Debener, Jörn; Heinke, Volker; Kriebel, Johannes

Detecting insurance fraud using supervised and unsupervised machine learning

Journal of Risk and Insurance

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Debener, Jörn; Heinke, Volker; Kriebel, Johannes (2023) : Detecting insurance fraud using supervised and unsupervised machine learning, Journal of Risk and Insurance, ISSN 1539-6975, Wiley, Hoboken, NJ, Vol. 90, Iss. 3, pp. 743-768, https://doi.org/10.1111/jori.12427

Received: 2 September 2022 | Revised: 21 January 2023 | Accepted: 3 April 2023

DOI: 10.1111/jori.12427

ORIGINAL ARTICLE

Journal of Risk and Insurance

Detecting insurance fraud using supervised and unsupervised machine learning

Jörn Debener | **Volker Heinke** | **Johannes Kriebel**

Finance Center Münster, University of Münster, Münster, Germany

Correspondence
Johannes Kriebel, Finance Center Münster, University of Münster, Münster, Germany.
Email: johannes.kriebel@wiwi.uni-muenster.de

Abstract

Fraud is a significant issue for insurance companies, generating much interest in machine learning solutions. Although supervised learning for insurance fraud detection has long been a research focus, unsupervised learning has rarely been studied in this context, and there remains insufficient evidence to guide the choice between these branches of machine learning for insurance fraud detection. Accordingly, this study evaluates supervised and unsupervised learning using proprietary insurance claim data. Furthermore, we conduct a field experiment in cooperation with an insurance company to investigate the performance of each approach in terms of identifying new fraudulent claims. We derive several important findings. Unsupervised learning, especially isolation forests, can successfully detect insurance fraud. Supervised learning also performs strongly, despite few labeled fraud cases. Interestingly, unsupervised and supervised learning detect new fraudulent claims based on different input information. Therefore, for implementation, we suggest understanding supervised and unsupervised methods as complements rather than substitutes.

KEYWORDS

insurance fraud detection, machine learning, supervised learning, unsupervised learning

JEL CLASSIFICATION
G22, C40, C52

1 | INTRODUCTION

Fraud is a major issue for insurance companies. Insurance Europe (2019), the European insurance and reinsurance federation, estimates that total fraudulent claims in Europe in 2017 amounted to approximately €13 billion. These massive costs affect not only insurance companies but also honest policyholders via their impact on insurance premiums (Viaene et al., 2007). Beyond the cost savings associated with identifying fraudulent claims, effective detection of insurance fraud also acts as a deterrent that is critical to the insurance market (Picard, 1996; Tennyson & Salsas-Forn, 2002). Therefore, from the perspective of insurance companies and honest policyholders alike, it is important to investigate which method can best detect insurance claim fraud.

Insurance claim fraud is a complex and multifaceted phenomenon, and detecting it is typically very time- and cost-intensive (Viaene et al., 2007). Therefore, an active strand of literature aims to detect insurance claim fraud more efficiently using statistical methods. Until recently, the vast majority of these studies used methods based on supervised learning, such as logistic regressions and artificial neural networks (e.g., Caudill et al., 2005; Viaene et al., 2002; Wang & Xu, 2018). Supervised learning methods use labeled data to learn to distinguish between fraudulent and nonfraudulent claims. Although these methods are generally very efficient when given a sufficient amount of labeled data, they face some challenges in the context of insurance fraud detection, as recently emphasized by Gomes et al. (2021). First, few insurance claims are typically identified as fraudulent due to the substantial effort required to manually detect fraud and fraud being relatively rare. This can make robust estimation a problem. Second, the training data likely contains fraudulent claims that were not detected (see also Brockett et al., 2002), making supervised learning methods prone to replicating existing fraud identification mechanisms. This means new and unknown fraud patterns potentially remain undetected. A remedy could be using methods from the unsupervised learning domain, that is, methods that identify patterns and anomalies in the data without requiring labels. However, thus far, relatively few studies have investigated the potential of unsupervised learning methods for detecting claim fraud. In a very recent example, Gomes et al. (2021) promote a more comprehensive discussion of unsupervised learning by providing important arguments for using unsupervised deep learning in fraud detection and also proposing a corresponding variable importance measure. This emphasizes the need for more evidence to guide the choice between unsupervised and supervised learning.[1]

Thus, consideration of the vivid extant literature surrounding insurance fraud detection reveals three important research gaps. First, there remains limited research on unsupervised learning for insurance fraud detection. Second, modern unsupervised and supervised learning methods have not been directly compared in terms of detecting insurance claim fraud.

[1]Although a comparison is outside the scope of Gomes et al. (2021), the study does compare unsupervised and supervised learning in the context of detecting credit card fraud. However, it is not clear ex ante whether the observations from this comparison are transferable to insurance fraud detection because the patterns and mechanisms of credit card fraud and insurance fraud may differ substantially.

Although the scarcity of available labels could advantage unsupervised learning, findings from credit card fraud detection as another fraud detection task favor supervised learning. Third, given that insurance companies will usually miss some fraud cases based on their existing processes, such a comparison would ideally also be based on an experiment using previously unsuspected cases that are then assessed based on model predictions, a research setting that remains unexplored. From a theoretical perspective, unsupervised learning and supervised learning could both have strengths: When the problem of new, unknown fraud patterns dominates, unsupervised learning could be advantageous; when fraud patterns are not detected comprehensively enough by existing fraud detection mechanisms, supervised learning could offer the benefit of rigorously identifying such claims.

We address these issues using different unsupervised and supervised learning methods to detect insurance fraud. Our main unsupervised learning method is isolation forests, a state-of-the-art unsupervised method that efficiently detects anomalies in data while maintaining relatively low complexity (Hariri et al., 2019; Liu et al., 2008). Our main supervised method is extreme gradient boosting (XGBoost), a modern classifier that performs strongly in many different applications (Shwartz-Ziv & Armon, 2022). We use these methods to analyze a proprietary data set from a German insurance company comprising 7750 automobile insurance claims from 2020 to 2021. After training the machine learning models on the data, we use three steps of analysis to evaluate their insurance fraud detection performance. First, we use regression analysis to investigate the overall discriminatory power of the machine learning-based fraud scores for identifying fraudulent claims. Second, we use precision@k, a measure typically used in the domain of information retrieval (Metzler & Bruce Croft, 2007), to specifically identify what share of the claims with the highest fraud scores are actually fraudulent. Precision@k reflects particularly well the special situation of insurance companies, which usually have only limited resources available for the in-depth investigation of claims (Dionne et al., 2009). Third, we conduct a field experiment to investigate whether the machine learning approaches identify new fraud cases. In the experiment, the insurance company examines those claims assigned a high fraud score by our machine learning methods that the insurance company did not previously suspect to be fraudulent.

The first two steps of our analysis reveal that unsupervised isolation forests have high predictive power for insurance claim fraud. Supervised XGBoost also performs strongly, despite few claims being labeled as fraudulent. Among the claims with the highest fraud scores, both methods feature a substantial share of detected fraudulent claims. Interestingly, the isolation forests outperform unsupervised neural networks based on deep learning and clustering-based anomaly detection. XGBoost outperforms supervised neural networks. Meanwhile, the field experiment reveals that both the unsupervised and supervised learning methods identify previously undetected fraudulent claims. However, the supervised XGBoost outperforms the unsupervised isolation forests. Studying the claims detected by each model in more detail reveals that supervised and unsupervised learning identify different cases to some extent. An analysis of the most important features of each model using SHapley Additive exPlanations (SHAP) provides insight into the mechanics of these machine learning models, showing that both approaches emphasize different input information.

We contribute to the literature in three important ways. First, we extend the sparse literature on unsupervised learning for insurance fraud detection by demonstrating the usefulness of isolation forests, a powerful yet straightforward unsupervised machine learning

method. Beyond Gomes et al. (2021), who highlight the usefulness of unsupervised deep learning for fraud detection in general, modern unsupervised learning has only rarely been studied in this important context. Second, we compare the performance of unsupervised learning and supervised learning methods in terms of detecting insurance fraud using proprietary insurance claim data. While the usefulness of supervised learning for insurance fraud detection has recently been questioned by Gomes et al. (2021), who make a good argument given the few labels that are characteristic for the task, its performance on insurance data has not yet been directly compared with unsupervised learning. Therefore, the literature leaves it unclear whether unsupervised or supervised learning is better-suited performance-wise. Third, we use a field experiment to assess the performance of machine learning models for fraud detection on new claims. Although the strong results for unsupervised learning align with the expectation for unsupervised learning to identify new fraud patterns, supervised learning generates even stronger results, supporting arguments that supervised learning can contribute to the rigorous identification of common fraud patterns, which existing detection mechanisms tend to overlook. Therefore, supervised and unsupervised methods should be considered complements rather than substitutes.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of the existing literature on insurance fraud and its detection using statistical methods. Section 3 introduces the data set. Section 4 presents our methodological approach for detecting insurance fraud using unsupervised and supervised machine learning and introduces our threefold evaluation approach. Section 5 presents the results from our evaluation approach. Section 6 concludes the paper.

2 | LITERATURE REVIEW

2.1 | A theoretical view on insurance fraud

Fraud is an important social issue that has received a lot of attention from academic researchers. It can be defined as "obtaining something of value or avoiding an obligation by means of deception" (Duffield & Grabosky, 2001). In general, fraud results from the interplay of two factors: the motivation of a perpetrator to defraud and an opportunity for the perpetrator to do so (Viaene & Dedene, 2004). The motivation to defraud can be separated into, for example, criminal energy, prestige and greed, and the (economic) situation of the perpetrator (Duffield & Grabosky, 2001). The opportunity to defraud can be considered a combination of a suitable target and the absence of effective protection mechanisms (Cohen & Felson, 1979; Duffield & Grabosky, 2001; Viaene & Dedene, 2004).

Insurance companies are particularly susceptible to fraud, especially in their processing of claims (Viaene & Dedene, 2004). This is due to the asymmetry of the information available to the insurer and the insured. The insured typically have private information, giving them an information advantage over the insurer (Derrig, 2002). This makes insurers a suitable target for motivated perpetrators. Notably, many studies have recognized that claim fraud is a particularly pressing problem in the context of automobile insurance (Picard, 1996; Viaene & Dedene, 2004; Weisberg & Derrig, 1991).

To mitigate this information asymmetry, insurance companies implement audit strategies, according to which they decide whether or not a claim should be investigated. Because investigating whether a claim is legitimate is cost- and time-intensive, this process is a form of

costly state-verification (e.g., Dionne et al., 2009; Picard, 1996; Viaene et al., 2007). The process of claim handling in insurance companies typically starts with an early screening of the claim. This screening determines whether the claim will be processed routinely or handed over to a special investigation unit. The special investigation unit comprises experts who examine the claim in detail before deciding whether the claim should be paid or whether negotiations or legal proceedings should be initiated.[2] However, given the need to process claims as quickly as possible, insurance companies often lack the time and resources required to investigate a lot of cases intensively, which results in the absence of an effective protection mechanism (Viaene et al., 2007). Therefore, insurance fraud is typically regarded as a low-risk, high-reward game (Derrig, 2002; Viaene & Dedene, 2004). This problem also highlights the importance of an efficient early screening that can identify cases with a high likelihood of fraud so that the special investigation unit can allocate its valuable resources to these cases. Our paper addresses this problem by evaluating and comparing the performance of state-of-the-art unsupervised and supervised machine learning methods for identifying fraudulent claims. Using artificial intelligence to more rigorously detect fraud can help deter fraud, benefiting insurers and their honest customers (Tennyson & Salsas-Forn, 2002).

To further disentangle the complex phenomenon of insurance fraud, a typical classification separates soft and hard fraud (Derrig, 2002). Soft fraud, or build-up, reflects situations in which policyholders take advantage of an opportunity (e.g., an unplanned accident) by misrepresenting and inflating claims (Crocker & Morgan, 1998; Picard, 1996; Viaene & Dedene, 2004). Meanwhile, hard fraud describes situations in which policyholders deliberately plan and execute their actions by, for example, staging or inventing accidents (Crocker & Morgan, 1998; Picard, 1996; Viaene & Dedene, 2004). This sort of criminal behavior is characterized by a particularly strong motivation that could result from financial strain or psychological elements, such as perceived power or ego (Duffield & Grabosky, 2001).

2.2 | A methodological view on insurance fraud

Detecting claim fraud in domains such as automobile insurance and healthcare insurance using statistical methods has long been the subject of academic research (e.g., Artís et al., 1999, 2002; Caudill et al., 2005; Johnson & Nagarur, 2016; Riedinger & Major, 2002; Viaene et al., 2007, 2002; Weisberg & Derrig, 1991). Although many studies have used traditional linear approaches, applying machine learning methods to detect insurance claim fraud has become increasingly popular. Current literature argues that this will play an important role in creating more efficient fraud detection mechanisms (Bauer et al., 2021), which could increase the market value of insurance companies (Fritzsch et al., 2021). Machine learning can be defined as a process in which "a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems" (Russell & Norvig, 2020). It generally allows more flexibility in modeling relationships, making it particularly well suited to application in a complex domain, such as insurance fraud detection.

Many studies that use machine learning to detect insurance fraud focus on methods from the domain of supervised learning (e.g., Johnson & Khoshgoftaar, 2019; Liang et al., 2019;

[2]See Viaene et al. (2007) for more details and a graphical representation of this process.

Óskarsdóttir et al., 2022; Sundarkumar et al., 2015; Wang & Xu, 2018). In supervised learning, algorithms learn about the relationship between variables using labeled observations. In contrast, unsupervised learning algorithms learn about patterns in the data without requiring labels (Dixon et al., 2020). In theory, both approaches have advantages and disadvantages with respect to identifying insurance claim fraud. When a lot of labeled data is available (i.e., the insurance company has identified and recorded a lot of fraudulent claims in the past), supervised learning should be able to detect suspicious claims efficiently. However, in the context of insurance claim fraud, there are often only very few detected fraud cases in the data because detecting fraudulent claims requires manual verification and is cost- and time-intensive (Gomes et al., 2021; Viaene et al., 2007), potentially hindering robust model estimations. Furthermore, the data likely contain fraudulent claims that have not been detected as such by the insurance company (Brockett et al., 2002; Gomes et al., 2021). Therefore, supervised learning algorithms might primarily identify fraudulent claims that are similar to previous fraudulent claims, thereby overlooking as-yet undetected fraud. Unsupervised learning addresses these challenges by not requiring labels. However, unsupervised learning has the disadvantage of discovering patterns in the data that are not automatically meaningful in terms of detecting fraud. Therefore, selecting variables that enter unsupervised learning models requires domain knowledge (Stripling et al., 2018).

Surprisingly, unsupervised learning methods have been largely overlooked for a long time in the context of insurance fraud detection. In an early study, Brockett et al. (1998) use Kohonen's self-organizing feature maps, a rather complex unsupervised clustering algorithm, to detect fraudulent automobile insurance claims. Brockett et al. (2002) suggest a principal component analysis of RIDIT scores (PRIDIT) and test this approach again on automobile insurance claims. However, PRIDIT requires strict assumptions regarding predictor variables, namely, a monotonically positive relationship between predictor variables and fraud (Brockett et al., 2002). More recently, Nian et al. (2016) propose unsupervised spectral ranking to detect anomalies in automobile insurance claims and test their approach on data from the years 1994 to 1996. Stripling et al. (2018) use isolation forests—an unsupervised anomaly detection method that will be introduced in more detail in Section 3—to generate features for detecting worker's compensation fraud. Jiang et al. (2021) use isolation forests to tackle the problem of drug reselling as a type of healthcare insurance fraud. Bauder et al. (2018) also focus on healthcare insurance fraud detection and compare the performance of different unsupervised learning methods. More broadly, Gomes et al. (2021) propose an approach to detect fraud in various domains such as insurance claims and credit card payments based on unsupervised deep learning. Their approach further allows identifying the most important variables for this task. Most recently, Vosseler (2022) suggests a Bayesian histogram anomaly detector for fraud detection in general and tests this approach on insurance claim data. Duval et al. (2023) further use unsupervised learning, including isolation forests, to derive anomaly profiles of driving behavior and identified a predictive relationship to the probability of automobile insurance claims. Tumminello et al. (2023) develop filter rules to identify the criminal infrastructures of fraudsters in extensive networks.

3 | DATA

To investigate how state-of-the-art unsupervised and supervised learning models detect insurance claim fraud, we use a proprietary data set obtained from a German insurance company. The data include 7750 automobile insurance claims placed between January 2020 and April 2021. The claims concern damages to the policyholders' cars resulting from collisions

with objects (e.g., road signs and parking garages) other than cars.[3] To fit the supervised and unsupervised machine learning models, we use a training sample that contains 60% of the data. Then, we use the test sample containing the remaining 40% of the data to evaluate the usefulness of the fraud detection approaches.

Studying prediction methods on proprietary insurance claim data is very useful. Publicly available data on insurance claim fraud is rare. Empirical studies on insurance fraud detection often refer to the same data sets. For example, Artís et al. (2002), Caudill et al. (2005), and Ai et al. (2013) analyze a small Spanish data set comprising 1995 claims, and Viaene et al. (2002) and Ai et al. (2013) analyze a small US data set featuring 1399 claims. Therefore, our data set offers a particularly interesting new perspective on investigating the potential of using machine learning methods for insurance claim fraud detection. Summary statistics of the variables in our data set that are used as machine learning features and the dependent variables that are used for performance evaluation are presented in Table 1.

The data set contains the following variables: The policyholder's *Claimed amount*, the *Age of the car* at the time of the claim, the *Power of the engine* of the car, the *Mileage of the car* at the time of the claim, the monthly *Premium for the car* paid by the policyholder for insuring the car, the *Dunning level* of the policyholder (ranging from 0 for no dunning to 1 for payment reminder to 3 for dunning), whether the insurer has identified potential *Misconduct* by the policyholder (where 1 indicates misconduct and 0 indicates no misconduct), the *No-claim class* of the policyholder (ranging from −1 for the lowest no-claim class, which indicates not much driving experience and/or recently filed claims, to 50, the highest no-claim class, which indicates no previously filed claims over a long period), the *Deductible* of the policyholder, the *Age of the contract* between the insurance company and the policyholder at the time of the claim, and whether there is a *Fraud record entry* for the policyholder at the time of the claim (ranging from 1 for no entry to 2–6 for different kinds of entries).[4]

To evaluate the machine learning models, we use two dummy variables from the data set that address two different stages of the aforementioned fraud detection process used by insurance companies (see Section 2.1). First, we use a variable indicating whether the early screening by a fraud coordinator has evaluated the claim as highly suspicious and the insurance company's special investigation unit has then investigated the claim (*susp*). Second, we use a variable indicating whether the special investigation unit has successfully shown that the claim was fraudulent (*fraud*). The reason for using *susp* as a dependent variable in addition to *fraud* is that insurance companies often lack the resources to prove that a claim is fraudulent or even decide that the potential benefit of proving the claim fraudulent will not exceed the legal costs. These legal and economic considerations affect the labeling of claims, making the inclusion of highly suspicious claims in the validation analysis common in studies that investigate the potential for statistical methods to detect insurance claim fraud (Stripling et al., 2018).

[3]In a robustness check, we later also present analyses for two additional data sets from the same insurance company. The first additional data set contains claims related to damages resulting from collisions with other cars, and the second additional data set contains claims related to glass damage. Because the fraud patterns likely differ between these claim types, we investigate the data sets separately. In our main analysis, we focus on claims from collisions between cars and objects because the insurance company has indicated the particular potential for fraud detection within this claim type and the desire to conduct the field experiment with this claim type.

[4]Unavailable variable values have been replaced with high negative values to allow for imputation to avoid missing any data. However, the machine learning methods can distinguish that values were not provided in the data initially in the case that this information is important.

TABLE 1 Summary statistics.

Variable	*N*	Mean	SD	Min	Median	Max
Machine learning features						
Claimed amount (in EUR)	6603	3129.806	2844.511	1	2700	54,000
Age of the car (in days)	7733	1643.940	1418.302	4	1190	14,154
Mileage of the car (in km)	7750	48,027.570	59,996.390	0	26,000	569,202
Power of the engine (in kW)	7679	123.991	67.117	10	110	2240
Premium for the car (in EUR)	7720	278.843	517.400	0.083	93.667	11,634
Age of the contract (in days)	7653	2907.401	3380.896	0	1776	18,221
Dunning level (in categories)	7750	0.498	0.832	0	0	3
No-claim class (in categories)	7683	14.284	12.322	−1	10	50
Deductible (in EUR)	7352	219.189	102.848	0	225	1325
Misconduct (in categories)	7750	0.630	0.483	0	1	1
Fraud record entry (in categories)	7750	1.064	0.437	1	1	6
Dependent variables						
susp (in categories)	7750	0.021	0.143	0	0	1
fraud (in categories)	7750	0.004	0.066	0	0	1

Note: This table reports summary statistics for the input variables used in the machine learning models and the variables used for evaluating the machine learning models. "*N*" denotes the number of nonmissing values, "Mean" the mean, "SD" the standard deviation, "Min" the minimum, "Median" the median, and "Max" the maximum.

In our data, 162 claims were marked as highly suspicious (training data, 103; test data, 59) and 34 were proven cases of fraud (training data, 21; test data, 13). The number of actual fraudulent cases (including unidentified cases) may be much higher. These low numbers reflect how difficult it is for insurance companies to identify claim fraud and indicate the potential challenges associated with training supervised learning algorithms to detect insurance claim fraud.

4 | METHODOLOGY

4.1 | Unsupervised fraud detection

Our analysis uses isolation forests as an unsupervised learning method for insurance fraud detection. Isolation forests, especially extended isolation forests, are a novel tree-based ensemble method that aims to detect anomalies in data (Hariri et al., 2019; Liu et al., 2008). Compared with isolation forests, other methods that aim to detect anomalies typically define a normal observation before identifying anomalies as anything that deviates from that normality (Chandola et al., 2009). The problem with this approach is that the methods used are optimized to detect normal observations rather than anomalies, which makes them less efficient in their actual intent (Liu et al., 2008). In contrast, isolation forests are designed to directly identify anomalies, even using higher-dimensional data (including textual data). Thus, the concept is

built on an explicit definition of anomalies. According to Liu et al. (2008), anomalies have two key characteristics: namely, they are few and different. "Few" means that anomalies are the minority in the data. "Different" means that anomalies typically differ substantially in appearance from the rest of the observations.

Isolation forests incorporate these two characteristics to describe an ensemble of specific decision trees (so-called isolation trees) in which every tree hosts an iterative process (see Liu et al., 2008 for a more detailed description). The first step in this process involves the random selection of a characteristic of the data. In the case of automobile insurance claim fraud, this could be, for example, the mileage of the car at the time of the accident. Next, a value between the maximum and the minimum for the characteristic in the sample is randomly selected. According to this random value, all observations are split. This process is repeated until all observations are separated (or the predefined maximum number of splits is reached). The more anomalous the case, the faster the isolation. Thus, the length of the isolation process is used as an anomaly score. This process is depicted in Figure 1.

As proposed by Hariri et al. (2019), extended isolation forests offer the novelty of also using hyperplanes, which are not orthogonal with regard to a single characteristic. When using the example of two characteristics, as plotted in Figure 1, the extended isolation forest would mostly include lines with varying slopes for separation (in addition to horizontal and vertical lines). This makes the extended isolation forest more flexible in terms of separating anomalies from normal cases. Where the traditional isolation forest proposed by Liu et al. (2008) featured more edgy separations between normal and anomalous areas, the extended isolation forest separates these areas more smoothly, enabling it to more powerfully identify anomalies. This work uses the terms extended isolation forests and isolation forests to refer to extended isolation forests.

Developing unsupervised machine learning methods for fraud detection is implicitly based on the assumption that anomalous cases are more likely to be fraudulent. However, in the context of insurance claim fraud, anomalous cases may not automatically be informative of fraud. As such, unsupervised learning approaches to fraud detection need to incorporate characteristics that have meaningful anomalies. Consequently, developing unsupervised

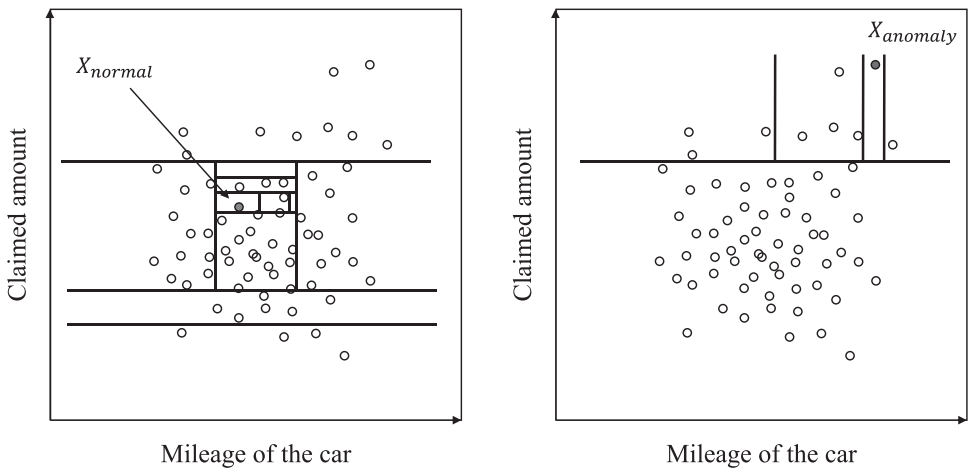

FIGURE 1 Visualization of an isolation forest. This figure displays how isolation forests identify anomalies in data sets based on a two-dimensional example.

prediction models must be informed by domain knowledge that enables the developed model to identify meaningful anomalies (Stripling et al., 2018). The models that we develop are based on these considerations. Additionally, to produce a single anomaly score, the mean of the two model outcomes is used in the analyses.

The first isolation forest aims to detect whether the combination of the claimed amount and the characteristics of the car is anomalous. This specifically addresses cases of inflated reported damage (Picard, 1996). We include as important input variables the *Claimed amount* together with variables that relate to the true financial size of the claim to identify inflated claims. Because the financial size of the claim depends on the value of the car, we include as further variables the *Age of the car*, *Mileage of the car*, and *Power of the engine* as well as the insurance *Premium for the car* paid by the policyholder, all of which closely relate to the car's value (Lessmann et al., 2010).

The second isolation forest aims to identify claims where the policyholder has an anomalously high economic motivation for fraud or demonstrates signs of criminal behavior. For this model, we include the following variables: *Age of the contract* as an indicator of intent because a claim soon after a contract's initiation potentially generates positive net income for the policyholder (Grabosky & Duffield, 2001); the *Dunning level* as a measure of the financial strain of the policyholder, which is also a common driver of criminal behavior (Duffield & Grabosky, 2001); the *No-claim class* and the *Deductible* as further indicators of economic incentive for the policyholder to defraud; and, as indicators of criminal energy, whether there was potential *Misconduct* (e.g., the driver was not allowed to drive the car under the insurance contract) and a *Fraud record entry* reported for the policyholder. Figure 2 presents both models and their associated variables. To find even more meaningful anomalies, we use the lower half of the age of the contract, deductibles, and the no-claim class by winsorizing these values at 50% to set higher values to the 50% quantile. Accordingly, the model focuses on newly entered contracts, low no-claim classes, and low deductibles, which reflect circumstances with higher economic motivation for fraud.

4.2 | Supervised fraud detection

To compare the performance of unsupervised learning with that of supervised learning, we implement XGBoost. XGBoost is a tree-based method (like the isolation forest), but it is designed to be trained on labeled data. XGBoost reflects the idea of gradient-boosted

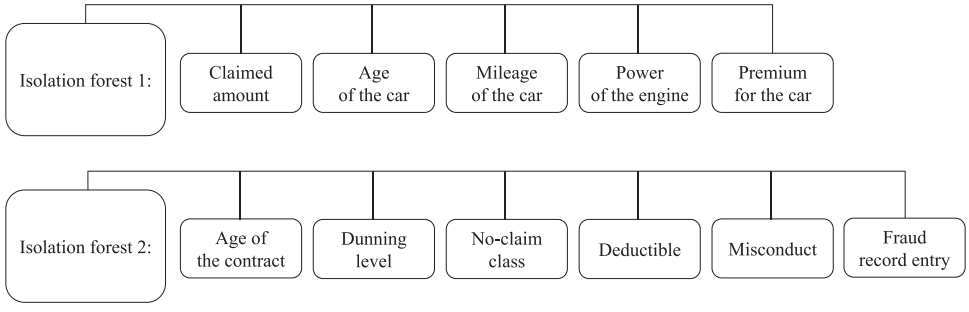

FIGURE 2 Overview of isolation forest models. This figure displays the variables used in the two isolation forest models developed in this study.

decision trees with a focus on speed and performance. XGBoost has recently been shown to perform favorably on similar classification tasks, such as default prediction (Gunnarsson et al., 2021). XGBoost has also recently been observed to perform strongly on tabular data, even compared with state-of-the-art deep learning models (Shwartz-Ziv & Armon, 2022).

The individual trees of the XGBoost are based on the idea of classification and regression trees proposed by Breiman et al. (1984), a machine learning algorithm that iteratively splits data into smaller subsets based on characteristics in the data during training. In the case of automobile insurance claim fraud, an example characteristic might be the mileage of a car. The algorithm begins with a single node that contains all observations. The classification trees (such as trees for identifying insurance claim fraud) select characteristics that produce the largest reduction in some measure of purity in the nodes of the tree aimed at producing nodes that ideally comprise only one class. The selected characteristic is used to split the observations into two groups. Then, the process is repeated recursively for the two new nodes until arriving at a stopping rule. The prediction of the method is derived by the node the observation is finally moved into. In classification problems, this is given by the class frequency across the observations within the node.

Similar to methods such as bagging (Breiman, 1996) and random forests (Breiman, 2001), XGBoost is also based on the idea of ensemble learning. Ensemble learning is a technique for creating multiple models and combining them to produce a final model that is more accurate than the individual models. The ensemble is created by training multiple models on subsamples of the data and then combining the predictions of the individual models. Ensemble learning is particularly common for tree-based methods as it can reduce overfitting and smoothen predictions over the space of characteristics. As suggested in Chen and Guestrin (2016), XGBoost also represents an ensemble of trees. The algorithm sequentially builds trees that address specific errors of trees earlier in the line. The algorithm offers further regularization to ultimately perform strongly. Meanwhile, the XGBoost features several tunable hyperparameters. We tune as hyperparameters the maximum depth of the trees, the subsample size used for each tree, the learning rate, the number of characteristics used by each tree, the regularization, and the number of trees.[5]

Regarding the task at hand, we use all input variables for the XGBoost. We build two XGBoost models. The first is trained to detect highly suspicious claims. The second is trained to detect proven fraudulent claims.[6]

4.3 | Fraud scores

Table 2 displays summary statistics on the fraud scores obtained from the machine learning models described in Sections 4.1 and 4.2. $if.score_1$, $if.score_2$, and $mean\_if.score$, respectively, represent the fraud scores from isolation forest 1, isolation forest 2, and the aggregate isolation forest model. Values range from 0 to 1, with values exceeding 0.5 indicating more anomalous claims and higher values generally indicating stronger anomalies.[7] $xgb.score_{susp}$ and

[5]Before training the supervised models, we randomly separate a validation sample comprising 33% of the observations in the training sample. We use this validation sample to find the hyperparameters (e.g., the maximum depth of trees for the XGBoost model) that best allow the supervised models to detect insurance fraud.
[6]The first model is trained using *susp* as the dependent variable. The second model is trained using *fraud* as the dependent variable.

TABLE 2 Summary statistics—Fraud scores.

Variable	*N*	Mean	SD	Min	Median	Max
Fraud scores						
$if.score_1$	7750	0.428	0.066	0.356	0.406	0.730
$if.score_2$	7750	0.470	0.060	0.369	0.472	0.657
mean_if.score	7750	0.449	0.045	0.363	0.444	0.672
$xgb.score_{susp}$	7750	0.024	0.037	0.007	0.014	0.723
$xgb.score_{fraud}$	7750	0.004	0.026	0.001	0.001	0.758

Note: This table reports summary statistics for the fraud scores resulting from the machine learning models. "*N*" denotes the number of nonmissing values, "Mean" the mean, "SD" the standard deviation, "Min" the minimum, "Median" the median, and "Max" the maximum.

$xgb.score_{fraud}$ represent the fraud scores from the first XGBoost, trained on *susp* as the dependent variable, and the second XGBoost, trained on *fraud* as the dependent variable. Values range from 0 to 1, and higher values indicate a higher likelihood of fraud.

4.4 | Evaluation approach

4.4.1 | Regression analysis

Our threefold procedure for evaluating the performance of the methods is described in the following paragraphs. The first evaluation step assesses the informativeness of the fraud scores for discriminating between fraudulent and nonfraudulent claims in the overall test data. For this, we use a logistic regression of the derived fraud scores to assess their correlation with the presence of fraud cases. We consider logistic regression models that employ two previously introduced dependent variables: *susp*, indicating whether claims are considered as highly suspicious of fraudulence and *fraud*, indicating whether claims are proven fraudulent.

The model that uses *susp* as a dependent variable identifies whether the fraud detection score $ml.score_i$, derived from a machine learning model for claim i, provides insight into whether domain experts are more likely to decide to send the claim to the special investigation unit, as indicated in Equation (1). The domain experts' decision is based on two criteria: their own consideration of the claim and a fraud score derived from the insurance company's expert system.

$$P(susp_i = 1|ml.score_i) = \frac{exp(\beta_0 + \beta_1 ml.score_i)}{1 + exp(\beta_0 + \beta_1 ml.score_i)}. \quad (1)$$

Next, the second regression model regresses a dummy indicating whether a claim was later proven fraudulent by the special investigation unit, *fraud*, as shown in Equation (2). In this

[7]Anomaly scores from the isolation forests are a function of the number of splits necessary to isolate a case. Values around 0.5 represent the length of the isolation process that is expected for a normal observation in the data.

analysis, the coefficients for the machine learning-based fraud scores should be positive and significant when they contain information pertinent to detecting insurance fraud.

$$P(fraud_i = 1|ml.score_i) = \frac{exp(\beta_0 + \beta_1 ml.score_i)}{1 + exp(\beta_0 + \beta_1 ml.score_i)}. \quad (2)$$

4.4.2 | Precision@k

Although the regression analysis in the first evaluation step assesses the general informativeness of fraud scores for identifying claim fraud, the processes for identifying claim fraud that is embedded in insurance operations can naturally only examine a limited number of potentially fraudulent claims. To further evaluate the machine learning methods for fraud detection, we next conduct an analysis using precision@k, a measure of prediction quality designed specifically for such situations and typically employed in the domain of information retrieval (Metzler & Bruce Croft, 2007). The measure evaluates the quality of an algorithm by aggregating how many of the k first cases ranked highest were actually positive, that is, in our study context, how many claims were highly suspicious or proven fraudulent. Accordingly, the measure reflects how many proven fraudulent claims will be identified by including machine learning-based fraud scores in the fraud detection process and investigating the top-ranked claims.

We evaluate the fraud scores again using the two dependent variables *susp* and *fraud*. For the k used in the precision@k measure, we use 100, 200, and 500, meaning that 100, 200, and 500 claims would hypothetically be delivered to the special investigation unit. To assess the statistical significance of the individual precision@k values, we utilize the fact that the number of successfully identified claims within k draws is binomially distributed for a given proportion of highly suspicious and actual fraudulent claims in the test sample. As such, a significance level can be specified for the likelihood that the determined number of correctly identified claims occurred by chance.

4.4.3 | Field experiment

The first two evaluation steps are generally appropriate for determining how well the machine learning-based fraud scores can identify previously detected fraudulent claims. However, the nature of fraud makes it most likely that not all fraudulent claims will be identified by the existing fraud detection process. This makes it particularly interesting to consider the performance of machine learning methods in the context of revealing fraudulent claims not yet identified as such. This crucial assessment of the quality of predictions is obviously not possible using only observational data. Accordingly, as a third evaluation step, we conduct a field experiment in collaboration with the insurance company responsible for the study data. This also represents a unique opportunity to explore the success of the machine learning methods for fraud detection in a realistic operational environment.

The design of the field experiment is such that we are able to select the 100 claims from our existing data set that are neither proven fraudulent nor deemed highly suspicious at the time of data provision but which receive the highest fraud scores from the machine learning models.

Meanwhile, to compare the supervised and unsupervised learning approaches, we submit 50 claims from each type of machine learning method (isolation forests as the unsupervised method and XGBoost as the supervised method) to the insurance company, which does not know which claims are identified by the unsupervised method and which claims are identified by the supervised method. The 100 claims are presented in a randomly sampled order for examination by the insurance company's special investigation unit. The special investigation unit is asked to evaluate whether the claims are highly suspicious and demand further examination based on the communication that this assessment is designed to be consistent with the classification of a highly suspicious claim (*susp*) used in the previous steps of the validation analysis. The field experiment is conducted with a 9-month lag to data provision. The number of claims correctly identified as highly suspicious is used here to measure the quality of the machine learning methods.

It is ex ante not clear whether supervised or unsupervised learning is superior in this setting. When there are many new and unknown fraud patterns present, such claims are likely best detected by unsupervised learning because unsupervised learning is not trained on labels produced by existing fraud detection mechanisms. However, when the fraud patterns present are predominantly not entirely new but have been overlooked by existing fraud detection mechanisms, supervised learning is likely superior due to having been trained to identify such claims rigorously.

4.5 | Benchmark methods

To check whether our results are unique to isolation forests and XGBoost, we compare the results of these methods to results from benchmark machine learning methods. For this comparison, we use state-of-the-art autoencoders and more traditional *k*-medoids clustering as alternative unsupervised machine learning methods and feed-forward artificial neural networks as an alternative supervised machine learning method.

Autoencoders are based on deep learning. While training, they encode input variables—meaning they reduce their dimensionality—before decoding the encoded variables again to reproduce the inputs.[8] This can be used to detect anomalies. Observations that differ strongly from the decoded output of the autoencoder are anomalous, which can be considered informative of fraud. The autoencoders used for comparison contain three hidden layers of which the first hidden layer contains 15 neurons, the second hidden layer with a lower dimensionality contains 10 neurons, and the third hidden layer contains 15 neurons. The hidden layers have a hyperbolic tangent activation function and the output layer has a linear activation function. The autoencoders are trained using an Adam optimizer and a mean squared error loss. The usefulness of autoencoders for fraud detection has recently been highlighted by Gomes et al. (2021). For comparison with the isolation forests, we train two distinct autoencoder models to obtain fraud scores before combining these scores using the mean (*mean_ae.score*).

We further apply *k*-medoids clustering—a widely used clustering algorithm introduced by Kaufman and Rousseeuw (1990). We use *k*-medoids clustering in combination with the Gower distance (Gower, 1971), which can deal with categorical variables. As common in anomaly

[8]For a detailed discussion of autoencoders, see, for example, Baldi (2012).

detection studies using clustering, fraud scores are derived as the distance of a claim to the center of its nearest cluster (Chandola et al., 2009). The optimal number of clusters for each model, which is two in our case, is selected based on the Silhouette coefficient suggested by Rousseeuw (1987). Again, we train two distinct models and combine the obtained fraud scores using the mean (*mean_km.score*).

With respect to the benchmark method for supervised learning, we use artificial neural networks. Neural networks comprise layers of neurons, with the weights determined based on the training data to fit the binary classification of a claim being or not being highly suspicious of fraud ($ann.score_{susp}$) or fraudulent ($ann.score_{fraud}$). The networks have the number of layers, the number of neurons, the dropout rate, the extent of l1l2-regularization, and the batch size as hyperparameters. The artificial neural networks use rectified linear unit activation functions in the hidden layers and a sigmoid activation function in the output layer. The artificial neural networks are trained using an Adam optimizer and a binary cross entropy loss.

5 | RESULTS

5.1 | Regression approach

This section presents the empirical results evaluating the performance of the machine learning methods. The results of our first evaluation step, in which we investigate the informativeness of the machine learning-based fraud scores for detecting automobile insurance claim fraud, appear in Table 3.

Where highly suspicious claims represent the dependent variable (upper panel in Table 3), columns one and two reveal that the fraud scores from both isolation forests ($if.score_1$ and $if.score_2$) are strongly significant predictors of highly suspicious claims. Combining the two isolation-forest-based fraud scores into one score (*mean_if.score*) increases the McFadden pseudo-R^2 compared with using only one of the two fraud scores (column three). Turning to the fraud score produced by an XGBoost model trained to identify highly suspicious claims ($xgb.score_{susp}$), this supervised learning approach also generates fraud scores that are strongly significant predictors (column four).

When using the proven fraudulent claim as the dependent variable (lower panel in Table 3), $if.score_1$ and *mean_if.score* maintain their level of statistical significance (columns one and three). $if.score_2$ remains a significant predictor, but its significance drops to the 10% level (column two). Interestingly, the same holds for $xgb.score_{fraud}$, the fraud score produced by an XGBoost trained on the variable *fraud* (column four).[9]

Overall, the regression analysis identifies that fraud scores generated from both unsupervised and supervised machine learning models are predictors of insurance fraud, meaning that higher fraud scores are typically associated with a higher likelihood of fraud. These results are supported when examining two other classes of automobile insurance claims, namely, claims associated with collisions with other cars and claims associated with glass damage. The results for these claim types appear in the Supporting Information.

[9]We note that the logistic regression makes the implicit assumption that the logarithm of the odds for being a fraudulent claim is a univariate linear function of the scores. This assumption could possibly be more beneficial for the isolation forest compared with the XGBoost. To compare the relation of their scores with the fraud variables, we calculated nonparametric Spearman and Kendall correlations. The relation to the occurrence of proven fraudulent claims is similar for both methods. The relation to the occurrence of highly suspicious claims is stronger for the XGBoost than for the isolation forest. We refrain from further analyzing the comparative strength of the methods in this step and refer the reader to the results using the precision@k and the field experiment for comparison.

TABLE 3 Performance evaluation: Regression analysis.

	Highly suspicious claim (*susp*)			
	(1)	**(2)**	**(3)**	**(4)**
$if.score_1$	8.894***			
	(1.401)			
$if.score_2$		13.951***		
		(2.407)		
mean_if.score			18.721***	
			(2.213)	
$xgb.score_{susp}$				15.002***
				(1.806)
Constant	−7.959***	−10.826***	−12.753***	−4.474***
	(0.688)	(1.247)	(1.105)	(0.165)
Observations	3099	3099	3099	3099
McFadden pseudo-R^2	0.051	0.056	0.101	0.083
	Proven fraudulent claim (*fraud*)			
	(1)	**(2)**	**(3)**	**(4)**
$if.score_1$	11.935***			
	(2.704)			
$if.score_2$		9.220*		
		(4.836)		
mean_if.score			19.948***	
			(4.221)	
$xgb.score_{fraud}$				6.469*
				(3.433)
Constant	−11.026***	−9.955***	−14.942***	−5.515***
	(1.424)	(2.444)	(2.164)	(0.283)
Observations	3099	3099	3099	3099
McFadden pseudo-R^2	0.096	0.022	0.108	−0.014

Note: This table displays the results for running a logistic regression of *susp* and *fraud*, respectively, on $if.score_1$, $if.score_2$, *mean_if.score*, $xgb.score_{susp}$, and $xgb.score_{fraud}$ for claims resulting from collisions. The table reports standard errors in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The results for the benchmark methods appear in Table 4. The autoencoder produces fraud scores that are significant predictors of highly suspicious claims (*mean_ae.score*), but the predictive power for proven fraudulent claims is not significant anymore. Autoencoders likely perform more weakly on this task because the large amount of data that autoencoders typically require may not be available in contexts such as that of our analysis. This problem has also been discussed by Gomes et al. (2021). *k*-Medoids clustering, as a second and more traditional unsupervised learning method, produces

fraud scores that are significant predictors of highly suspicious and proven fraudulent claims (*mean_km.score*). In the case of supervised neural networks, their fraud scores are significant predictors of highly suspicious claims ($ann.score_{susp}$) but, similarly to autoencoders, not significant predictors of proven fraudulent claims ($ann.score_{fraud}$).

5.2 | Precision@k approach

While the general informativeness of machine learning-based fraud scores is a good indication of their usefulness for detecting claim fraud, insurance companies typically face a situation in which they only have the resources to investigate a few claims, which are naturally those that

TABLE 4 Performance evaluation: Regression analysis (benchmark methods).

	Highly suspicious claim (*susp*)		
	(1)	**(2)**	**(3)**
mean_ae.score	4.422***		
	(1.140)		
mean_km.score		10.781***	
		(1.270)	
$ann.score_{susp}$			90.207***
			(10.953)
Constant	−5.195***	−8.529***	−4.433***
	(0.368)	(0.626)	(0.180)
Observations	3099	3099	3099
McFadden pseudo-R^2	0.023	0.124	0.108
	Proven fraudulent claim (*fraud*)		
	(1)	**(2)**	**(3)**
mean_ae.score	3.031		
	(2.214)		
mean_km.score		10.290***	
		(2.552)	
$ann.score_{fraud}$			42.119
			(65.845)
Constant	−6.313***	−9.855***	−11.909
	(0.710)	(1.272)	(10.094)
Observations	3099	3099	3099
McFadden pseudo-R^2	0.008	0.073	−0.021

Note: This table displays the results for running a logistic regression of *susp* and *fraud*, respectively, on *mean_ae.score*, *mean_km.score*, $ann.score_{susp}$, and $ann.score_{fraud}$ for claims resulting from collisions. The table reports standard errors in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

are most likely to be fraudulent. The savings from identifying fraud should ideally exceed the cost necessary to investigate the claims. Therefore, assessing the number of fraudulent claims among those claims that receive the highest fraud scores from machine learning-based methods is particularly interesting for real operations. The precision@k-derived results that reflect these considerations appear in Table 5.

We present results for three different variants of k (100, 200, and 500 in columns one, two, and three) and for the two variables of interest, *susp* (highly suspicious claims) and *fraud* (proven fraudulent claims). Randomly drawing and investigating 100 claims would have produced two highly suspicious claims. However, investigating the top 100 claims according to the isolation-forests-based fraud scores reveals 15 claims labeled highly suspicious (upper panel in Table 5). This number (15 claims) is highly statistically significant. Furthermore, increasing the number of identified highly suspicious claims in comparison to randomly drawing claims by factor 7.5 (for *mean_if.score*) is also economically significant. Investigating the top 100 XGBoost claims reveals 16 highly suspicious claims. For $k = 200$ and 500, $xgb.score_{susp}$ is superior to *mean_if.score* by two and six claims. Therefore, the supervised model slightly outperforms the combined isolation forest, despite the small number of claims labeled as suspicious in the training data. For proven fraudulent claims (lower panel in Table 5), randomly drawing and investigating 100 claims would have resulted in an average number of proven fraudulent claims below one. When investigating the top 100 claims, the isolation forest-based fraud score *mean_if.score* is on par with $xgb.score_{fraud}$. This is similar for $k = 200$ and 500. However, $if.score_2$ only detects a significant amount of proven fraudulent claims for $k = 500$.

TABLE 5 Performance evaluation: Precision@k.

Highly suspicious claim (*susp*)			
$if.score_1$	7***	15***	26***
$if.score_2$	8***	11***	25***
mean_if.score	15***	23***	33***
$xgb.score_{susp}$	16***	25***	39***
Random selection	2	4	10
k	100	200	500
Proven fraudulent claim (*fraud*)			
$if.score_1$	3***	4**	8***
$if.score_2$	0	1	5*
mean_if.score	3***	5***	8***
$xgb.score_{fraud}$	3***	5***	7***
Random selection	0	1	2
k	100	200	500

Note: This table displays the precision@k results for $k = 100$, 200, and 500 using *susp* and *fraud* as target variables, respectively, and $if.score_1$, $if.score_2$, *mean_if.score*, $xgb.score_{susp}$, and $xgb.score_{fraud}$ as ranking variables for claims resulting from collisions. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Overall, these results demonstrate that the upper tail of the distributions of fraud scores generated by both unsupervised and supervised learning methods contain important information for identifying fraudulent claims. Analyzing the other claim types presented in the Supporting Information reveals that these results are largely qualitatively unchanged, with XGBoost performing particularly strongly for claims from glass damage.

Turning to the benchmark methods (Table 6), the autoencoder-based fraud score is very helpful for detecting highly suspicious claims but less helpful for detecting proven fraudulent claims. Although the results for identifying highly suspicious claims are nearly on par with the results produced by isolation forests, substantially worse results are observed in the context of identifying proven fraudulent claims. Again, this is likely due to the large amount of data that autoencoders require. *k*-Medoids clustering provides slightly stronger results than autoencoders for both variables of interest but still falls behind the isolation forest. With respect to supervised learning, XGBoost performs slightly better than the supervised neural network-based model.

While precision@k represents a particularly appropriate evaluation measure in the context of insurance claim fraud, it is worth also considering other performance measures that are typically used to assess classification methods. These results appear in the Supporting Information. Here, isolation forests and XGBoost again show strong results, with a slight advantage observed for XGBoost.

The analysis then also investigates how a combination of the isolation forest and the XGBoost performs. For this step, we calculate the mean rank of *mean_if.score* and $xgb.score_{susp}$ ($if\_xgb.score_{susp}$) and $xgb.score_{fraud}$ ($if\_xgb.score_{fraud}$), respectively. The precision@k-results are presented in Table 7. Most interestingly, the new measure exceeds the quality of predictions of the individual models in Table 5 in almost all cases. Regarding the detection of highly suspicious cases, the combined measure detects more highly suspicious cases than the isolation forest for all three levels of k ($17 > 15$, $25 > 23$, and $40 > 33$). The combined measure detects more cases than the XGBoost for k equal to 100 and 500 ($17 > 16$ and $40 > 39$) and is on the

TABLE 6 Performance evaluation: Precision@k (benchmark methods).

Highly suspicious claim (*susp*)			
mean_ae.score	13***	17***	32***
mean_km.score	12***	21***	35***
$ann.score_{susp}$	14***	17***	32***
Random selection	2	4	10
k	100	200	500
Proven fraudulent claim (*fraud*)			
mean_ae.score	2*	2	5*
mean_km.score	1	3*	8***
$ann.score_{fraud}$	3***	3*	4
Random selection	0	1	2
k	100	200	500

Note: This table displays the precision@k results for $k = 100$, 200, and 500 using *susp* and *fraud* as target variables, respectively, and *mean _ae.score*, *mean_km.score*, $ann.score_{susp}$, and $ann.score_{fraud}$ as ranking variables for claims resulting from collisions. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

TABLE 7 Performance evaluation: Precision@k (combined measures).

Highly suspicious claim (*susp*)			
if _xgb.score$_{susp}$	17***	25***	40***
Random selection	2	4	10
k	100	200	500
Proven fraudulent claim (*fraud*)			
if _xgb.score$_{fraud}$	4***	6***	10***
Random selection	0	1	2
k	100	200	500

Note: This table displays the precision@k results for $k = 100$, 200, and 500 using *susp* and *fraud* as target variables, respectively, and the mean rank of *mean_if.score* and *xgb.score*$_{susp}$ (*if _xgb.score*$_{susp}$) and *xgb.score*$_{fraud}$ (*if _xgb.score*$_{susp}$), respectively, as ranking variables for claims resulting from collisions. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

same level for a *k* of 200 (25 cases). Regarding proven fraud cases, the combined measure consistently outperforms the individual fraud scores (isolation forest: $4 > 3$, $6 > 5$, and $10 > 8$; XGBoost: $4 > 3$, $6 > 5$, and $10 > 7$). These results suggest to use the scores produced from supervised learning and unsupervised learning in combination. This idea will be further investigated in Section 5.3.

5.3 | Field experiment

The first two steps of our performance evaluation approach have usefully demonstrated that both machine learning methods (i.e., unsupervised isolation forests and supervised XGBoost) can detect fraudulent claims. However, the previous analyses likely underestimate the true number of fraudulent claims that can be detected using machine learning methods because the observational data available (and, in fact, any observational data on insurance claims) likely contains fraudulent claims undetected as such by the insurance company. Therefore, we now provide the field experiment results, in which the insurance company examines the 100 claims from collisions with objects that have been assigned the highest fraud scores by an unsupervised isolation forest model and a supervised XGBoost model trained on *susp*. The experiment uses the 50 claims with the highest isolation forest scores and the 50 claims with the highest XGBoost scores, considering only claims not previously identified as highly suspicious by the insurance company. These claims are then evaluated by the special investigation unit. The results from this field experiment are summarized in Table 8 and Figure 3.

The results from our field experiment show that both machine learning models have been able to detect a high number of claims not previously identified as highly suspicious but evaluated as highly suspicious during the field experiment by the special investigation unit. The numbers are quite strong. Among 50 claims identified by the isolation forest model as anomalous, 22 claims have been assessed as highly suspicious of fraud. Interestingly, XGBoost demonstrates an even better performance: among 50 claims identified by the XGBoost approach as anomalous, 29 claims have been evaluated as highly suspicious of fraud. Thus, a

TABLE 8 Performance evaluation: Field experiment.

	Overall	Isolation forest	XGBoost
Submitted claims	100	50	50
Highly suspicious claims	51	22	29
Unique highly suspicious claims	–	6	13

Note: This table displays the results from the field experiment. Fifty claims that were initially labeled as not highly suspicious but judged as highly suspicious by the isolation forest and 50 claims that were initially labeled as not highly suspicious but judged as highly suspicious by the XGBoost were submitted to the special investigation unit of the insurance company to assess whether these claims are indeed highly suspicious of fraud.

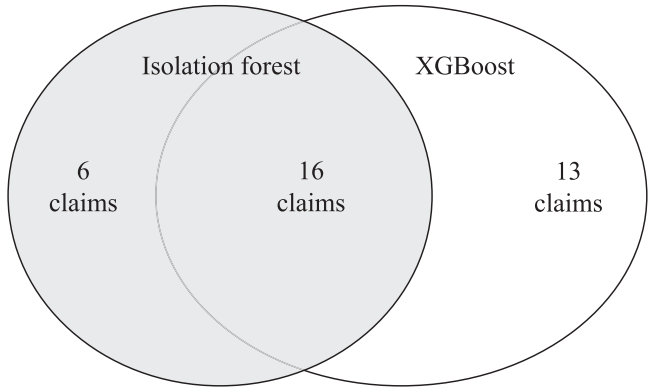

FIGURE 3 Venn diagram field experiment. This figure displays the results from the field experiment, that is, the new claims detected by the isolation forest and the XGBoost, in the form of a Venn diagram. XGBoost, extreme gradient boosting.

total of 51 of the 100 submitted claims have been found to be highly suspicious by the special investigation unit. These results generally indicate the considerable potential of using machine learning methods for insurance claim fraud detection and further confirm the results of the previous two evaluation steps. Moreover, the strong performance of XGBoost on new claims gives it a greater advantage over isolation forests compared with the first two steps of the evaluation.

It is of considerable interest to investigate whether the two methods identify similar fraud patterns—and therefore detect the same highly suspicious claims—or whether the methods differ in this regard. This inquiry is depicted in the final row of Table 8 and in Figure 3 in the form of a Venn diagram. Among the 51 claims found to be highly suspicious by the special investigation unit, 16 claims were identified by both machine learning methods. However, the unsupervised approach detected six highly suspicious claims that were not detected by the supervised approach, and the supervised approach detected 13 claims that were not detected by the unsupervised approach.[10] Thus, the fraud patterns identified by the two machine learning approaches are apparently dissimilar. To more deeply understand this finding, we consider two possible explanations for why additional highly suspicious claims have been detected in the

[10]Thus, 16 claims were identified by both methods, six only by the isolation forest, and 13 only by the XGBoost, resulting in a total of 51 detected highly suspicious claims.

field experiment. The first possibility is that new fraud patterns detected by the machine learning models are outside the scope of the insurance company's fraud detection mechanism. The second possibility is that highly suspicious claims have been detected by the machine learning models that did not differ completely from already known fraud patterns but were simply overlooked by the insurance company's existing fraud detection mechanism due to the limited available resources. In this context, it is likely that the XGBoost has identified highly suspicious claims that have avoided the scrutiny of the existing mechanism, while the isolation forest has identified highly suspicious claims with novel fraud patterns. This important finding indicates that the two machine learning approaches (i.e., supervised and unsupervised) are not substitutes but instead complementary approaches that can be combined to increase the efficiency of insurance claim fraud detection.

5.4 | Explaining machine learning models for fraud detection

To further understand how the fraud patterns identified by the two machine learning approaches differ, we investigate which machine learning features were most important for each approach in terms of identifying fraudulent claims in the field experiment. To make transparent the mechanics of the processes employed by these approaches (i.e., the commonly criticized black box of machine learning), we use a state-of-the-art explainable artificial intelligence technique called SHAP (Lundberg & Lee, 2017). SHAP values provide insight into how much a certain feature contributes to a particular prediction made by a machine learning model. To derive the contribution of a feature for a particular prediction, SHAP calculates how much the predicted value changes if a feature is added to the model. The resulting SHAP values can also be aggregated over many predictions to obtain insights into how the model works in general. We calculate SHAP values for the 50 claims identified as suspicious by the isolation forests and for the 50 claims identified as suspicious by the XGBoost and present the results of this analysis in Figure 4.

As the top left corner of the figure shows, the amount claimed by the policyholder and the premium paid by the policyholder for insuring a car are the most important features for the first isolation forest in its detection of the 50 most highly anomalous claims. The power of the engine of the car is also an important feature for the algorithm. This indicates that anomalous combinations of these three features are particularly informative of fraud. For the second isolation forest, the deductible, the dunning level, and the age of the contract are particularly important features for identifying fraudulent claims. The no-claim class and the previous fraud record entries are also important features. Overall, the second isolation forest suggests that indicators of intent, economic incentives, and financial strain are particularly informative for identifying fraud and may be even more relevant than indicators of criminal energy.

For the XGBoost, the premium of the car, the claimed amount, and the mileage of the car are the three most important features. Although the first two features are also most important for the first isolation forest, the mileage of the car plays a much bigger role for the XGBoost. Furthermore, the deductible, the most important feature for the second isolation forest, has almost no importance for the XGBoost. Furthermore, the dunning level, the second most important feature for the second isolation forest, is of rather low importance for the XGBoost.

Overall, the SHAP analysis confirms that supervised and unsupervised learning approaches emphasize different features when detecting claim fraud. This aligns with the field experiment

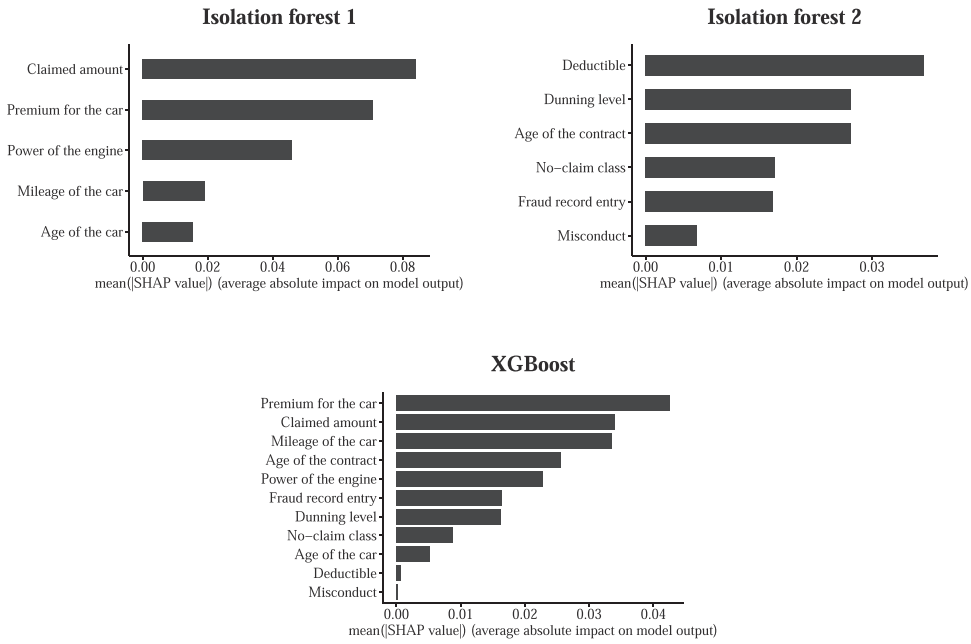

FIGURE 4 Feature importance via SHAP. This figure displays the variable importance for the isolation forests and the XGBoost using SHAP. SHAP, SHapley Additive exPlanations; XGBoost, extreme gradient boosting.

finding that the two approaches did not identify the same fraudulent claims, instead acting as complementary fraud detection mechanisms.

6 | CONCLUSION

Given the economic value associated with preventing fraud, it is important for insurance companies to implement efficient and effective processes to identify fraudulent claims. Besides being in the financial interest of insurance companies, this also benefits honest insurance holders by lowering insurance premiums. The extant literature has often assessed supervised learning methods to detect insurance fraud. Supervised learning has several potential shortcomings: There are usually few labeled cases, and unknown fraud patterns could remain undetected. Unsupervised learning, especially anomaly detection, addresses these issues. However, while unsupervised learning has been studied in several other fraud detection contexts, the literature on insurance fraud detection has paid relatively little attention to unsupervised learning. Moreover, there is little empirical evidence that can guide the decision between using supervised and unsupervised learning for insurance fraud detection. Because insurance companies are particularly interested in detecting new cases, in addition to using observational data, it is important to also study nonobservational data to understand the differences between unsupervised and supervised learning in this context. For example, more comprehensive detection of partially known patterns could benefit supervised learning, and the detection of new patterns could benefit unsupervised learning.

This study has considered the application of isolation forests, a convenient but effective unsupervised learning algorithm that can be used for fraud detection, and XGBoost, a fast and effective current supervised machine learning algorithm. We further compared these approaches to neural-network-based and clustering-based fraud detection algorithms. We have considered how the supervised and unsupervised learning methods perform in terms of identifying insurance fraud in a large proprietary data set and in terms of identifying insurance fraud in a field experiment. Our results generally emphasize the usefulness of unsupervised learning for insurance companies (particularly when no labeled data are available). However, even when limited labeled data are available, the supervised learning approach performs strongly, on par with the unsupervised learning approach. Our results further suggest that both, supervised learning and unsupervised learning methods detect fraud cases that have not been identified by existing mechanisms so far and that the detected cases partly differ. Moreover, explainable artificial intelligence methods reveal that the supervised and unsupervised learning methods use different input information. As such, our results indicate that supervised and unsupervised learning methods should be considered complements rather than substitutes.

ACKNOWLEDGMENTS

We thank Joan Schmit (the editor), one anonymous senior editor, two anonymous referees, Lars Beckmann, Andreas Pfingsten, Judith Schneider, Wenjun Zhu, and the participants at the 17th Credit Scoring and Credit Control Conference, the 2021 International Conference on Operations Research, the 7th Rostock Conference on Service Research, the 42nd International Conference on Information Systems (ICIS), and the 2022 Global AI Finance Research Conference for providing us with very helpful comments and suggestions. We further thank Constanze Schlesinger for excellent research assistance. Open Access funding enabled and organized by Projekt DEAL.

ORCID


Johannes Kriebel http://orcid.org/0000-0003-1249-5189

REFERENCES

Ai, J., Brockett, P. L., Golden, L. L., & Guillén, M. (2013). A robust unsupervised method for fraud rate estimation. *Journal of Risk and Insurance*, *80*(1), 121–143.

Artís, M., Ayuso, M., & Guillén, M. (1999). Modelling different types of automobile insurance fraud behaviour in the spanish market. *Insurance: Mathematics and Economics*, *24*(1–2), 67–81.

Artís, M., Ayuso, M., & Guillén, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, *69*(3), 325–340.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of the ICML Workshop on Unsupervised and Transfer Learning* (pp. 37–49).

Bauder, R., da Rosa, R., & Khoshgoftaar, T. (2018). Identifying medicare provider fraud with unsupervised machine learning. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 285–292).

Bauer, D., Leverty, J. T., Schmit, J., & Sydnor, J. (2021). Symposium on insure-tech, digitalization, and big-data techniques in risk management and insurance. *Journal of Risk and Insurance*, *88*(3), 525–528.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole.

Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., & Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance*, *69*(3), 341–371.

Brockett, P. L., Xia, X., & Derrig, R. A. (1998). Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance*, *65*(2), 245–274.

Caudill, S. B., Ayuso, M., & Guillén, M. (2005). Fraud detection using a multinomial logit model with missing information. *Journal of Risk and Insurance*, *72*(4), 539–550.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 1–58.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*(4), 588–608.

Crocker, K. J., & Morgan, J. (1998). Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts. *Journal of Political Economy*, *106*(2), 355–375.

Derrig, R. A. (2002). Insurance fraud. *Journal of Risk and Insurance*, *69*(3), 271–287.

Dionne, G., Giuliano, F., & Picard, P. (2009). Optimal auditing with scoring: Theory and application to insurance fraud. *Management Science*, *55*(1), 58–70.

Dixon, M. F., Halperin, I., & Bilokon, P. (2020). *Machine learning in finance*. Springer.

Duffield, G., & Grabosky, P. (2001). The psychology of fraud. In *Trends and Issues in crime and criminal justice* (Vol. 199). Australian Institute of Criminology.

Duval, F., Boucher, J.-P., & Pigeon, M. (2023). Enhancing claim classification with feature extraction from anomaly-detection-derived routine and peculiarity profiles. *Journal of Risk and Insurance*, (forthcoming).

Fritzsch, S., Scharner, P., & Weiß, G. (2021). Estimating the relation between digitalization and the market value of insurers. *Journal of Risk and Insurance*, *88*(3), 529–567.

Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, *88*(3), 591–624.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

Grabosky, P. N., & Duffield, G. M. (2001). Red flags of fraud. In *Trends and issues in crime and criminal justice* (Vol. 200). Australian Institute of Criminology.

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305.

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1479–1489.

Insurance Europe. (2019). *Insurance fraud: Not a victimless crime*. https://www.insuranceeurope.eu/publications/703/insurance-fraud-not-a-victimless-crime/Insurance%20fraud%20-%20not%20a%20victimless%20crime.pdf

Jiang, X., Lin, K., Zeng, Y., & Yang, F. (2021). Medical insurance medication anomaly detection based on isolated forest proximity matrix. In *2021 16th International Conference on Computer Science & Education (ICCSE)* (pp. 512–517). IEEE.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Medicare fraud detection using neural networks. *Journal of Big Data*, *6*(1), 1–35.

Johnson, M. E., & Nagarur, N. (2016). Multi-stage methodology to detect health insurance claim fraud. *Health Care Management Science*, *19*(3), 249–260.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Lessmann, S., Listiani, M., & Voß, S. (2010). Decision support in car leasing: A forecasting model for residual value estimation. In *Proceedings of the 31st International Conference on Information Systems (ICIS)*.

Liang, C., Liu, Z., Liu, B., Zhou, J., Li, X., Yang, S., & Qi, Y. (2019). Uncovering insurance fraud conspiracy with network learning. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1181–1184).

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).

Metzler, D., & Bruce Croft, W. (2007). Linear feature-based models for information retrieval. *Information Retrieval*, *10*(3), 257–274.

Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, *2*(1), 58–75.

Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., & Reynkens, T. (2022). Social network analytics for supervised fraud detection in insurance. *Risk Analysis*, *42*(8), 1872–1890.

Picard, P. (1996). Auditing claims in the insurance market with fraud: The credibility issue. *Journal of Public Economics*, *63*(1), 27–56.

Riedinger, D., & Major, J. (2002). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. *Journal of Risk and Insurance*, *69*(3), 309–324.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90.

Stripling, E., Baesens, B., Chizi, B., & vanden Broucke, S. (2018). Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud. *Decision Support Systems*, *111*, 13–26.

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

Tennyson, S., & Salsas-Forn, P. (2002). Claims auditing in automobile insurance: Fraud detection and deterrence objectives. *Journal of Risk and Insurance*, *69*(3), 289–308.

Tumminello, M., Consiglio, A., Vassallo, P., Cesari, R., & Farabullini, F. (2023). Insurance fraud detection: A statistically validated network approach. *Journal of Risk and Insurance*, (forthcoming).

Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., & Dedene, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, *176*(1), 565–583.

Viaene, S., & Dedene, G. (2004). Insurance fraud: Issues and challenges. *The Geneva Papers on Risk and Insurance—Issues and Practice*, *29*(2), 313–333.

Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, *69*(3), 373–421.

Vosseler, A. (2022). Unsupervised insurance fraud prediction based on anomaly detector ensembles. *Risks*, *10*(7), 132.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87–95.

Weisberg, H. I., & Derrig, R. A. (1991). Fraud and automobile insurance: A report on bodily injury liability claims in massachusetts. *Journal of Insurance Regulation*, *9*(4), 497–541.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*, *90*, 743–768. https://doi.org/10.1111/jori.12427