumulative benefits of integrating multiple preprocessing techniques and their overall impact on classification performance.

4.3 Experimental Results and Discussion

In this section, we provide a comprehensive examination of experimental results which emphasizes the critical impact of various data preprocessing techniques on classifier performance. The experiments followed a systematic approach, beginning with dataset preprocessing, splitting it into training and testing sets in an $80-20$ ratio,

and then testing classifiers across different preprocessing scenarios. Performance metrics including accuracy, F1 score, recall, precision, and AUC, were recorded and are detailed in Table 5. Notably, for imbalanced datasets, the AUC and F1 score were prioritized over accuracy, as these metrics better address class distribution, and reduce bias when one class is overrepresented.

Experiment 1 In the baseline scenario without data transformations, each classifier was tested for raw predictive capacity. Results showed that the Decision Tree (DT) classifier excelled with an AUC of 68.15%, followed by the RF model with AUC score 52.14% in the first dataset, while Gradient Boosting (GB) performed better in the second dataset, achieving an AUC of 68.08%. These scores reflect each model's basic efficacy without enhancements from data transformation.

Experiment 2 The study next investigated the effects of data discretization on classification performance by applying the KBD technique. The combination of KBD + DT demonstrated notable improvement on the first dataset with an AUC of 69.89%, while KBD + GB led on the second dataset with an AUC of 68.43%. This increase in AUC indicates that data discretization can sharpen the classifier's ability to handle class distinctions.

Experiment 3 To assess the impact of feature selection, we applied two methods, KBest and RFE. This experiment evaluated the classifiers with reduced feature sets into two scenarios. In the first dataset, the combination KBest + DT attained the highest AUC of 75.35%, surpassing the baseline, while KBest + GB performed best on the second dataset with an AUC of 67.57%. These results highlight that targeted feature selection can improve classification accuracy by reducing noise and focusing on the most informative attributes.

Experiment 4 This experiment integrated both discretization and feature selection, showing enhanced results across both datasets. Specifically, the (KBD + KBest + DT) combination achieved the highest AUC of 77.34% in the first dataset, while (KBD + KBest + GB) scored 67.89% in the second dataset. These findings suggest that the combination of discretization and feature selection strengthens the model by simultaneously reducing dimensionality and emphasizing essential features.

Experiment 5 We examined the effects of three resampling techniques (Undersampling, Oversampling, and SMOTE) on classifier performance in this experiment with three scenarios, which aimed to address class imbalance. The results show the (Oversampler + RF) combination demonstrated the highest performance among classifiers with AUC of 95.5% in the first dataset and 88.17% in the second dataset. These outcomes confirm that resampling methods, especially Oversampling, can dramatically improve model performance by balancing class distribution and enabling the model to learn from minority classes more effectively.

Experiment 6 This final experiment assessed the combined influence of discretization, feature selection, and resampling on classifier efficacy. Six scenarios were

tested, revealing that (KBD + RFE + Over + RF) combination achieved the highest AUC scores of 99.26% on the first dataset and 89.29% on the second. This result marks a substantial improvement over other scenarios suggesting that integrating all three preprocessing techniques is a powerful strategy for maximizing classifier accuracy and reliability in distinguishing between classes.

To provide a clear and concise overview, Tables 6 and 7 summarize this comparison, focusing on AUC score and F1 score as the primary metrics for selecting the top-performing scenario in each experiment. This metric quantifies the model's ability to differentiate between classes, enabling efficient identification of the models that demonstrate superior discriminatory power within the specific context of each experiment. Figures 5 and 6 show the ROC plot for the best scenario within each experiment in both datasets, as follows: (A) Experiment 1, (B) Experiment 2, (C) Experiment 3, (D) Experiment 4, (E) Experiment 5, (F) Experiment 6.

The results show how each preprocessing step affects the model's ability to classify data effectively, with specific techniques leading to significant improvements in performance metrics for both datasets. Resampling techniques, particularly Oversampling and SMOTE tend to show a notable increase in model performance and highlight their effectiveness in addressing class imbalance issues.

The detailed analysis of Table 5 reveals those specific combinations of preprocessing techniques consistently improved model accuracy across experiments. For instance, decision trees showed optimal performance when paired with KBD discretization, RFE feature selection, and SMOTE resampling, a synergy that provided superior accuracy and generalization.

The pattern holds across models. For Random Forest, combining KBD discretization, RFE feature selection, and Oversampling consistently yielded the highest scores across datasets, affirming that these preprocessing strategies enhance performance, particularly for tree-based models. For ensemble methods like AdaBoost and Gradient Boosting, the best results were also obtained by pairing KBD discretization, RFE or KBest feature selection, and SMOTE resampling, suggesting that these techniques offer considerable benefits for ensemble models in maintaining accuracy and preventing overfitting. Similarly, the Bagging classifier performed best with KBD discretization, RFE feature selection, and Oversampling, showing a preference for this resampling technique over SMOTE in managing class imbalance effectively.

4.4 Statistical test analysis

Different scenarios arise from various resampling and feature selection procedures, and these variations directly influence the performance of classifiers. Identifying the best strategy becomes challenging when different datasets possess unique characteristics, and varying data preprocessing options can impact classifier accuracy and robustness. Due to this complexity, determining the optimal combination of methods for evaluating and comparing classifiers across different data preprocessing situations necessitates a systematic approach.

Statistical significance tests including the ANOVA, and the Friedman test are helpful for comparing performance outcomes objectively and carefully evaluating these

Table 5 The performance of the classification models on different scenarios

Scenarios	Dataset)1(					Dataset (2)				
	AC	F1	RC	PR	AUC	AC	F1	RC	PR	AUC
DT	0.7979	0.2592	0.5477	0.1698	0.6815	0.6846	0.4601	0.4695	0.4511	0.6202
KBD+DT	0.7824	0.2634	0.6030	0.1685	0.6989	0.6992	0.4842	0.4932	0.4755	0.6375
Kbest+DT	0.6313	0.2384	0.8945	0.1376	0.7538	0.6977	0.4795	0.4864	0.4728	0.6345
RFE+DT	0.7980	0.2539	0.5327	0.1667	0.6745	0.6934	0.4742	0.4831	0.4657	0.6304
Under +DT	0.7027	0.7059	0.7213	0.6911	0.7029	0.6051	0.6033	0.6134	0.5935	0.6053
Over +DT	0.9131	0.9165	0.9561	0.8800	0.9032	0.8543	0.8565	0.9379	0.8151	0.8530
SMOTE+DT	0.8936	0.8945	0.9049	0.8844	0.8936	0.7558	0.7591	0.7577	0.7606	0.7558
KBD+Kbest+DT	0.6154	0.2427	0.9548	0.1390	0.7734	0.7065	0.4912	0.4949	0.4875	0.6431
KBD+RFE+DT	0.7746	0.2692	0.6432	0.1702	0.7135	0.6875	0.4533	0.4525	0.4541	0.6172
KBD+Kbest+Under +DT	0.7541	0.7807	0.8852	0.6983	0.7555	0.6406	0.6339	0.6357	0.6322	0.6405
KBD+Kbest+Over +DT	0.7860	0.8121	0.9274	0.7223	0.7864	0.8637	0.8757	0.9453	0.8156	0.8623
KBD+Kbest+SMOTE+DT	0.8196	0.8382	0.9367	0.7584	0.8199	0.7905	0.7933	0.7915	0.7952	0.7905
KBD+RFE+Under +DT	0.7405	0.7460	0.7705	0.7231	0.7409	0.6369	0.6453	0.6747	0.6184	0.6377
KBD+RFE+Over +DT	0.9107	0.9160	0.9595	0.8581	0.9108	0.8524	0.8665	0.9427	0.8017	0.8510
KBD+RFE+SMOTE+DT	0.9226	0.9209	0.9108	0.9519	0.9225	0.7806	0.7862	0.7941	0.7784	0.7804
RF	0.9361	0.0837	0.0452	0.5625	0.5214	0.7598	0.4343	0.3220	0.6667	0.6287
KBD+RF	0.9342	0.0379	0.0201	0.3333	0.5087	0.7487	0.4536	0.3644	0.6006	0.6336
Kbest+RF	0.9364	0.0485	0.0251	0.7143	0.5122	0.7516	0.4386	0.3390	0.6211	0.6280
RFE+RF	0.9339	0.0377	0.0201	0.3077	0.5085	0.7545	0.4403	0.3373	0.6338	0.6296
Under +RF	0.7676	0.7749	0.8087	0.7437	0.7680	0.6770	0.6716	0.6747	0.6685	0.6769
Over +RF	0.9550	0.9550	1.0000	0.9401	0.9550	0.8824	0.8874	0.8720	0.8753	0.8817
SMOTE+RF	0.9474	0.9473	0.9474	0.9471	0.9474	0.8167	0.8122	0.7805	0.8466	0.8173
KBD+Kbest+RF	0.9351	0.0385	0.0201	0.4444	0.5092	0.7428	0.4444	0.3593	0.5824	0.6280
KBD+RFE+RF	0.9342	0.0379	0.0201	0.3333	0.5087	0.7322	0.3947	0.3051	0.5590	0.6043
KBD+Kbest+Under +RF	0.7351	0.7621	0.8579	0.6856	0.7364	0.7061	0.7061	0.7212	0.6916	0.7064
KBD+Kbest+Over +RF	0.8027	0.8257	0.9367	0.7381	0.8031	0.8815	0.8889	0.9329	0.8488	0.8807
KBD+Kbest+SMOTE+RF	0.8386	0.8544	0.9499	0.7764	0.8389	0.8322	0.8271	0.7902	0.8677	0.8329

Table 5 (continued)

Scenarios	Dataset)1(					Dataset (2)				
	AC	F1	RC	PR	AUC	AC	F1	RC	PR	AUC
KBD+RFE+Under +RF	0.7000	0.6959	0.6940	0.6978	0.6999	0.7006	0.6876	0.6729	0.7029	0.7001
KBD+RFE+Over +RF	**0.9926**	**0.9926**	**1.0000**	**0.9953**	**0.9926**	**0.8928**	**0.8986**	**0.9349**	0.8650	**0.8921**
KBD+RFE+SMOTE+RF	0.9674	0.9665	0.9429	0.9913	0.9673	0.8312	0.8241	0.7785	**0.8755**	0.8321
AdaBoost	0.9303	0.0271	0.0151	0.1364	0.5042	0.7783	0.5215	0.4220	0.6822	0.6716
KBD+AdaBoost	0.9293	0.0180	0.0101	0.0870	0.5014	0.7821	0.5357	0.4390	0.6870	0.6794
Kbest+AdaBoost	0.9290	0.0352	0.0201	0.1429	0.5059	0.7734	0.5089	0.4102	0.6704	0.6646
RFE+AdaBoost	0.9293	0.0180	0.0101	0.0870	0.5014	0.7773	0.5204	0.4220	0.6785	0.6709
Under +AdaBoost	0.7514	0.7767	0.8743	0.6987	0.7527	0.7161	0.7194	0.7435	0.6969	0.7167
Over +AdaBoost	0.7581	0.7862	0.8918	0.7029	0.7584	0.7326	0.7397	0.7479	0.7317	0.7324
SMOTE+AdaBoost	0.8670	0.8726	0.9129	0.8357	0.8672	0.7833	0.7849	0.7785	0.7914	0.7833
KBD+Kbest+AdaBoost	0.9290	0.0352	0.0201	0.1429	0.5059	0.7787	0.5220	0.4220	0.6841	0.6719
KBD+RFE+AdaBoost	0.9296	0.0356	0.0201	0.1538	0.5062	0.7700	0.4881	0.3831	0.6726	0.6541
KBD+Kbest+Under +AdaBoost	0.7622	0.7963	0.9399	0.6908	0.7641	0.7207	0.7252	0.7528	0.6995	0.7213
KBD+Kbest+Over +AdaBoost	0.7601	0.7929	0.9208	0.6962	0.7605	0.7326	0.7420	0.7570	0.7276	0.7322
KBD+Kbest+SMOTE+AdaBoost	0.7672	0.7941	0.9001	0.7104	0.7675	0.8160	0.8184	0.8163	0.8206	0.8160
KBD+RFE+Under +AdaBoost	0.7541	0.7786	0.8743	0.7018	0.7553	0.7243	0.7297	0.7602	0.7015	0.7250
KBD+RFE+Over +AdaBoost	0.9317	0.9328	0.9142	0.6998	0.9321	0.8096	0.8089	0.8031	0.8052	0.8093
KBD+RFE+SMOTE+AdaBoost	0.9401	0.9401	0.9229	0.9175	0.9402	0.8140	0.8142	0.8020	0.8167	0.8142
GB	0.9364	0.0392	0.0201	0.8000	0.5099	0.7899	0.5369	0.4254	0.7275	0.6808
KBD+GB	0.9371	0.0490	0.0251	1.0000	0.5126	0.7914	0.5435	0.4339	0.7273	0.6843
Kbest+GB	0.9364	0.0392	0.0201	0.8000	0.5099	0.7870	0.5274	0.4153	0.7227	0.6757
RFE+GB	0.9364	0.0392	0.0201	0.8000	0.5099	0.7826	0.5214	0.4136	0.7052	0.6721
Under +GB	0.7595	0.7925	0.9290	0.6911	0.7613	0.7389	0.7476	0.7900	0.7095	0.7399
Over +GB	0.8089	0.8333	0.9575	0.7376	0.8093	0.7571	0.7673	0.7883	0.7474	0.7566
SMOTE+GB	0.8917	0.8965	0.9402	0.8566	0.8918	0.8104	0.8103	0.7974	0.8237	0.8106
KBD+Kbest+GB	0.9351	0.0291	0.0151	0.4286	0.5068	0.7865	0.5339	0.4271	0.7119	0.6789

Table 5 (continued)

Scenarios	Dataset)1(					Dataset (2)				
	AC	F1	RC	PR	AUC	AC	F1	RC	PR	AUC
KBD+RFE+GB	0.9351	0.0196	0.0101	0.4000	0.5045	0.7739	0.4868	0.3746	0.6950	0.6543
KBD+Kbest+Under +GB	0.7676	0.8009	0.9454	0.6948	0.7695	0.7225	0.7308	0.7695	0.6958	0.7234
KBD+Kbest+Over +GB	0.7789	0.8117	0.9557	0.7055	0.7794	0.7498	0.7626	0.7909	0.7362	0.7492
KBD+Kbest+SMOTE+GB	0.8093	0.8349	0.9672	0.7345	0.8097	0.8418	0.8416	0.8274	0.8664	0.8421
KBD+RFE+Under +GB	0.7703	0.7981	0.9180	0.7059	0.7719	0.7225	0.7350	0.7862	0.6900	0.7238
KBD+RFE+Over +GB	0.9581	0.9236	0.9454	0.8297	0.9584	0.8336	0.8493	0.7837	0.8178	0.8328
KBD+RFE+SMOTE+GB	0.9619	0.9605	0.9491	0.9941	0.9618	0.8382	0.8350	0.8059	0.8662	0.8387
Bagging	0.9342	0.0731	0.0402	0.4000	0.5180	0.7433	0.4437	0.3576	0.5845	0.6278
KBD+Bagging	0.9313	0.0619	0.0352	0.2593	0.5141	0.7453	0.4604	0.3797	0.5849	0.6358
Kbest+Bagging	0.9351	0.0385	0.0201	0.4444	0.5092	0.7555	0.4815	0.3966	0.6126	0.6480
RFE+Bagging	0.9303	0.0444	0.0251	0.1923	0.5089	0.7467	0.4435	0.3525	0.5977	0.6287
Under +Bagging	0.7081	0.7143	0.7377	0.6923	0.7084	0.7025	0.6964	0.6970	0.6957	0.7023
Over +Bagging	0.9414	0.9414	1.0000	0.9330	0.9414	0.8631	0.8690	0.8362	0.8646	0.8624
SMOTE+Bagging	0.9438	0.9443	0.9554	0.9334	0.9438	0.8024	0.7971	0.7642	0.8331	0.8031
KBD+Kbest+Bagging	0.9351	0.0385	0.0201	0.4444	0.5092	0.7477	0.4902	0.4237	0.5814	0.6507
KBD+RFE+Bagging	0.9316	0.0944	0.0553	0.3235	0.5237	0.7273	0.4218	0.3475	0.5366	0.6136
KBD+Kbest+Under +Bagging	0.7595	0.7855	0.8907	0.7026	0.7609	0.6815	0.6735	0.6710	0.6760	0.6813
KBD+Kbest+Over +Bagging	0.8034	0.8270	0.9419	0.7370	0.8038	0.8782	0.8861	0.9329	0.8438	0.8773
KBD+Kbest+SMOTE+Bagging	0.8364	0.8513	0.9391	0.7784	0.8366	0.8293	0.8241	0.7876	0.8642	0.8299
KBD+RFE+Under +Bagging	0.6865	0.6831	0.6831	0.6831	0.6864	0.7034	0.6895	0.6729	0.7070	0.7027
KBD+RFE+Over +Bagging	0.9891	0.9892	1.0000	0.9787	0.9892	**0.8786**	**0.8866**	**0.9349**	0.8431	**0.8776**
KBD+RFE+SMOTE+Bagging	0.9657	0.9648	0.9443	0.9863	0.9556	0.8299	0.8246	0.7870	**0.8659**	0.8306

Table 6 The comparison of the best scenario's performance on the experiments for dataset (1)

Experiment	The best scenario	Evaluation Metrics	
		AUC	F1 score
Experiment (1)	DT	0.6815	0.2592
Experiment (2)	KBD + DT	0.6989	0.2634
Experiment (3)	Kbest + DT	0.7538	0.2384
Experiment (4)	KBD + Kbest + DT	0.7734	0.2427
Experiment (5)	Over + RF	0.9550	0.9550
Experiment (6)	KBD + RFE + Over + RF	**0.9926**	**0.9926**

Table 7 The comparison of the best scenario's performance on the experiments for dataset (2)

Experiment	The best scenario	Evaluation Metrics	
		AUC	F1 score
Experiment (1)	GB	0.6808	0.5369
Experiment (2)	KBD + GB	0.6843	0.5435
Experiment (3)	Kbest + GB	0.6757	0.5274
Experiment (4)	KBD + Kbest + GB	0.6789	0.5339
Experiment (5)	Over + RF	0.8817	0.8874
Experiment (6)	KBD + RFE + Over + RF	**0.8921**	**0.8986**

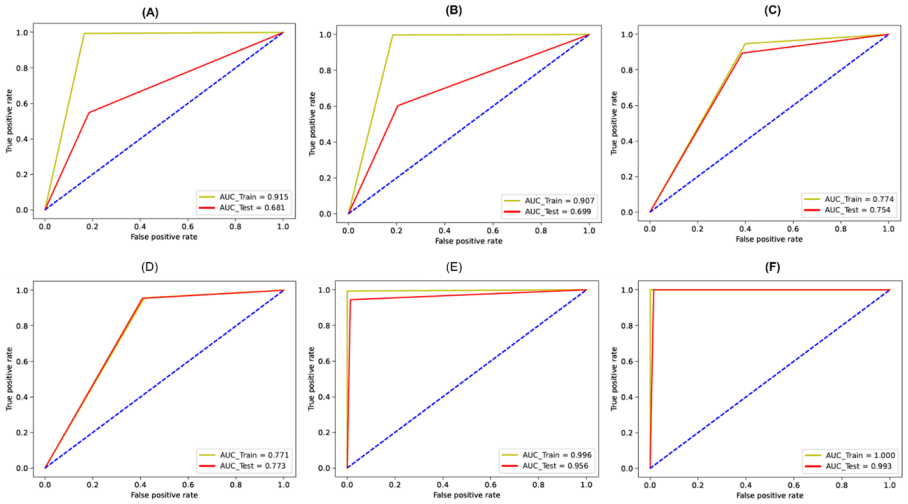

Fig. 5 The ROC plots for all the best scenarios in Dataset (1)

differences. As indicated in Table 8, we used both tests to look at how well classifiers did when given different combinations of resampling and feature selection. We were especially interested in the AUC values for each approach within each dataset. We were able to reject the null hypothesis because the p-value was less than the 0.05 threshold, which indicates that there are statistically significant differences. The results confirm the alternative hypothesis, which states that different scenarios within each dataset perform significantly differently. Insights like these are crucial for guid-

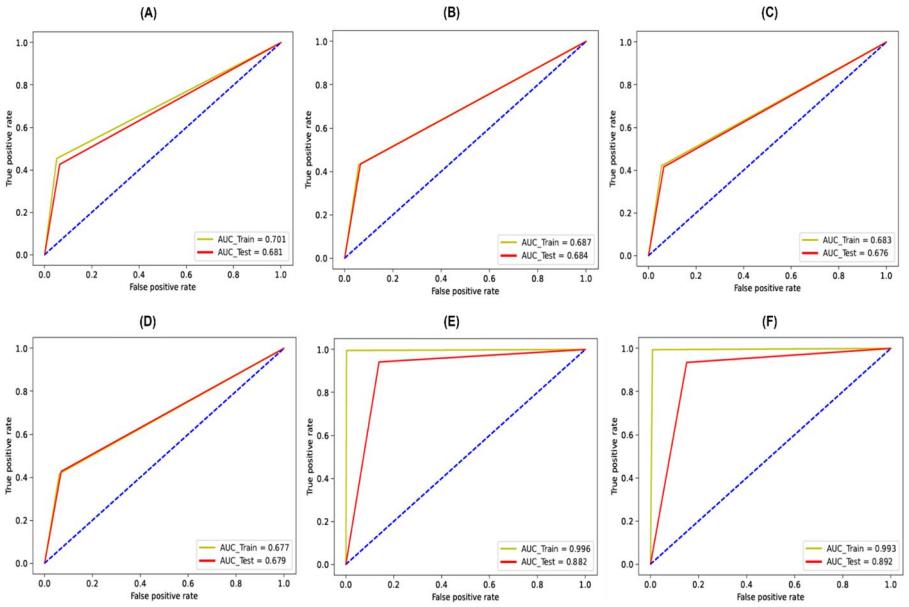

Fig. 6 The ROC plots for all the best scenarios in Dataset (2)

Table 8 Statistical analysis of classifier performance: ANOVA and Friedman test results

Test	ANOVA		Friedman	
Dataset	F	*P*-value	Chi-Square	*P*-value
Dataset (1)	22.997	5.54e-19***	61.9118	5.40e-08
Dataset (2)	18.18	1.49e-16***	61.0468	7.68e-08

ing feature selection and resampling strategies that maximize classifier efficacy for any specific dataset setup.

After confirming that the different approaches perform significantly differently within each dataset using ANOVA and the Friedman Test, further analysis is needed to identify the best scenario for each dataset. Tables 9 and 10 offer a comprehensive evaluation of various predictive models and preprocessing strategies, systematically analyzed across multiple experimental scenarios to identify the most effective combinations based on Median performance values, sum of ranks, and Friedman test rankings. The median metric reflects the central tendency of model performance, while the sum of ranks and rank columns provide insights into the relative performance of each scenario based on the Friedman test.

As shown in Table 9, the top-performing scenario for Dataset (1) is KBD + RFE + Over + ML, which achieves the highest rank (Rank 1) with a median AUC of 0.9584 and a sum of ranks of 75. This is closely followed by KBD + RFE + SMOTE + ML, which obtains the second rank (Rank 2) with a median AUC of 0.9456, and Under + ML, which achieves the third rank (Rank 3) with a median AUC of 0.8936. Notably, scenarios incorporating Recursive Feature Elimination (RFE) and oversampling techniques consistently outperform other approaches, indicating that the combination of feature selection and data balancing significantly improves model performance. In

Table 9 Comparison of predictive models across scenarios for dataset (1)

Experiment	Scenario	Median	Sum of rank	Rank
Experiment (1)	Raw data + ML	0.5180	19	13
Experiment (2)	KBD + ML	0.5126	16	14
Experiment (3)	Kbest + ML	0.5099	21	10.5
	RFE + ML	0.5089	7	15
Experiment (4)	KBD + Kbest + ML	0.7527	35	9
	KBD + RFE + ML	0.9032	58	4
Experiment (5)	Under + ML	0.8936	63	3
	Over + ML	0.5092	21	10.5
	SMOTE + ML	0.5087	20	12
Experiment (6)	KBD + Kbest + Under + ML	0.7609	44	7
	KBD + Kbest + Over + ML	0.7864	50	6
	KBD + Kbest + SMOTE + ML	0.8199	57	5
	KBD + RFE + Under + ML	0.7409	37	8
	KBD + RFE + Over + ML	0.9584	75	1
	KBD + RFE + SMOTE + ML	0.9456	71	2

contrast, simpler methods such as Raw data + ML and KBD + ML rank considerably lower (Ranks 13 and 14, respectively), underscoring the critical role of advanced preprocessing techniques. These findings highlight the substantial impact of preprocessing choices on model performance with the best results achieved through the integration of feature selection, and data balancing.

The results presented in Table 10 for Dataset (2) reveal that the KBD + RFE + Over + ML scenario delivers the highest performance attaining the top rank (Rank 1) with a median value of 0.8510 and a sum of ranks of 69. This is closely followed by KBD + Kbest + SMOTE + ML, which secures the second rank (Rank 2) with a median AUC of 0.8299, and KBD + RFE + ML which achieves the third rank (Rank 3) with a median AUC of 0.8530. In contrast, simpler approaches such as Raw data + ML (Rank 12, median = 0.6287) and SMOTE + ML (Rank 15, median = 0.6172) perform poorly underscoring the importance of advanced preprocessing strategies.

In summary, this analysis demonstrates that combining various preprocessing techniques can significantly improve classification model performance. The optimal combination, particularly for distinguishing tasks such as insurance fraud detection is found in Experiment 6 with the (KBD + RFE + Over + RF) scenario which showed the highest AUC score across both datasets. This combination effectively balances precision, and recall highlights the value of using integrated preprocessing approaches to optimize classifier accuracy and robustness. These findings provide valuable insights into the critical role of data preprocessing in optimizing classifier performance and can guide future research in selecting appropriate preprocessing strategies for specific model families.

Table 10 Comparison of predictive models across scenarios for dataset (2)

Experiment	Scenario	Median	Sum of rank	Rank
Experiment (1)	Raw data + ML	0.6287	18	12
Experiment (2)	KBD + ML	0.6375	28	10
Experiment (3)	Kbest + ML	0.6480	17	13.5
	RFE + ML	0.6304	17	13.5
Experiment (4)	KBD + Kbest + ML	0.7023	32	9
	KBD + RFE + ML	0.8530	63	3
Experiment (5)	Under + ML	0.8031	54	6
	Over + ML	0.6507	26	11
	SMOTE + ML	0.6172	6	15
Experiment (6)	KBD + Kbest + Under + ML	0.7064	39	8
	KBD + Kbest + Over + ML	0.8623	62	4.5
	KBD + Kbest + SMOTE + ML	0.8299	65	2
	KBD + RFE + Under + ML	0.7027	41	7
	KBD + RFE + Over + ML	0.8510	69	1
	KBD + RFE + SMOTE + ML	0.8306	62	4.5

4.5 Explaining Predictive Models Using SHAP Analysis

Interpreting the results of predictive models is a critical step in understanding their behavior and ensuring their reliability in real-world applications (de Souza et al., 2024). As a state-of-the-art interpretability framework, SHAP (SHapley Additive exPlanations) analysis measures the contribution of each feature to the model's predictions, which helps us understand the variables underlying this higher performance. Through the examination of SHAP values, we can identify the crucial features impacting the model's findings, verify the significance of chosen features, and derive practical insights into the dataset's underlying patterns (Lundberg & Lee, 2017). This interpretability step not only enhances trust in the model but also provides valuable guidance for refining preprocessing strategies and improving future predictive performance. As shown in Table 6 for dataset (1), the best-performing scenario as identified through comprehensive evaluation is KBD + RFE + Over + RF.

The SHAP analysis results are illustrated via SHAP values in Fig. 7, which determine the influence of features on the model's final results. SHAP values range approximately −0.4 to 0.4, indicating the strength and direction of each feature's impact on the model. Positive values (to the right) boost the model's output, whereas negative values (to the left) decrease it. This clarifies the significance of various features and their effects on the outcomes. Figure 7 illustrates the vertical ranking of features with the most significant positioned at the top and the least significant at the bottom. The primary features such as X31, X13, and X16 have the greatest SHAP values which indicate their substantial influence on the model's predictions.

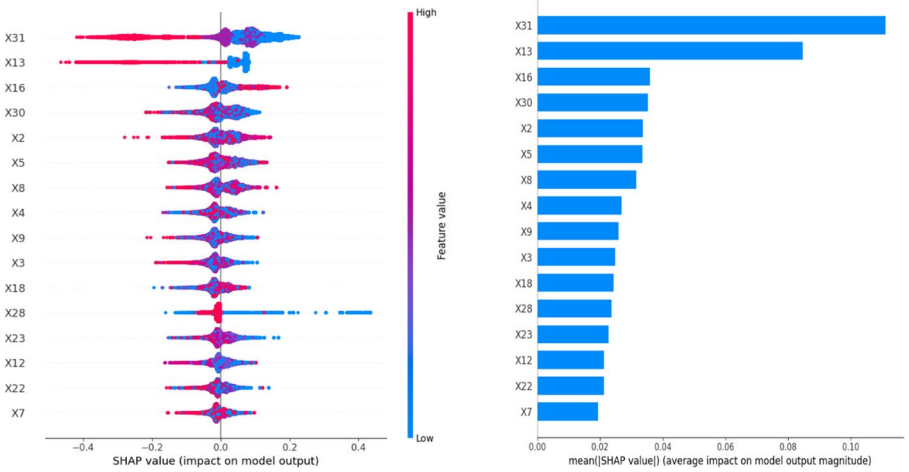

Fig. 7 SHAP Analysis: Feature importance and average impact on model output for insurance fraud dataset

According to the results, the variables X31, which represent the insurance coverage type, X13, which show the person responsible for the accident, and X16, which indicates the car pricing category, had the most significant impact with SHAP values ranging from 0.2 to 0.4. These features are likely the primary drivers of the model's performance and align with the success of the best-performing scenario (KBD + RFE + Over + RF), which combines feature selection and domain-specific knowledge. On the other hand, features like X12, X22, and X7 that have lower SHAP values make a small contribution which indicates that they are either unnecessary or of little relevance.

4.6 The Limitation of Study

While this study has developed a robust predictive system for insurance fraud detection through the integration of statistical and machine learning methods, several important limitations should be acknowledged. Most notably, our current analysis is constrained by the absence of detailed economic variables in the publicly available dataset, which precludes comprehensive economic impact assessments such as cost-benefit matrices, ROI analysis, or quantitative modeling of insurance-specific phenomena like moral hazard and adverse selection. These analyses would require access to proprietary financial metrics (e.g., claim-specific cost structures, policyholder premium histories, and loss ratios) that were beyond the scope of our data. We have therefore prioritized the development and validation of the core detection algorithm within the constraints of available data.

However, we recognize these economic dimensions as critical for demonstrating real-world implementation value, and propose that future research should: (1) establish industry partnerships to access sensitive financial data for comprehensive economic modeling; (2) develop integrated frameworks that combine fraud prediction with cost-impact analysis; and (3) investigate the behavioral economics aspects of

insurance fraud through richer policyholder datasets. These extensions would significantly enhance the practical utility of fraud detection systems while addressing the important intersections between technical prediction capabilities and business value demonstration that this study has identified.

The cross-sectional nature of our dataset prevents comprehensive temporal validation, as we lack longitudinal records of claim timestamps, historical fraud patterns, and scheme evolution documentation. This limitation restricts our ability to examine crucial dynamic aspects including seasonal fraud trends, behavioral adaptation patterns among fraudsters, and temporal model performance degradation - all of which are essential for maintaining detection accuracy in production environments. We emphasize that addressing these temporal dimensions through industry partnerships for longitudinal data collection represents a critical future research direction to bridge the gap between experimental validation and operational deployment.

Another limitation of our study is the lack of sensitive demographic data (e.g., race, gender) due to privacy constraints, preventing fairness assessments using metrics like demographic parity. While this protects claimant privacy, it limits our ability to evaluate potential biases—a critical concern given the insurance sector's vulnerability to discriminatory outcomes. Future work with appropriate data should rigorously audit predictions across demographic groups and implement fairness-aware modeling techniques to ensure equitable fraud detection systems.

5 Conclusion

In conclusion, this study provides an in-depth examination of fraud detection within the insurance industry through utilizing a multi-faceted approach that integrates diverse data preprocessing techniques and classification algorithms. The analysis investigates how various classifiers, feature selection methods, data discretization techniques, and resampling strategies impact the performance of fraud detection models. The empirical results reveal the significant advantages of combining these methods, with the Experiment 6 combination (KBD + RFE + Oversampling + Random Forest) demonstrating the highest efficacy in detecting fraudulent claims, as evidenced by superior metrics like AUC and F1-score. This approach highlights the importance of an integrated strategy in fraud detection, which shows that a well-rounded application of data mining and machine learning techniques can greatly enhance fraud detection accuracy and help insurers minimize financial losses.

The implications of this framework are substantial for both industry practitioners and policymakers. By adopting such a structured and data-driven approach, insurance companies can bolster their fraud detection systems which strengthen overall financial integrity and operational resilience. In addition, this framework not only serves as a guide for selecting optimal predictive models but also as a foundation for establishing more robust industry practices in fraud prevention.

While our study provides an effective framework for fraud detection, its cross-sectional design limits temporal validation due to the absence of longitudinal data (e.g., claim timestamps, fraud pattern evolution) and the lack of detailed economic variables (e.g., cost structures, policyholder financial histories) restricts cost-benefit

analyses, ROI quantification, and modeling of insurance-specific phenomena like moral hazard and adverse selection. Future research should extend this work by: (1) incorporating temporal analyses to assess seasonal trends, fraudster adaptation, and model decay; (2) developing ensemble/hybrid classifiers with advanced feature discretization and resampling strategies to improve robustness; (3) establishing industry partnerships to enable real-world validation with longitudinal datasets: and (4) Future economic modeling requires insurer collaboration to access financial data (claim costs, payment histories) for cost-benefit analysis, ROI quantification, and moral hazard/adverse selection studies. Additionally, testing novel feature selection techniques and adaptive learning methods will be critical to address evolving fraud tactics. These advancements would bridge the gap between experimental research and operational deployment, ultimately enhancing the insurance sector's resilience and sustainability. Our findings lay a foundation for these efforts, equipping practitioners with scalable tools while paving the way for a more secure and trustworthy fraud detection ecosystem.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability The data that support the findings of this study are available at this link (https://github.com/AhmedKhalil91/classification-model.git).

Declarations

Conflict of interest The author declares that has no relevant financial or non-financial interests to disclose.

References

Aivaz, K. A., Florea, I. O., & Munteanu, I. (2024). Economic fraud and associated risks: An integrated bibliometric analysis approach. *Risks*, *12*(5), 74.

Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., Khan, Z., & Ashraf, A. (2023). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research*, *327*(2), 633–657. https://doi.org/10.1007/s10479-022-05015-5

Alsuwailem, A. A. S., Salem, E., & Saudagar, A. K. J. (2023). Performance of different machine learning algorithms in detecting financial fraud. *Computational Economics*, *62*(4), 1631–1667. https://doi.org/10.1007/s10614-022-10314-x

Amirruddin, A. D., Muharam, F. M., Ismail, M. H., Tan, N. P., & Ismail, M. F. (2022). Synthetic minority over-sampling technique (SMOTE) and logistic model tree (LMT)-adaptive boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (*Elaeis guineensis*) using spectroradiometers and unmanned aerial vehicles. *Computers and Electronics in Agriculture, 193*, 106646. https://doi.org/10.1016/j.compag.2021.106646

Baesens, B., Höppner, S., Ortner, I., & Verdonck, T. (2021). RobROSE: A robust approach for dealing with imbalanced data in fraud detection. *Statistical Methods & Applications, 30*(3), 841–861. https://doi.org/10.1007/s10260-021-00573-7

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal, 3*, 100071.

Barry, L., & Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society, 7*(1), Article 2053951720935143. https://doi.org/10.1177/2053951720935143

Basit, M. S., Khan, A., Farooq, O., Khan, Y. U., & Shameem, M. (2022). Handling Imbalanced and Overlapped Medical Datasets: A Comparative Study. *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, 1–7. https://doi.org/10.1109/IMPACT55510.2022.10029111

Ben Jabeur, S., Stef, N., & Carmona, P. (2023). Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Computational Economics*, *61*(2), 715–741. https://doi.org/10.1007/s10614-021-10227-1

Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, *2*(4), 156–162.

Cappiello, A. (2020). Risks and Control of Insurance Undertakings. In A. Cappiello (Ed.), *The European Insurance Industry: Regulation, Risk Management, and Internal Control* (pp. 7–29). Springer International Publishing. https://doi.org/10.1007/978-3-030-43142-6_2

Cerda, P., & Varoquaux, G. (2022). Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering, 34*(3), 1164–1176. https://doi.org/10.1109/TKDE.2020.2992529

Das, S., Datta, S., Zubaidi, H. A., & Obaid, I. A. (2021). Applying interpretable machine learning to classify tree and utility pole related crash injury types. *IATSS Research*, *45*(3), 310–316. https://doi.org/10.1016/j.iatssr.2021.01.001

de Souza, M., de Castro, J. G., Peng, D. T., & Gartner, I. R. (2024). A machine learning-based analysis on the causality of financial stress in banking institutions. *Computational Economics, 64*(3), 1857–1890. https://doi.org/10.1007/s10614-023-10514-z

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 1–5.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2020). A secure AI-driven architecture for automated insurance systems: Fraud detection and risk measurement. *Ieee Access : Practical Innovations, Open Solutions, 8*, 58546–58558. https://doi.org/10.1109/ACCESS.2020.2983300

Gupta, S., Modgil, S., Bhattacharyya, S., & Bose, I. (2022). Artificial intelligence for decision support systems in the field of operations research: Review and future scope of research. *Annals of Operations Research*, *308*(1), 215–274. https://doi.org/10.1007/s10479-020-03856-6

Hanafy, M., & Ming, R. (2021). Improving imbalanced data classification in auto insurance by the data level approaches. *International Journal of Advanced Computer Science and Applications, 12*(6), 493–499.

Hassan, A. K. I., & Abraham, A. (2013). Computational intelligence models for insurance fraud detection: A review of a decade of research. *Journal of Network and Innovative Computing*, *1*(2013), 341–347.

Hassan, A. K. I., & Abraham, A. (2016a). Modeling insurance fraud detection using ensemble combining classification. *International Journal of Computer Information Systems and Industrial Management Applications*, *8*, 257–265.

Hassan, A. K. I., & Abraham, A. (2016b). Modeling insurance fraud detection using imbalanced data classification. In N. Pillay, A. P. Engelbrecht, A. Abraham, du M. C. Plessis, V. Snášel, & A. K. Muda (Eds.), *Advances in nature and biologically inspired computing* (pp. 117–127). Springer International Publishing.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Hoyt, R. E., Mustard, D. B., & Powell, L. S. (2006). The effectiveness of state legislation in mitigating moral hazard: Evidence from automobile insurance. *The Journal of Law & Economics, 49*(2), 427–450. https://doi.org/10.1086/501092

Itri, B., Mohamed, Y., Mohammed, Q., & Omar, B. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–4. https://doi.org/10.1109/ICDS47004.2019.8942277

Jovanovic, D., Antonijevic, M., Stankovic, M., Zivkovic, M., Tanaskovic, M., & Bacanin, N. (2022). Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. *Mathematics*, *10*(13), 2272.

Khalil, A. A., Liu, Z., & Ali, A. A. (2022a). Using an adaptive network-based fuzzy inference system model to predict the loss ratio of petroleum insurance in Egypt. *Risk Management and Insurance Review, 25*(1), 5–18. https://doi.org/10.1111/rmir.12200

Khalil, A. A., Liu, Z., Salah, A., Fathalla, A., & Ali, A. (2022b). Predicting insolvency of insurance companies in Egyptian market using bagging and boosting ensemble techniques. *Ieee Access : Practical Innovations, Open Solutions, 10*, 117304–117314. https://doi.org/10.1109/ACCESS.2022.3210032

Khalil, A. A., Liu, Z., Fathalla, A., Ali, A., & Salah, A. (2024a). Machine learning based method for insurance fraud detection on class imbalance datasets with missing values. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3468993

Khalil, A., Liu, Z., & Ali, A. (2024b). Precision in insurance forecasting: Enhancing potential with ensemble and combination models based on the adaptive neuro-fuzzy inference system in the Egyptian insurance industry. *Applied Artificial Intelligence, 38*(1), 2348413. https://doi.org/10.1080/08839514.2024.2348413

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 25–36.

Kowshalya, G., & Nandhini, M. (2018). Predicting Fraudulent Claims in Automobile Insurance. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 1338–1343. https://doi.org/10.1109/ICICCT.2018.8473034

Lakshmanarao, A., Srisaila, A., & Kiran, T. S. R. (2022). Machine Learning and Deep Learning framework with Feature Selection for Intrusion Detection. *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–5.

Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., & Zhang, C. (2021). CleanML: A study for evaluating the impact of data cleaning on ML classification tasks. *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 13-24. https://doi.org/10.1109/ICDE51399.2021.00009

Liu, J., Wu, C., & Li, Y. (2019). Improving financial distress prediction using financial network-based information and GA-based gradient boosting method. *Computational Economics, 53*(2), 851–872. https://doi.org/10.1007/s10614-017-9768-3

Liu, T., Zhu, X., Pedrycz, W., & Li, Z. (2020). A design of information granule-based under-sampling method in imbalanced data classification. *Soft Computing*, *24*(22), 17333–17347. https://doi.org/10.1007/s00500-020-05023-2

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*, 1–10.

NAIC (2022). *Special Investigative Unit (SIU) guidelines.*

Nordin, S. Z. S., Wah, Y. B., Haur, N. K., Hashim, A., Rambeli, N., & Jalil, N. A. (2024). Predicting automobile insurance fraud using classical and machine learning models. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(1), 911–921.

Park, Y., & Kwon, T. Y. (2024). Ensemble with divisive bagging for feature selection in big data. *Computational Economics*. https://doi.org/10.1007/s10614-024-10741-y

Pearsall, J. (1999). *Concise Oxford Dictionary 10th Ed*. Oxford UP.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Piovezan, R. P. B., de Andrade Junior, P. P., & Ávila, S. L. (2023). Machine learning method for return direction forecast of exchange traded funds (ETFs) using classification and regression models. *Computational Economics*. https://doi.org/10.1007/s10614-023-10385-4

Prasasti, I. M. N., Dhini, A., & Laoh, E. (2020). Automobile Insurance Fraud Detection using Supervised Classifiers. *2020 International Workshop on Big Data and Information Security (IWBIS)*, 47–52. https://doi.org/10.1109/IWBIS50925.2020.9255426

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. In *2017 international conference on circuit, power and computing technologies (ICCPCT)* (pp. 1–6). https://doi.org/10.1109/ICCPCT.2017.8074258

Saylor, A. T. (2023). *Recommendations to Enhance Deterrence and Detection of Opportunistic Auto Insurance Fraud for Insurance Professionals and Industry Partners* [Master Thesis, University of Wisconsin – Platteville]. https://minds.wisconsin.edu/bitstream/handle/1793/84189/Saylor,%20Andrew.pdf?sequence=1

Singh, S. K., & Chivukula, M. (2020). A commentary on the application of artificial intelligence in the insurance industry. *Trends in Artificial Intelligence, 4*(1), 75–79.

Singh, A., & Jain, A. (2019). Adaptive credit card fraud detection techniques based on feature selection method. In S. K. Bhatia, S. Tiwari, K. K. Mishra, & M. C. Trivedi (Eds.), *Advances in computer communication and computational sciences* (pp. 167–178). Springer Singapore.

Srivatsan, S., & Santhanam, T. (2021). Early onset detection of diabetes using feature selection and boosting techniques. *ICTACT Journal on Soft Computing*, *12*(1), 2474–2485.

Subudhi, S., & Panigrahi, S. (2018). Effect of class imbalanceness in detecting automobile insurance fraud. *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, 528–531. https://doi.org/10.1109/ICDSBA.2018.00104

Subudhi, S., & Panigrahi, S. (2020). Use of optimized fuzzy c-means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University - Computer and Information Sciences, 32*(5), 568–575. https://doi.org/10.1016/j.jksuci.2017.09.010

Sundarkumar, G. G., Ravi, V., & Siddeshwar, V. (2015). One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1–7. https://doi.org/10.1109/ICCIC.2015.7435726

Tayebi, M., & Kafhali, E., S (2024). Performance analysis of metaheuristics based hyperparameters optimization for fraud transactions detection. *Evolutionary Intelligence*, *17*(2), 921–939.

Turban, E. (2011). *Decision support and business intelligence systems*. Pearson Education India.

Visalakshi, S., & Radha, V. (2014). A literature review of feature selection techniques and applications: Review of feature selection in data mining. *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 1–6.

Wang, P., & Chen, X. (2020). Three-way ensemble clustering for incomplete data. *IEEE Access : Practical Innovations, Open Solutions, 8*, 91855–91864.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95. https://doi.org/10.1016/j.dss.2017.11.001

Xiaolong, X., Wen, C., & Yanfei, S. (2019). Over-sampling algorithm for imbalanced data classification. *Journal of Systems Engineering and Electronics*, *30*(6), 1182–1191. https://doi.org/10.21629/JSEE.2019.06.12

Zhang, X., Yu, L., & Yin, H. (2024). An ensemble resampling based transfer adaboost algorithm for small sample credit classification with class imbalance. *Computational Economics*. https://doi.org/10.1007/s10614-024-10690-6