Novel Approach to deal with Data Imbalance in Automobile Insurance Fraud Data

Abdus Saboor Gaffari Mohammed b00105302@aus.edu
Department of Computer Science & Engineering
Master of Science in Machine Learning
American University of Sharjah
Sharjah, UAE

Alizar Farhan b00106512@aus.edu
Department of Computer Science & Engineering
Master of Science in Computer Engineering
American University of Sharjah
Sharjah, UAE

Khalifa Alshamsi b00078654@aus.edu
Department of Computer Science & Engineering
Master of Science in Machine Learning
American University of Sharjah
Sharjah, UAE

Abstract

Insurance fraud is one of the most critical issues in the automobile fraud industry, as it not only results in significant financial losses to the insurance companies but also results in higher premium costs to the policy holders. To overcome the drawbacks of traditional methods, researchers have implemented machine learning techniques to efficiently detect fraudulent claims. But as there are less number of fraudulent insurance claims when compared to non-fraudulent claims, the automobile insurance claims data is mostly imbalanced, leading to diminished performance of ML models. To overcome these issues traditional oversampling and undersampling techniques are applied. Our research introduces a novel approach to balance the data by generating synthetic samples with the use of TVAE. We evaluated the performance of the the approach by training XGBoost and Random Forest classification models with data balance using TVAE and other traditional techniques such as SMOTE, ADASYN, Random Over Sampling and Random Under Sampling. In addition to balancing the data, we also improved the performance of the models by augmenting the data with more synthetic samples generated using the same TVAE. Our results show that using TVAE for balancing and augmenting the training data makes the model better ant generalizing to unseen patterns.

Keywords: TVAE, Machine Learning, Automobile Insurance Fraud Detection, insurance fraud, data imbalance, synthetic data, SMOTE, XGBoost, Random Forest

1 Introduction

The insurance sector is one of the oldest and largest financial sectors, providing financial protection from losses to individuals. Insurance is an agreement between an insurer and insured party (policyholder), in which the insurer promises financial compensation to the insured party in cases of an accident or specific loss (Viaene and Dedene, 2004), given that

the insured party has placed a claim. This process of claiming financial compensation poses as an opportunity for malicious actors, who place false claims. These fraudulent claims are referred to as insurance fraud. One of the areas of insurance sector that is particularly vulnerable to frauds is the automobile insurance.

Automobile insurance fraud is a growing concern for the insurance industry as it significantly affects both the insurers and policyholders. Fraudulent activities can be in various forms such as, over-inflated damage claims involving excessive repair costs, staged accidents including preplanned collisions involving one or more parties, and report fraudulent injuries to exploit medical coverage benefits. As these frauds go undetected, they increase the cost that is incurred to the insurance companies, who inturn increase the premiums of policies they provide, which ultimately leads to the honest customers in paying more for their policies. According to a report by Forbes (Kilroy, 2024), insurance frauds imposed severe economic consequences, with an estimated annual cost of $308.6 billion to the U.S. economy. This increased financial burden and cost for insurance companies is ultimately passed on to honest policyholders, who face increased premiums ranging between $400 and $700 annually per household.

In order to detect these activities in earlier times, insurance companies have sought out to investigating the claims by traditional methods. But these traditional methods rely on manual reviews and rule-based systems, which are time consuming, costly and prone to human error. The also introduces unwarranted scrutiny of legitimate claims (Andersson et al., 2020; Bermúdez et al., 2008). To tackle these issues, experts have come up with many statistical models to detect insurance frauds like the one presented by Belhadji et al., 2000, which are still lag in performance and efficiency.

The advent of machine learning (ML) and artificial intelligence (AI) has introduced transformative approaches to fraud detection. By leveraging advanced techniques such as anomaly detection, predictive modelling, and behaviour analysis, insurers can identify patterns indicative of fraud. However, these methods face a significant challenge: data imbalance. Fraudulent claims typically represent a small fraction of the total dataset, leading models to focus disproportionately on the majority class (genuine claims). This imbalance reduces the model's sensitivity to fraudulent cases, resulting in poor detection rates and overfitting problems (Phua et al., 2012). Additionally, the lack of publicly available high-quality datasets, due to privacy concerns, further complicates efforts to improve fraud detection systems, as researchers often struggle with insufficient or skewed data.

1.1 Contribution and Plan

As we have already discussed, the main issues that pertain to automobile insurance fraud detection using machine learning is the unavailability of data and the imbalance of fraud and non-fraud cases within the data. These issues forms the basis of our research question:

- How can we improve the performance of existing machine learning techniques?
- How can we deal with the issue of data imbalance?

To address both these questions, we plan to use synthetic data generated using TVAE (Xu et al.). To tackle the first question we generated synthetic records for both the classes

using TVAE and augmented them to the original data, effectively increasing the size of the dataset. By creating more synthetic samples in the training data we aim to help make the model more robust and better generalizable to the unseen test data.

To tackle the second question, we used the TVAE again, but in this case, we only generated synthetic data for the minority class, making sure that we only generated just enough samples so that the number of fraud and non-fraud cases become equal. Using these two approaches, we plan to prove how TVAE can be used for effeciently balance the data and achieve better performance at the same time using oversampled synthetic data. Finally, for the classification task, we plan to use ensemble decision tree models like XGBoost and Random forest, which have proven to work well the the dataset we have chosen.

To summarize, this study focuses on addressing the issue of data imbalance in automobile insurance fraud detection while improving the performance of machine learning techniques. Specifically, it explores the generation of synthetic datasets, the application of advanced ML models, and the challenges posed by limited data availability. The research aims to contribute to the development of robust and efficient fraud detection frameworks that not only mitigate the economic impact of fraud but also enhance the experience of genuine policyholders.

The rest of the paper is organized in the following manner, Section 2 will introduce to some key concepts that the paper deals with followed by Section 3 talking about some of the related works in the same domain and their challenges. Section 4 will introduce to the chosen datset, the proposed methodology, and data preprocessing techniques employed. This is followed by the details on evaluation method used, results and discussions on the results in Section 5. Finally Section 6 will provide the conclusions and future directions.

2 Background

2.1 Data Imbalance

Data Imbalance refers to a situation in machine learning and data analysis where the distribution of classes in a dataset is highly uneven, with one class significantly outnumbering the other(s). This is common in real-world scenarios such as fraud detection, medical diagnosis, and anomaly detection, where the minority class (e.g., fraudulent claims, rare diseases) is often the most critical but under-represented. Data imbalance poses challenges for machine learning models, as they tend to favour the majority class, leading to poor performance in predicting the minority class. To address this, techniques such as oversampling (e.g., SMOTE Chawla et al., 2002) or undersampling can be employed. Additionally, cost-sensitive learning and ensemble methods are also effective in enhancing the model's focus on minority class predictions (He and Garcia, 2009). Proper evaluation metrics like F1-score, precision-recall curves, and area under the precision-recall curve (AUC-PR) are essential for assessing performance on imbalanced datasets (Fernández et al., 2018). By addressing data imbalance, models can better generalize and improve performance in critical applications, ensuring fairness and reliability in decision-making tasks.

2.2 Data Leakage

Data leakage in machine learning occurs when the attributes and features of the test data are introduced in the training data, and these attributes or features won't be available during the actual prediction stage of the model. Data leakage lead to having a good or excellent performance in the training and testing phase, but fails in the production phase. The common causes of data leakage include improper preprocessing, flawed feature engineering, or cross-validation errors Kaufman et al. (2012). Data leakage can also be a result of oversampling the data before the data is split for training. When the data is split into train and test after oversampling, the data samples that are very similar, or identical in some cases, to the test data are introduced in the training data which leads the model to overfit to the test resulting in good testing performance. Data leakage is of more concern in cases like fraud and fault identification as these involve detection of anomalies and outliers. Data leakage in these cases causes the model to memorize the fraudulent data rather than understanding the underlying patterns (Baesens et al., 2021).

2.3 Synthetic Data

Synthetic Data refers to artificially generated data that imitates actual-world information at the same time as keeping its statistical and structural format. According to Patki et al., 2016b, synthetic data may be created by the usage of algorithms, simulations, or generative models like GANs (Generative Adversarial Networks) is used when real information is inaccessible, constrained, or sensitive. In simple terms, synthetic data acts alternatively for real records. For example, in training gadget learning models, synthetic data allows builders to test their algorithms without exposing non-public or private information. Due to privacy constraints, it is typically utilized in fields like healthcare, finance, and autonomous structures to stability privacy, price, and scalability while ensuring information diversity and application.

3 Related Works

This literature review explores various balancing techniques from recent studies on predicting auto insurance fraud, highlighting their contributions and limitations, followed by a critical analysis. These methods will serve as baseline methods for comparisons.

One of the early studies Patel and Subudhi (2019) proposed a methodology for detecting fraudulent auto insurance claims using an Extreme Learning Machine (ELM). To ensure consistent scaling and reduce the impact of differing feature ranges on classifier performance, numerical features were normalized to a range between 0 and 1 using Min-Max normalization. The ELM model consists of a single-hidden-layer feed-forward neural network with neurons and their weights randomly initialized, emphasizing high training time. A sigmoid activation function was employed for classification. The effectiveness of the model depends on two critical hyperparameters: the number of hidden layer nodes (Q) and the regularization parameter (C), which were optimized using grid search to identify the best-performing combination. The study utilized the "carclaims.txt" dataset and divided it into training and testing sets (80%/20%) while maintaining the class distribution through stratified random sampling. Consequently, experimental results highlighted the ELM's performance,

achieving sensitivity and specificity of 47.40% and 74.98%, respectively, validated through 10-fold cross-validation. The proposed ELM method outperformed existing models such as Probabilistic Neural Network (PNN), Multi-Layer Perceptron (MLP), Decision Tree (DT), and Group Method of Data Handling (GMDH).

Harjai et al. (2019) evaluated the effectiveness of combining the Synthetic Minority Oversampling Technique (SMOTE) with the Random Forest (RF) classifier for detecting automobile insurance fraud. The methodology began with data preprocessing including cleaning steps such as removing redundant data, duplicates, and outliers, as well as transforming seven time-related features into two standardized date-time features in the YYYY-MM-DD format. Due to the severe class imbalance in the dataset "carclaims.txt" (6% minority class), SMOTE was applied to balance the data before splitting it into an 80/20 training and testing set. Subsequently, the non-parametric Random Forest classifier was employed known for leveraging its ability to identify important features through its bagging approach. This involved constructing 100 parallel decision trees, with each tree trained on different subsets of features using random sampling with replacement, and determining the final classification through majority voting. The proposed approach achieved an accuracy of 94.3%, precision of 98.6%, recall of 45.1%, and an F1-score of 61.9%, validated using 10-fold cross-validation. These results highlight the effectiveness of Random Forest and SMOTE in addressing imbalanced datasets for fraud detection.

Similarly, in another study, Salmi and Atif (2022) conducted a comprehensive evaluation comprising six experiments to assess the effectiveness of SMOTE and Random Over-Sampling Examples (ROSE) as oversampling techniques in combination with Random Forest (RF) and Logistic Regression (LR) classifiers for predicting automobile insurance fraud. Data preprocessing involved removing identification and illogical variables, such as zeros in 'Age,' 'MonthClaimed,' and 'DayOfWeekClaimed,' and applying label encoding to categorical features. Two feature subsets were examined: one with 23 features derived from a prior study, and another with 5 features ('BasePolicy,' 'Fault,' 'VehicleCategory,' 'AddressChange_Claim,' and 'AccidentArea') selected for their high correlation with the target variable using a Chi-Square test. The dataset, "carclaims.txt," was split into training and testing sets using a 75/25 proportional stratified sampling method to preserve class distribution in both sets. Recognizing the challenges of oversampling or undersampling techniques, such as overfitting or loss of relevant information, respectively, the authors employed SMOTE and ROSE to oversample the minority class in the training set to match the majority class. SMOTE generated synthetic samples using five nearest neighbours ($k = 5$), while ROSE applied a "smooth bootstrap" approach. Results demonstrated that RF consistently outperformed LR across all feature and oversampling combinations. SMOTE and ROSE produced comparable metrics, with the best performance achieved using RF and ROSE on the 23-feature subset (Accuracy: 64.36%, Recall: 93.07%, Specificity: 62.53%, F1-Score: 23.84%) and RF with SMOTE on the 5-feature subset (Accuracy: 61.34%, Recall: 95.24%, Specificity: 61.35%, F1-Score: 22.79%). These results highlight RF's strong performance and the effectiveness of both oversampling methods.

Padhi and Panigrahi (2020) proposed a novel hybrid approach that combines balancing techniques with ensemble algorithms to address the challenge of class imbalance in efficiently detecting insurance fraud claims. The authors highlighted that using either random oversampling or under-sampling techniques alone is insufficient to generate a balanced class

distribution or reduce skewness; additionally, these techniques tend to cause overfitting. Their hybrid method for balancing involves under-sampling the majority class (non-fraud) using fuzzy C-means clustering (FCM). This technique calculates the Euclidean distance between all data points in the majority class and the generated clusters, marking data points exceeding a specified threshold distance as outliers. The threshold distance was computed using the box-and-whisker technique. Simultaneously, the SMOTE technique was employed to oversample the minority class, increasing its instances to match those of the majority class. This dual approach not only balances the dataset but also excludes outliers and reduces skewness, thereby improving the dataset quality for model training. Subsequently, the balanced dataset was divided into standard 80/20% splits for training and testing. The training set was fed into three ensemble supervised classification techniques: support vector machine (SVM), multilayer perceptron (MLP), and K-nearest neighbours (KNN). The final classification was determined using the majority voting technique. The study utilized a real-world automobile insurance dataset, "carclaims.txt," and achieved average performance metrics of 94.2% recall, 73.0% specificity, and 81.2% accuracy using a 10-fold cross-validation technique. These results demonstrate a significant improvement in performance compared to the original imbalanced dataset.

A study by Wongpanti and Vittayakorn (2024) proposed a hybrid approach utilizing a one-dimensional Convolutional Neural Network (1D-CNN) for its excellent spatial feature extraction capabilities in combination with Conditional Tabular Generative Adversarial Networks (CTGAN) to enhance auto insurance fraud classification. The authors highlighted CTGAN's advantages over previous synthetic data generation techniques, such as Adaptive Synthetic Sampling (ADASYN), Random Oversample, and traditional Generative Adversarial Networks. GANs use a generator to generate synthetic data and a discriminator to distinguish it from real data. This process iterates until the discriminator can no longer differentiate between real and synthetic data, thereby producing realistic synthetic samples. CTGAN builds upon GANs but is specifically designed to handle categorical features effectively, making it particularly suitable for tabular datasets. The preprocessing steps involved removing multicollinearity using Pearson correlation to eliminate highly correlated features, followed by standardizing numerical features using the z-score method. Categorical features were encoded using label encoding for ordinal variables and one-hot encoding for nominal variables. The study employed the "carclaims.txt" dataset and performed a performance comparison between CTGAN with 1D-CNN as the classifier and the widely used SMOTE technique with the same classifier. Results demonstrated that CTGAN significantly outperformed SMOTE and ADASYN across all performance metrics, establishing it as a superior data balancing technique. CTGAN generated more realistic synthetic data with less noise and handled highly categorical datasets more effectively than SMOTE and ADASYN.

All except the first work implemented balancing techniques. Although these proposed methods aim to mitigate the issue of imbalanced datasets and improve overall performance metrics, achieving higher F1 scores remains a significant challenge. It is also beneficial to use these techniques as baseline methods for performance comparisons. Table 1 summarizes the balancing techniques reviewed in the literature, highlighting their respective advantages and disadvantages.

Upon further analysis of the papers, we identified biases in their methodologies. Papers such as (Harjai et al., 2019; Padhi and Panigrahi, 2020) and those discussed in a survey by

Schrijver et al. (2024) applied balancing techniques before splitting the data into training and test sets. This approach introduces data leakage that artificially inflates performance metrics and undermines the credibility of the reported results. While this may make the models appear effective, they are likely to fail in real-world production scenarios. To ensure unbiased and reliable evaluations, it is vital to first split the dataset and then apply balancing techniques to prevent data leakage and maintain the integrity of the results.

Table 1: Summary of various oversampling and undersampling methods used in related works

Work	**Balance Technique**	**Advantages**	**Disadvantages**
Patel and Subudhi, 2019	FCM + SMOTE	Under sampling via removing outliers + Generate synthetic samples to balance datasets	High computational cost due to clustering. Can introduce noise or synthetic data artifacts, leading to overfitting.
Salmi and Atif, 2022; Patel and Subudhi, 2019	Random oversampling	Easy to construct	Can lead to overfitting by duplicating existing data
Salmi and Atif, 2022; Patel and Subudhi, 2019	Random undersampling	Easy to construct, reduces computational cost	Can lead to loss of valuable information from the majority class
Patel and Subudhi, 2019; Harjai et al., 2019; Salmi and Atif, 2022; Wongpanti and Vittayakorn, 2024	SMOTE	Generate synthetic samples to balance datasets, widely used	Can introduce noise or synthetic data artifacts, leading to overfitting
Salmi and Atif, 2022	ROSE	Generates more diverse synthetic samples	Computationally intensive and may add noise to the dataset
Wongpanti and Vittayakorn, 2024	ADASYN	Generate synthetic samples for harder-to-classify instances	May skew data distribution, leading to overfitting in some cases
Wongpanti and Vittayakorn, 2024	GAN	Generates realistic synthetic data by mimicking data distribution	Requires extensive training and is computationally expensive

Wongpanti and Vittayakorn, 2024	CTGAN	Generates realistic synthetic data by mimicking data distribution, effectively handles categorical data	More complex and computationally intensive than GAN

4 Methodology

4.1 Dataset

The dataset used in this research is the *carclaims.txt* data that was originally made available as part of the Agnoss Knowledge Seeker product. The original copy of the data is no longer available for download, but many publicly available copies are available through GitHub[1] and Kaggle[2]. The dataset contains 15,420 samples of automobile insurance claims from an insurance company for the years from 1994 to 1996. Out of the 15,420, only 923 are fraudulent claims indicating a very high imbalance in the data (Figure 1). This particular dataset was chosen because there are not many publicly available datasets on automobile insurance fraud, and the ones that are available are very small. Another reason for choosing this data is that this has been extensively used in many other research on automobile insurance fraud detection (Schrijver et al., 2024; Salmi and Atif, 2022; Padhi and Panigrahi, 2020; Wongpanti and Vittayakorn, 2024; Harjai et al., 2019; Patel and Subudhi, 2019), making it easy for comparative analysis.

Table 2 shows the various columns of the data and their details. Out of 33 columns only two columns are numerical, Age and PolicyNumber, and the remaining columns are all categorical including the target feature FraudFound. The main drawback of this data is that due to its small size and high imbalance the model performances are usually low, leading researchers to employ oversampling techniques.

Table 2: Details about the columns of the dataset - *carclaims.txt*

Column	**Description**	**Type**	**Subtype**
PolicyNumber	Unique identifier for the policy	Numerical (PK)	Discrete
Month	Month when the incident occurred	Categorical	Ordinal
WeekOfMonth	Week of the month when the incident occurred	Categorical	Ordinal
DayOfWeek	Day of the week when the incident occurred	Categorical	Ordinal
Make	Make of the vehicle involved in the incident	Categorical	Nominal
AccidentArea	Area where the accident occurred (Urban or Rural)	Categorical	Nominal

1. https://github.com/Rashmi-77/Vehicle-Insurance-Fraud-Detection
2. https://www.kaggle.com/datasets/khusheekapoor/vehicle-insurance-fraud-detection

DayOfWeek-Claimed	Day of the week when the claim was made	Categorical	Nominal
MonthClaimed	Month when the claim was made	Categorical	Nominal
WeekOfMonth-Claimed	Week of the month when the claim was made	Categorical	Nominal
Sex	Gender of the policyholder	Categorical	Nominal
MaritalStatus	Marital status of the policyholder	Categorical	Nominal
Age	Age of the policyholder (or) policy [3]	Numerical	Discrete
Fault	Indicates fault (Policy Holder or Third Party)	Categorical	Nominal
PolicyType	Type of insurance policy	Categorical	Nominal
VehicleCategory	Category of the vehicle (e.g., Sport, Utility, Sedan)	Categorical	Nominal
VehiclePrice	Price range of the vehicle	Categorical	Ordinal
RepNumber	Identifier for the representative managing the case	Numerical	Discrete
Deductible	Deductible amount in the policy	Categorical	Ordinal
DriverRating	Driver rating (1 to 4)	Categorical	Nominal
Days:Policy-Accident	Days between policy start and accident	Categorical	Ordinal
Days:Policy-Claim	Days between policy start and claim	Categorical	Ordinal
PastNumberOf-Claims	Number of claims made in the past	Categorical	Ordinal
AgeOfVehicle	Age of the vehicle involved in the incident	Categorical	Ordinal
AgeOfPolicy-Holder	Age range of the policyholder	Categorical	Ordinal
PoliceReportFiled	Indicates if a police report was filed (Yes/No)	Categorical	Nominal
WitnessPresent	Indicates if a witness was present (Yes/No)	Categorical	Nominal
AgentType	Type of agent (Internal or External)	Categorical	Nominal
NumberOfSuppli-ments	Number of claim supplements	Categorical	Ordinal
AddressChange-Claim	Time since the last address change	Categorical	Ordinal
NumberOfCars	Number of cars in the policy	Categorical	Ordinal
Year	Year of the incident	Categorical	Nominal
BasePolicy	Basic policy coverage (e.g., Liability, Collision)	Categorical	Nominal
FraudFound	Indicates if fraud was found in the claim	Categorical	Nominal

3. This is not made clear in the data or any supporting sources

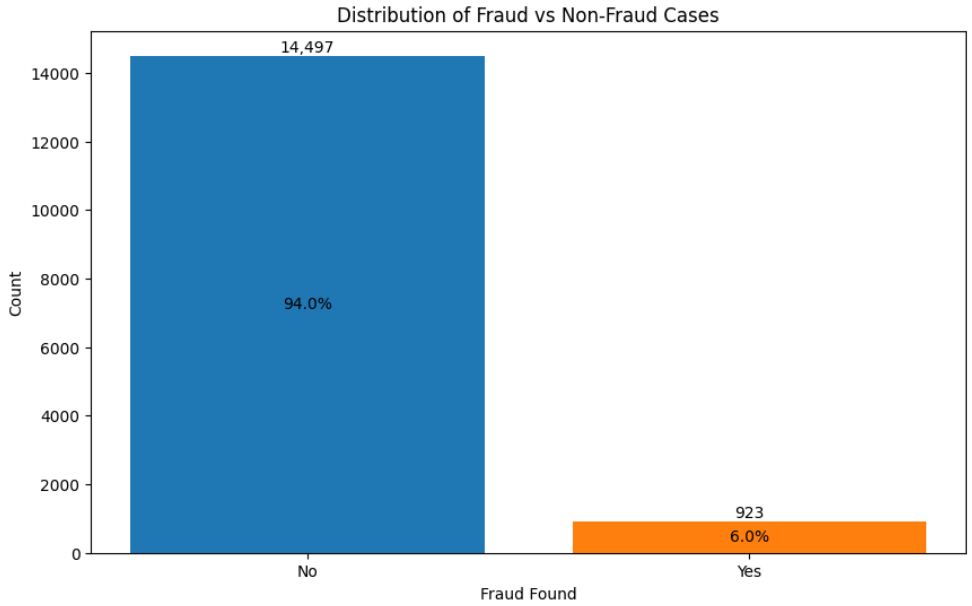

Figure 1: Data imbalance in carclaims.txt

4.2 Proposed Approach

The proposed model to balance the automobile insurance fraud data and to improve the performance of the ML models, uses a TVAE to generate synthetic data to balance the data and generate additional samples for training the ML models. As shown in Figure 2 in the first phase of training, after the dataset is split into train and test sets, the train set is used to train the TVAE. After the TVAE is trained, it is used to generate synthetic data that resembles the distributions of the original training data. The number of samples generated for each of the classes is determined by the imbalance between the classes and a predefined value that dictates the number of additional samples to be generated for both classes.

First the difference in the number of occurrences of the minority and majority classes is calculated. This difference is then added to the predefined value, this resulting value is used as the number of synthetic samples to be generated for the minority class using TVAE. Finally, the unchanged predefined value is used as the number of synthetic samples to be generated for the majority class. Combining these samples with the original training set gives us balanced and oversampled training data that can be used to train the classification models responsible for the prediction task.

The classification models chosen for the prediction task are the ensemble decision trees; XGBoost and Random Forest. These specific models were chosen as they are robust, scalable, and used heavily with the chosen dataset (Schrijver et al. (2024); Salmi and Atif (2022); Aiemsuwan and Srikamdee (2024); Owolabi et al. (2024)).

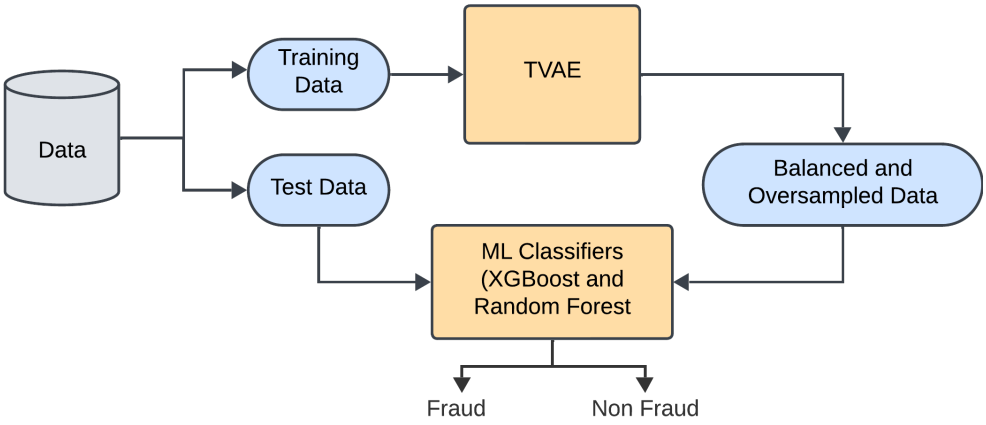

Figure 2: Block diagram of the proposed approach to balancing the training data and oversample both the classes using synthetic data generated using TVAE

4.3 Tabular Variational Autoencoder (TVAE)

TVAE is an adaptation of variational autoencoders that is designed to handle synthetic data generation for tabular data. It was proposed by Xu et al. as a baseline for comparative analysis against their generative model for tabular data, CTGAN. Both the models were designed to deal with the mixed types of data, continuous and discrete, often seen in tabular data. Continuous columns might have a group of values that occur more frequent and discrete or categorical columns may be imbalanced. Moreover, continuous data is usually non-Gaussian making then difficult to be normalized and tabular data usually contains multimodal distributions.

To deal with the issue of non-Gaussian continuous values TVAE and CTGAN use a *mode-specific* normalization and the discrete columns are *one-hot* encoded. The latent space of the VAE is modified as the joint distribution of $2N_c + N_d$, where N_c are the continuous columns and N_d are discrete columns. For the training and data generation task the continuous columns are assumed to follow a Gaussian distribution and discrete columns follow a categorical PMF. One drawback of TVAE over CTGAN, given in the paper by the authors, is the TVAE need to be trained on the actual data whereas CTGAN need not be.

The specific implementation of TVAE used is the one provided by SDV (Synthetic Data Vault, Patki et al., 2016a). SDV is a python library designed for creating synthetic tabular data.

4.4 Data Preprocessing

Before the data is fed to the machine learning models it need to be processed into format that can be easily understood and interpreted by the models. The TVAE implementation by SDV does not require any data preprocessing to be done from our side. The model does its own encoding and normalization depending on the type of features. In order to determine the type of features, SDV library requires us to pass the metedata of our data in the form of JSON. As we are using decision tree ensemble models like XGBoost and Random Forest,

we do not need to normalize or standardize the data, but we need to encode the features. Therefore, all the nominal features (As mentioned in Table 2) were one-hot encoded and all the ordinal features were label encoded.

5 Results and Analysis

5.1 Metrics

In this section, we evaluate the performance of the model using standard metrics such as accuracy, precision, recall, and F1-score. The formulas for these metrics are as follows:

Accuracy Calculates the proportion of correctly predicted instances out of the total observations. It reflects the overall effectiveness of the model but may be misleading in cases of imbalanced datasets.

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

Precision Calculates the proportion of correctly predicted positive instances out of all instances predicted as positive. In this context, it indicates the percentage of fraud predicted that are actually fraud.

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

Recall Calculate the proportion of actual positive instances that were correctly identified by the model. It indicates the percentage of truly fraud cases that were predicted as fraud.

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

F1 Score Calculates the harmonic mean of precision and recall. It provides a balanced evaluation of the model's performance.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.2 Evaluation

We will evaluate our proposed method by comparing the performance of the chosen classification models with different oversampling and undersampling techniques. The proposed approach only applies the oversampling on the training data after splitting the data for test and training, but in order to compare our performance with studies that performed oversampling before the split we have also evaluated additional results by oversampling before the split. The subsection 5.3.1 will discuss the results when the oversampling was performed after the split as mentioned in the proposed approach. The subsection 5.3.2 will discuss the results when the oversampling was applied before the split.

The TVAE implementation provided by SDV gives the flexibility to tune the model to our requirement through parameters like number of epochs, the encoder dimensions,

Table 3: Comparison of metrics with works mentioned in related works. All the works used the same carclaims.txt dataset

Work	Methods	Balance Techniques	Acc	Pre	Rec	F1
Patel and Subudhi (2019)	ELM	Not Mentioned	74.9%	–	–	–
Harjai et al. (2019)	Random Forest	SMOTE	94.3%	98.6%	45.1%	61.9%
Salmi and Atif (2022)	Random Forest	SMOTE ROSE	64.3% 61.34%	– 14.1%	93.07% 95.24%	23.8% 22.79%
Salmi and Atif (2022)	SVM MLP KNN	FCM (Under Sampling) SMOTE	81.2%	–	–	94.2%
Wongpanti and Vittayakorn (2024)	4-layer 1D-CNN	SMOTE CTGAN	81.3% 79.3%	13.6% 16.7%	39.8% 61.5%	20.3% 26.2%
Proposed	Random Forest XGBoost	TVAE	87.13%	28.66%	66.83%	40.12%

the decoder dimension, the latent space dimension etc. Therefore, to achieve the best results possible, hyper-parameter tuning was performed by employing Bayesian optimization methods provided by the libraries, *hyperopt* (Bergstra et al.) and *scikit-optimize*[4]. In addition to tuning the TVAE hyper-parameters, the parameters for XGBoost and Random Forest were also tuned, using the same library, for all the oversampling and undersampling methods, using the same search space.

5.3 Results

In this section we will look at the evaluation results of our proposed approach. In particular we will look at how the performance of XGBoost and RF are affected by different oversampling and undersampling techniques, and contrast them with the oversampling by using TVAE. The next two subsections will discuss the results with respect to balancing the data after the split and before the split respectively.

As shown in Table 3, when comparing the results of studies that first split the dataset and then applied balancing techniques (Salmi and Atif, 2022; Wongpanti and Vittayakorn, 2024), our proposed method achieves the highest F1 score, with a significant improvement

4. https://scikit-optimize.github.io/stable/index.html

of 14% over latest balance technique CTGAN. It also demonstrates the highest accuracy and precision, and the second-highest recall. Although the study by Salmi and Atif, 2022 outperformed our proposed method in terms of recall, their precision is less than half of our result, highlighting a trade-off in their approach. On the other hand, when comparing metrics to studies that balanced the data before splitting (Patel and Subudhi, 2019; Harjai et al., 2019; Padhi and Panigrahi, 2020), it is evident that their results are significantly inflated due to data leakage, which artificially enhances performance. Therefore, these results are not suitable for a fair and meaningful comparison.

5.3.1 Balance data after split

Figure 3 shows the performance metrics for XGBoost when the balancing of data was done only on the training data after the train-test split was performed. From the figure, we can gather that XGBoost was able to achieve the best F1 score, of 37.1% when the data balancing was done using TVAE. Moreover, when the data was augmented with additional synthetic samples by TVAE in addition to balancing, the model was able to achieve an F1 score of 40.1%. The TVAE balanced model was also able to achieve a good recall, indicating that the model is able to identify the fraudulent claims better with TVAE balanced data when compared to other traditional methods like SMOTE, ADASYN or ROS. Data balanced using RUS gives the best recall, but it comes at the cost of precision, meaning that the model is overfitting to the fraudulent cases as a result of the underrepresentation of genuine claims due to undersampling. The data balanced using ROSE achieves the best precision, but performs very poorly in terms of recall, indicating its inability to generalize to the test data. Whereas, the TVAE balanced data is able to achieve very good precision as well, leading to a good F1 score. Finally, in terms of accuracy, the TVAE balanced data achieves an accuracy of 89.1%, which is very close to the best accuracy provided by ROSE. Figure 4 shows the performance metrics for RF trained using data balanced after the train-test split. Similar to XGBoost, the TVAE balanced data was able to achieve a good F1 score but was not able to out perform ADASYN and ROS. But in terms of recall, TVAE achieved the best performance next in line to RUS, which as discussed earlier tends to overfit fraudulent cases leading to low precision. In terms of over all metrics RF seems to perform well with data balanced using ADASYN and ROS. The overall performance seems to indicate that the models trained on data balanced and augmented using TVAE tend to generalize better to the unseen test data. This might be due to the fact that TVAE generates synthetic data using a generative approach and distributions, whereas traditional approaches, like SMOTE, ADASYN or ROS oversample by creating duplicates of the data.

5.3.2 Balance data before split

Figure 5 shows the performance of XGBoost when it is trained on the data that is first balanced and then split into train and test. At a first glance its obvious that the results for all the balancing techniques are much better compared to the previous section. However, these results are due to the bias introduced by data leakage. The data balanced, and augmented, using synthetic data generated by TVAE achieved the best accuracy and F1 score in both the cases. The model trained on augmented data performs slightly better compared to the one trained on just balanced data (using TVAE). Following the trend

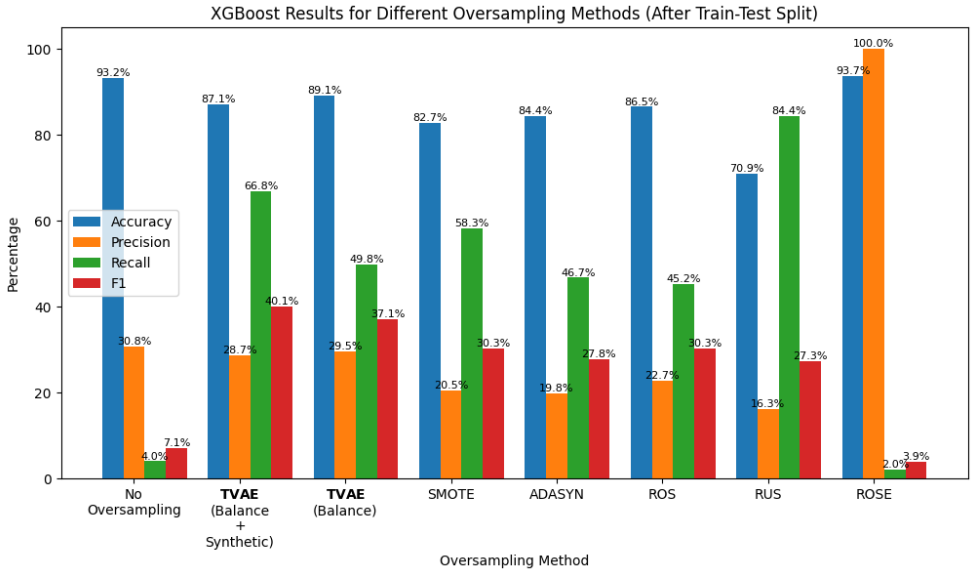

Figure 3: Performance of XGBoost classifier model on data balanced using different techniques after the train-test split

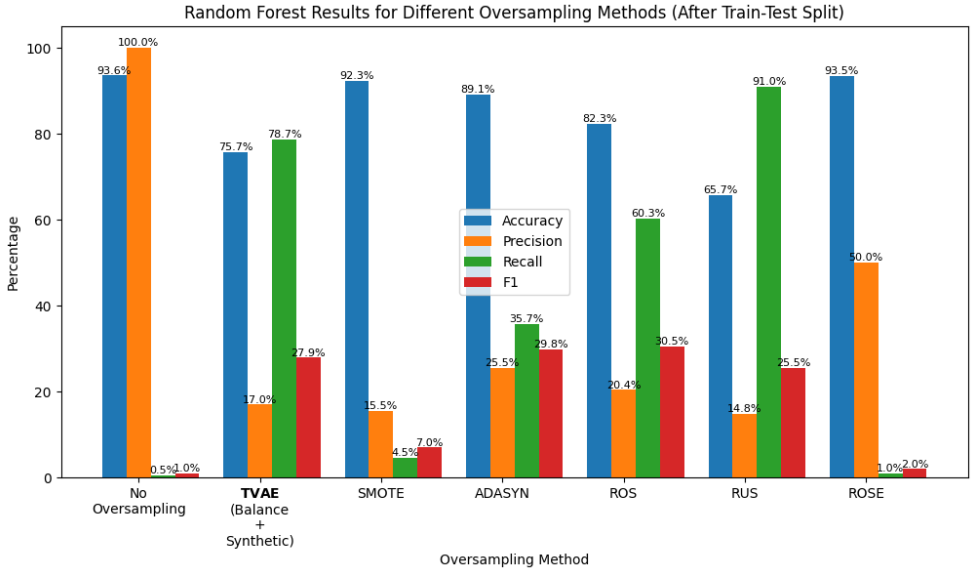

Figure 4: Performance of Random Forest classifier model on data balanced using different techniques after the train-test split

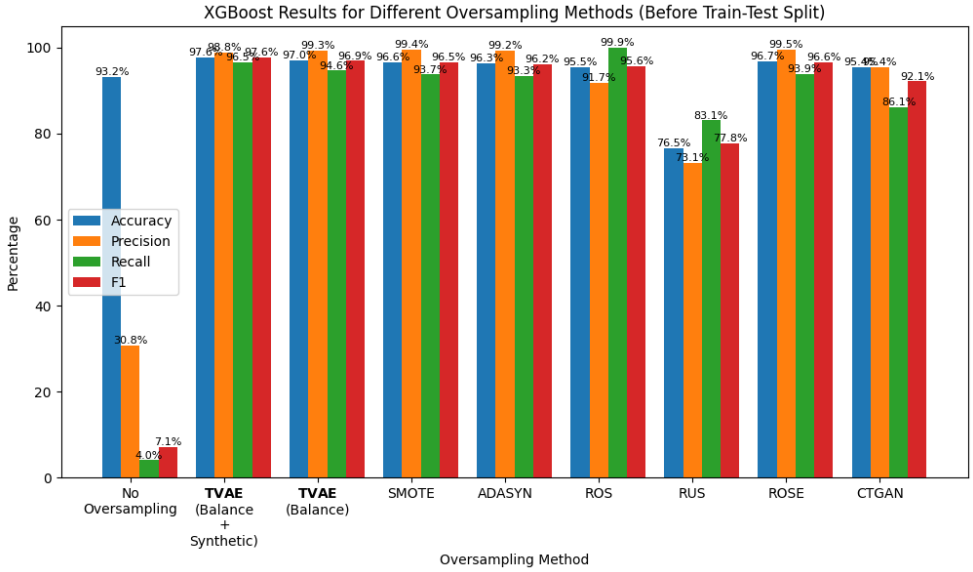

Figure 5: Performance of XGBoost classifier model on data balanced using different techniques before the train-test split

from previous results, RUS has the worst performance due under sampling of genuine cases. Figure 6 shows the performance of RF when its trained on the data that is balanced before the split. It can be seen that the results are very similar to the results with XGBoost.

6 Conclusion and Future Work

To summarize, this project addressed the challenge of severely imbalanced datasets in automobile insurance fraud detection while improving the performance of applied machine learning techniques. A comprehensive review of current balancing methods highlighted persistent challenges, such as handling high-categorical features and generating noisy synthetic data. This study utilized the "carclaims.txt" dataset, which is highly imbalanced and includes 30 categorical features. Preprocessing involved cleaning the data and applying one-hot and label encoding. To address the imbalance issue, TVAE was proposed as a balancing technique that generates realistic synthetic data and effectively handles categorical features. Experimental results demonstrated that TVAE, when combined with Random Forest and XGBoost, achieved the highest F1 score, both before and after splitting, surpassing other balancing techniques (CTGAN, ADAYSN, FCM, SMOTE, ROSE, RUS, and ROS). Furthermore, when more synthetic data generated using TVAE was added to the training data, the model performance improved slightly. These findings underscore the effectiveness of TVAE in mitigating dataset imbalance and improving fraud detection performance. This approach has broader applicability in fields like credit card fraud detection and beyond. Although we able to outperform existing state-of-the-art performance for the dataset, there is still room for improvement. Future research could explore advanced and broader hyperparameter optimization techniques to further fine-tune the parameters of TVA, XGBoost and

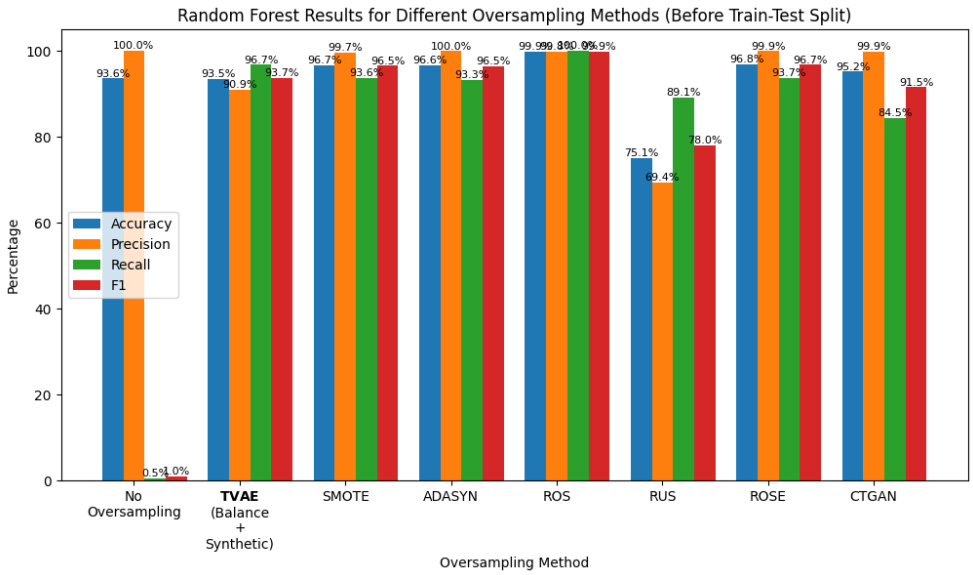

Figure 6: Performance of Random Forest classifier model on data balanced using different techniques before the train-test split

Random Forest. Additionally, modern clustering techniques could be leveraged to address the complexity of highly nuanced auto insurance fraud datasets. Incorporating ensemble learning methods may also yield significant improvements in detecting fraudulent claims. These efforts aim to advance the development of more precise and impactful methods for insurance fraud classification.

Acknowledgments and Disclosure of Funding

We would like to express our heartfelt gratitude to Dr. Alex Aklson for his guidance and American University of Sharjah for providing access to external GPU resources which significantly speeded up the training process

References

Phannana Aiemsuwan and Supawadee Srikamdee. A Novel Hybrid Method for Imbalanced Automobile Insurance Fraud Detection. In *2024 16th International Conference on Knowledge and Smart Technology (KST)*, pages 12–17, February 2024. doi: 10.1109/KST61284.2024.10499643.

Jonas Andersson, Andreas Olden, and Aija Rusina. Fraud detection by a multinomial model: Separating honesty from unobserved fraud. -12-31 2020.

Bart Baesens, Sebastiaan Höppner, Irene Ortner, and Tim Verdonck. robROSE: A robust approach for dealing with imbalanced data in fraud detection. *Statistical Meth-*

ods & Applications, 30(3):841–861, September 2021. ISSN 1613-981X. doi: 10.1007/s10260-021-00573-7.

El Bachir Belhadji, Georges Dionne, and Faouzi Tarkhani. A Model for the Detection of Insurance Fraud. *The Geneva Papers on Risk and Insurance. Issues and Practice*, 25(4): 517–538, 2000. ISSN 1018-5895.

J Bergstra, D Yamins, and D D Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures.

Ll. Bermúdez, J.M. Pérez, M. Ayuso, E. Gómez, and F.J. Vázquez. A bayesian dichotomous model with asymmetric link for fraud in insurance. *Insurance: Mathematics and Economics*, 42(2):779–786, 2008. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2007.08.002. URL `https://www.sciencedirect.com/science/article/pii/S0167668707000947`.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. Performance Measures. In Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera, editors, *Learning from Imbalanced Data Sets*, pages 47–61. Springer International Publishing, Cham, 2018. ISBN 978-3-319-98074-4. doi: 10.1007/978-3-319-98074-4_3.

S. Harjai, S. K. Khatri, and G. Singh. Detecting fraudulent insurance claims using random forests and synthetic minority oversampling technique. In *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, pages 123–128, Mathura, India, 2019. IEEE. doi: 10.1109/ISCON47742.2019.9036162.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6 (4), December 2012. ISSN 1556-4681. doi: 10.1145/2382577.2382579. URL `https://doi.org/10.1145/2382577.2382579`.

Ashley Kilroy. Insurance Fraud Statistics 2024. https://www.forbes.com/advisor/insurance/fraud-statistics/, March 2024.

Toluwalope Owolabi, Essa Q. Shahra, and Shadi Basurra. Auto-Insurance Fraud Detection Using Machine Learning Classification Models. In Xin-She Yang, R. Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Eighth International Congress on Information and Communication Technology*, pages 503–513, Singapore, 2024. Springer Nature. ISBN 978-981-99-3043-2. doi: 10.1007/978-981-99-3043-2_39.

S. Padhi and S. Panigrahi. Use of data mining techniques for data balancing and fraud detection in automobile insurance claims. In *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2019)*, volume 1044, pages 259–268, Singapore, 2020. Springer. doi: 10.1007/978-981-15-1084-7_22. URL `http://link.springer.com/10.1007/978-981-15-1084-7_22`.

D. K. Patel and S. Subudhi. Application of extreme learning machine in detecting auto insurance fraud. In *2019 International Conference on Applied Machine Learning (ICAML)*, pages 78–81, Bhubaneswar, India, 2019. IEEE. doi: 10.1109/ICAML48257.2019.00023.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016a. doi: 10.1109/DSAA.2016.49.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016b. doi: 10.1109/DSAA.2016.49.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Computers in Human Behavior*, 28(3): 1002–1013, May 2012. ISSN 07475632. doi: 10.1016/j.chb.2012.01.002.

M. Salmi and D. Atif. Using a data mining approach to detect automobile insurance fraud. In *Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, pages 55–66, Cham, Switzerland, 2022. Springer. doi: 10.1007/978-3-030-96302-6_5. URL `https://doi.org/10.1007/978-3-030-96302-6_5`.

Gilian Schrijver, Dipti K. Sarmah, and Mohammed El-hajj. Automobile insurance fraud detection using data mining: A systematic literature review. *Intelligent Systems with Applications*, 21:200340, March 2024. ISSN 2667-3053. doi: 10.1016/j.iswa.2024.200340.

Stijn Viaene and Guido Dedene. Insurance Fraud: Issues and Challenges. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 29(2):313–333, April 2004. ISSN 1468-0440. doi: 10.1111/j.1468-0440.2004.00290.x.

R. Wongpanti and S. Vittayakorn. Enhancing auto insurance fraud detection using convolutional neural networks. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 294–301, Phuket, Thailand, 2024. IEEE. doi: 10.1109/JCSSE61278.2024.10613702.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan.