MILLIMAN REPORT

AI-supported anomaly detection in insurance

December 2025

Leonid Zeldin
Bartosz Gaweda
Dmitrii Ivanitckii

Milliman

Table of contents

Introduction

Insurance companies depend on accurate, high-quality data to operate effectively, whether for pricing policies, processing claims, or managing risk. However, the large volume and complexity of insurance data, combined with its creation and handling across multiple systems and stakeholders, make it susceptible to anomalies. These anomalies may stem from human error, system integration issues, unusual customer behavior, or deliberate fraud. Detecting such irregularities is crucial for safeguarding operational integrity, ensuring compliance, and maintaining customer trust.

Traditional anomaly detection in insurance has often relied on manual review or rule-based systems. While these approaches benefit from expert knowledge and are transparent, they struggle to scale with growing datasets and cannot easily adapt to new or complex anomaly patterns. Recent advances in machine learning, particularly unsupervised methods, offer promising alternatives by learning the underlying structure of data and identifying deviations without requiring extensive labeled examples.

This work investigates the application of ensemble-based unsupervised learning models, specifically autoencoder and variational autoencoder ensembles, for anomaly detection in insurance data. By training multiple models with different architectures and combining their outputs, we aim to improve robustness, scalability, and detection accuracy. In addition, we explore a hybrid approach, where outputs from unsupervised models are used as features for supervised learning, enabling the system to incorporate expert feedback and historical labels for more targeted detection.

The study is structured in two main parts:

1. **Preliminary analysis:** We benchmark a range of classical machine learning algorithms (e.g., KNN-inspired methods, clustering approaches, one-class support vector machines, and Isolation Forest) and deep learning models (autoencoders and variational autoencoders) on open-source proxy datasets. This stage compares detection accuracy, scalability, and computational efficiency, highlighting the strengths and limitations of each method.
2. **Pilot project with real insurance data:** In collaboration with an insurance company, we apply the developed methods to automobile insurance contracts and claims data from the past five years, including a small set of confirmed fraud cases. The pipeline combines unsupervised ensembles for anomaly detection with supervised models - notably eXtreme Gradient Boosting (XGBoost) - to identify potential fraud. Data preparation steps include integration of contract and claim attributes, temporal feature engineering, categorical encoding, and normalization. The preliminary results are reviewed with domain experts to validate findings and refine the models.

Beyond fraud detection, the paper discusses broader application areas, such as data migration, risk monitoring, and operational data quality assurance. We also outline future directions, including the integration of large language models (LLMs) for nontabular data and more complex hybrid pipelines that combine multiple detection techniques.

Overall, this research demonstrates that ensemble-based deep learning methods - especially autoencoder ensembles - offer a robust, scalable, and adaptable solution for anomaly detection in the insurance industry. The pilot project confirms their potential to enhance fraud detection processes, reduce manual workload, and improve decision-making, while opening avenues for further innovation in data-driven risk management.

Business rationale

Insurance companies rely on accurate and trustworthy data to make informed decisions, whether for pricing policies, approving claims, or assessing risks. Data quality directly impacts customer experience, regulatory compliance, and financial performance.

However, insurance data is generated, updated, and transferred across multiple systems and departments, introducing opportunities for errors, inconsistencies, or deliberate manipulation. Anomalies can arise from human error, integration issues, unusual customer behavior, or fraud. These may be simple outliers, where one or more values fall far outside the normal range, or structural anomalies, where individually plausible values occur in highly unlikely combinations. Detecting such irregularities is essential for preventing fraud, maintaining data integrity, and ensuring reliable analytics.

Traditionally, anomaly detection in insurance relies on manual review and expert judgment, which allows nuanced case-by-case analysis. While effective for small datasets, manual methods are costly, time-consuming, and prone to oversight, especially as data volumes grow and patterns become more complex.

Rule-based systems represent the first level of automation and codify domain expertise into predefined thresholds and conditions (e.g., unusually high claim amounts, implausible customer ages, or duplicate records). These systems are transparent and easy to implement but rigid, often generating excessive false positives or missing complex anomalies. Updating rules to reflect new patterns can be cumbersome and limit scalability and adaptability.

Modern machine learning methods - particularly unsupervised anomaly detection - offer a more dynamic approach by learning the underlying structure of data and identifying unusual cases without requiring labeled examples. This makes them especially relevant in the insurance sector, where confirmed anomaly or fraud cases are rare and diverse.

The overarching aim of this work is to explore the areas of the insurance industry where unsupervised anomaly detection can provide significant value. Potential application domains include fraud detection, data migration quality assurance, operational risk monitoring, and data quality improvement. By systematically evaluating different algorithms and approaches, we seek to identify scenarios where these methods can enhance efficiency, scalability, and decision-making in insurance operations.

Foundational study: Benchmarking unsupervised anomaly detection methods in insurance data

As outlined above, the goal of this thesis is to identify and evaluate potential application areas for unsupervised anomaly detection within the insurance industry. Before applying these methods to real-world insurance datasets, it is essential to examine their performance, strengths, and limitations in a controlled environment. This chapter presents a foundational study that acts as the methodological starting point for the subsequent applied research in this work. The outcomes of this study help to pinpoint promising approaches, understand their characteristics, and make informed decisions about model selection for the later pilot project.

1. INTRODUCTION

This study is based on the paper “A machine learning-based anomaly detection framework in life insurance contracts,” by Andreas Groll, Akshat Khanna, and Leonid Zeldin (the Groll life insurance study), which investigates a range of unsupervised anomaly detection algorithms using synthetic datasets closely resembling real-life insurance data.[1] Recognizing the scarcity of publicly available, labeled datasets in the insurance sector, the authors focus on methods capable of detecting anomalies without prior knowledge of their exact form. The Groll life insurance study systematically compares classical techniques, such as KNN-inspired proximity measures, clustering algorithms, one-class support vector machines, and Isolation Forest, with modern deep learning approaches, including autoencoders and variational autoencoders. Performance is assessed using metrics such as detection accuracy, runtime, and scalability, and ensemble strategies are explored to increase robustness and mitigate the risk of poor performance from suboptimal architecture selection.

1. Journal of Insurance Issues, 48 (2025). Retrieved December 1, 2025, from https://www.jstor.org/stable/48842644.

The results show that Isolation Forest is the most efficient and reliable among classical methods, though it tends to over-detect anomalies, while autoencoder and variational autoencoder ensembles achieve the most robust and accurate detection across datasets of varying size, with autoencoder ensembles offering shorter runtimes. These findings provide a strong methodological foundation for this thesis, guiding the selection of algorithms for real-world applications and indicating that ensemble-based deep neural networks hold particular promise for scalable and reliable anomaly detection in the insurance domain.

With this background established, the following section describes the datasets and preprocessing steps used in the benchmark analysis, thereby ensuring a transparent and reproducible methodological framework for evaluating the selected algorithms.

2. DATA SOURCES AND PREPROCESSING

Due to the lack of open-access life insurance datasets, the study employs two open-source health insurance datasets as proxies. Dataset 1 is relatively small and composed of 986 records and 12 variables, such as age, height, weight, information about diseases, and the premium price; variables were expanded to 13 with the addition of BMI. Dataset 2 is much larger, with 25,000 records and 24 variables, such as personal information about the customer, information about their health level, and the structure of the policy; variables were expanded to 51 after one-hot encoding of categorical data. Both datasets undergo thorough preprocessing: Missing values are removed, categorical variables are converted using one-hot encoding, and features such as BMI are engineered where appropriate.

To facilitate rigorous method benchmarking, four obvious anomalies are manually introduced into each dataset. These anomalies are designed to mimic realistic yet detectable irregularities, therefore ensuring that the evaluation reflects practical challenges.

However, it is important to note several limitations arising from the use of synthetically generated data and anomalies. First, the datasets, as well as the manually inserted anomalies, are synthetic and therefore may not fully capture the complexity and irregularity of real-world insurance data. Synthetic data is typically generated based on predefined rules, which can result in smoother and more regular patterns compared to the unstructured scenarios commonly found in actual insurance datasets. Second, the inserted anomalies may be overly conspicuous, standing out distinctly from the rest of the data. This could potentially diminish the importance of less obvious anomalies, as they may appear less significant in comparison. Consequently, reproducing the results on real-world data may present additional challenges, and the performance of anomaly detection methods might differ when applied to genuinely unstructured and noisy datasets.

3. METHODS AND ALGORITHMS

This section provides an overview of various unsupervised anomaly detection algorithms, spanning classical machine learning techniques and modern deep learning approaches, as well as the metrics used to assess their performance. The methods themselves are not explained in detail here; the Groll life insurance study discussed above contains comprehensive descriptions and in-depth discussions.

3.1 Classical machine learning methods

3.1.1 Nearest Neighbors (KNN-inspired, proximity-based method)

Nearest Neighbors (NN), a KNN-inspired method, calculates an anomaly score for each point as the average (Euclidean) distance to its k nearest neighbors. A threshold, typically determined by the mean and standard deviation of these distances, is used to flag anomalies. While this approach is computationally efficient and easy to interpret, its effectiveness diminishes with increasing data dimensionality and complexity.

3.1.2 Clustering-based methods

- **K-MEANS:** The *k*-means partitions data into *k* clusters and classifies points far from cluster centers as anomalies. The optimal number of clusters is determined using the silhouette score. However, as the dataset size grows, *k*-means becomes computationally expensive and less effective.
- **DBSCAN:** Density-based spatial clustering of application with noise (DBSCAN) identifies clusters based on density and flags points in sparse regions as anomalies. The desired density is set by two parameters: an ε value that defines a ball of radius ε around every point and *k*, the desired number of points in each sphere. A point which has *k* or more neighbors within its ε-ball is called a core point. A point that is located in some core point's ball but is no core point itself (e.g., is an ε away from a core point but has less than *k* neighbors in its ε-ball) is called a border point. If a point has less than *k* neighbors in its ε-ball and no one of them is a core point, this point is classified as an anomaly. Some drawbacks of this method include ambiguity in classification of border points of two clusters, a choice of an appropriate distance and ε parameter of high-dimensional data, and failure to handle clusters of different density.
- **HDBSCAN:** Hierarchy density-based spatial clustering of applications with noise (HDBSCAN) builds on the DBSCAN algorithm by adding a hierarchical method to better find clusters in data that have different densities. The main idea is that areas with many data points are likely to form clusters, while sparser regions separate them. Unlike DBSCAN, which uses a single, fixed density threshold, HDBSCAN adjusts this threshold automatically, creating a hierarchy of clusters that reveals more details about the data. A key concept in HDBSCAN is the core distance, which measures how densely packed the area around each point is. This helps define the mutual reachability distance between points, making the algorithm less sensitive to local changes in density and preventing it from joining points that are not truly part of the same cluster.

Two main parameters in HDBSCAN are the minimum cluster size and κ. The minimum cluster size helps ensure that only meaningful clusters are found, while κ affects how local density is measured and influences the resulting hierarchy. To build the hierarchy, HDBSCAN replaces DBSCAN's fixed threshold with a process based on a minimum spanning tree (MST) constructed from mutual reachability distances. This tree represents all possible clusters at different density levels. By gradually removing the longest edges in the tree, HDBSCAN creates a nested structure of clusters at different resolutions. Small or weak clusters are filtered out, and the most stable clusters - those that persist over a wide range of thresholds - are chosen as the final result.

3.1.3 One-class support vector machines

Support vector machines (SVMs) can be adapted for anomaly detection through the use of one-class SVMs (OCSVM). In this setting, the objective is to characterize the distribution of "normal" data, enabling the identification of outliers that do not conform to established patterns.

A central component of SVMs is the kernel function, which facilitates the mapping of data into a high-dimensional feature space. This transformation allows the algorithm to construct boundaries that can separate normal data from potential anomalies even when the data exhibits complex, nonlinear structure.

Several kernel functions are commonly employed in OCSVMs. The linear kernel is suitable when normal data can be separated from anomalies using a simple hyperplane. However, real-world data often exhibits more intricate patterns, necessitating the use of nonlinear kernels. The radial basis function (RBF) kernel is frequently preferred in anomaly detection applications due to its capacity to create flexible, nonlinear decision boundaries that can encapsulate clusters of normal data. The polynomial kernel is another option, though it is generally reserved for cases where domain knowledge suggests that the data follows polynomial relationships.

3.1.4 Isolation Forest (tree-based method)

Isolation Forest (IF) detects anomalies by isolating data points through random partitioning. Points that can be separated with fewer splits are considered more anomalous. The method is efficient and well suited for large, high-dimensional datasets.

3.2 Deep learning methods

3.2.1 Autoencoder

Autoencoders (AEs) are neural networks trained to reconstruct their input. They consist of three main components: the encoder, the decoder, and the latent space. The encoder gradually reduces the dimensionality of the original data until it reaches the latent space, a compressed representation of the input. The decoder then uses this representation to reconstruct the original data points. The difference between the original and reconstructed data - the reconstruction error - indicates how well the model has captured the underlying data structure.

For anomaly detection, the key idea is that the AE learns to represent and reconstruct patterns that are common in the training data. If a new data point deviates significantly from these learned patterns, the model will struggle to reconstruct it, which results in a higher reconstruction error. Such points are flagged as potential anomalies. In our work, we use a feedforward AE, where information flows only in one direction (from input to output) without feedback to previous layers. Training is performed using backpropagation over multiple iterations, called epochs.

3.2.2 Variational autoencoder

The variational autoencoder (VAE) extends the basic AE by estimating the mean and standard deviation within the latent space. From these parameters, it samples points of the same size from a multidimensional normal distribution. The remaining components and procedures are the same as in the basic AE.

3.3 Measurement techniques

3.3.1 Silhouette score

The silhouette score evaluates how well a data point fits within its assigned cluster by comparing its average similarity to points in the same cluster with its similarity to points in other clusters. High values indicate that the point is well matched to its own cluster, while low or negative values suggest poor clustering or that the point may be an outlier. The average silhouette score across all data points is a useful indicator of the overall goodness of fit of the clustering model.

3.3.2 Anomaly score

In the IF method, the anomaly score measures how easily a data point can be separated from the rest of the dataset. This is determined by calculating the average path length of the point across all trees in the forest, with the path length being the number of splits required to isolate the point. Anomalies typically have shorter average path lengths because they differ significantly from the majority of the data and are easier to isolate. In contrast, normal points require more splits, resulting in longer average path lengths and lower anomaly scores.

It is important to note that the distribution of anomaly scores, rather than their overall average, is more informative for evaluating the model’s performance. In datasets where anomalies are rare, the average score across all points will be dominated by normal data and may remain low even if the model successfully identifies the few anomalies.

3.3.3 Reconstruction error

In an AE, the reconstruction error measures the difference between the original input data and the data reconstructed from the latent representation. It reflects how well the model has learned to capture the essential features of the input. A low reconstruction error indicates that the AE can accurately reproduce the input, while a high reconstruction error suggests that important patterns were not captured. In anomaly detection, unusual data points often lead to higher reconstruction errors because the model fails to represent them effectively. By monitoring reconstruction error and comparing it to a predefined threshold, anomalies can be identified based on deviations from the normal reconstruction error distribution.

One important caveat when using the average reconstruction error is that it can mask different behaviors across models. For example, two models can have the same average reconstruction error: One might reconstruct most points slightly inaccurately, while the other might perfectly reconstruct all normal points but fail significantly on a few anomalies. In such cases, relying solely on the average can be misleading, and the distribution of reconstruction errors should also be examined.

3.4 Ensemble learning

Ensemble learning combines the outputs of multiple models to improve overall performance and robustness. Instead of relying on a single model, different model architectures or training runs are employed, and their outputs are subsequently aggregated. This approach reduces the risk that selecting a suboptimal architecture will lead to degraded results.

4. EXPERIMENTAL SETUP AND EVALUATION

The evaluation covers all tested approaches, including ensemble learning, alternative methods, and single-model configurations, to provide a comprehensive comparison of performance across different anomaly detection techniques.

In our case, ensemble learning was applied to both AEs and VAEs to address the issue of misleading average reconstruction errors. Three models were created for each method. Since it is not known in advance which architecture will yield the best or worst reconstruction performance, using multiple diverse models enables more stable anomaly detection. By combining the reconstruction results from several models, the influence of individual weak models is minimized, and the detection rate for rare but significant deviations is improved. These ensemble-based results were directly compared against those obtained from the other approaches in the study.

Each method was assessed on its ability to detect four manually inserted anomalies in both datasets. Evaluation criteria included detection accuracy, computational efficiency, and scalability. Where applicable, automatic parameter tuning techniques, such as grid search and silhouette score maximization, were employed to optimize model configurations.

FIGURE 1: COMPARISON OF VARIOUS CLASSICAL ANOMALY DETECTION METHODS

METHODS	DATASET 1 (990 ROWS)			DATASET 2 (24,010 ROWS)		
	ANOMALY DETECTED OUT OF 4	TOTAL RUNNING TIME	SILLHOUETTE SCORE	ANOMALY DETECTED OUT OF 4	TOTAL RUNNING TIME	SILLHOUETTE SCORE
NN	3	<1 sec	-	2	1.1 sec	-
K-MEANS	3	1:05 min	0.95	0	2.5 h	0.57
DBSCAN	4	6.7 sec	0.93	DNF	> 5 h	-
HDBSCAN	4	4.6 sec	0.91	DNF	> 5 h	-
OCSVM	2	<1 sec	-	1	13.5 sec	-
ISOLATION FOREST	4	<1 sec	0.48*	4	8.5 sec	0.46*

Note:
DNF: Did not finish
*Average Anomaly Score

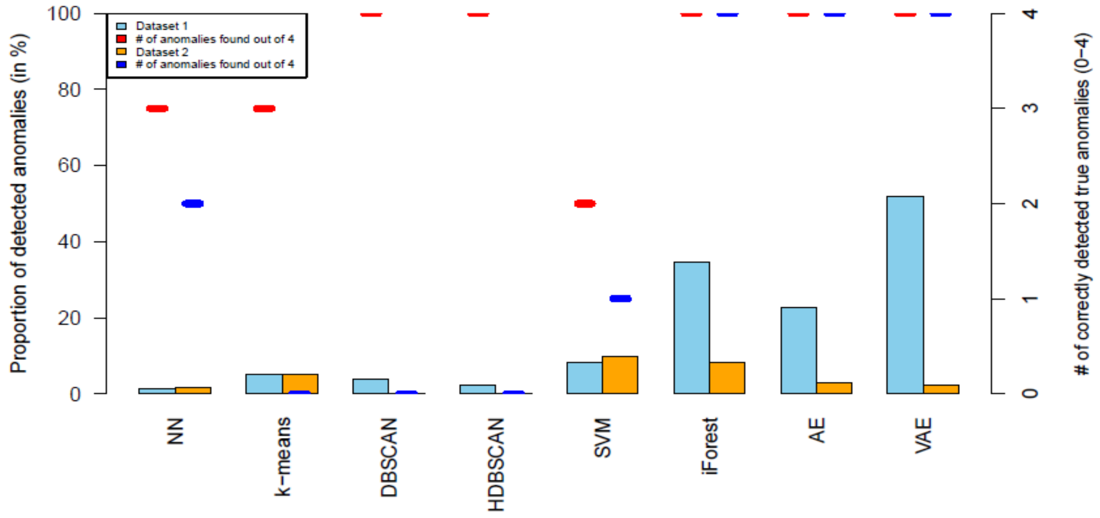
The first comparison table focused on detection accuracy, runtime, and clustering quality (via silhouette score) for classical methods such as NN, *k*-means, DBSCAN, HDBSCAN, OCSVM, and IF. For the first dataset, several methods - particularly HDBSCAN and IF - achieved complete detection of all four ground-truth anomalies with relatively short runtimes. DBSCAN and *k*-means also performed well but required longer execution times. In the second dataset, IF was the only classical method to detect all anomalies within seconds. *K*-means required over two hours and failed to detect any anomalies, and DBSCAN and HDBSCAN did not finish within the five-hour limit. These results highlight the scalability challenges of certain clustering-based methods when applied to larger datasets.

FIGURE 2: COMPARISON OF AUTOENCODER AND VARIATIONAL AUTOENCODER ENSEMBLES

METHODS	DATASET 1 (990 ROWS)			DATASET 2 (24,010 ROWS)		
	ANOMALY DETECTED OUT OF 4	TOTAL RUNNING TIME	TOTAL MODELS	ANOMALY DETECTED OUT OF 4	TOTAL RUNNING TIME	TOTAL MODELS
AUTOENCODER	4	2:22 min	3	4	6 min	3
VARIATIONAL AUTOENCODER	4	4:35 min	3	4	20 min	3

The second table compared AE and VAE ensembles. In both datasets, these ensemble approaches successfully detected all anomalies. However, despite both ensembles using three models, the AE ensemble consistently achieved shorter runtimes than the VAE ensemble: 2 minutes, 22 seconds versus 4 minutes, 35 seconds for the first dataset, and 6 minutes versus 20 minutes for the second dataset. This supports the hypothesis that ensemble learning mitigates architecture-selection risks, thereby ensuring stable detection performance across datasets of varying size.

FIGURE 3: PROPORTION OF ANOMALIES AND DETECTED MANUAL ANOMALIES (SECONDARY Y-AXIS)



The bars represent the relative proportion of points within each dataset that were classified as anomalous by the respective method. The small red and blue markers indicate the number of manually inserted anomalies (out of four) that were correctly detected by each method in the corresponding dataset.

The three strongest methods (IF, AE, and VAE) flagged a relatively high number of points as anomalous in the smaller dataset. In the larger dataset, this effect is less pronounced, particularly for the AE and VAE ensembles. Many of the classical methods struggled to perform well - or at all - on the larger dataset.

FIGURE 4: NUMBER OF TOTAL DETECTED ANOMALIES BY EACH METHOD

METHODS	DATASET 1 (990 ROWS)	DATASET 2 (24,010 ROWS)
NN	14	395
K-MEANS	50	1201
DBSCAN	38	DNF
HDBSCAN	22	DNF
OCSVM	83	2401
ISOLATION FOREST	342	1974
AUTOENCODER	224	705
VARIATIONAL AUTOENCODER	512	548

Figure 4 shows the absolute numbers of anomalies detected by each method for each dataset. In the first dataset, the three nominally strongest methods (IF, AE, and VAE) classified a relatively large share of points as anomalous. This may suggest that, on a small dataset, they were not able to fully capture and learn the underlying structures. In the larger dataset, the situation appears quite different, especially for AE and VAE, which flagged fewer points relative to the dataset size.

While some classical approaches, such as OCSVM and IF, also reported high anomaly counts, these often exceeded the known ground-truth anomalies by a wide margin, indicating a likely increase in false positives. AE and VAE ensembles produced more moderate counts overall, which may reflect stricter anomaly thresholds, though this cannot be confirmed without an explicit false-positive analysis.

Across both datasets, AE and VAE ensembles performed well in terms of runtime and detection of the manually inserted anomalies. Given that VAE appears to struggle more on smaller datasets, AE ensembles were chosen for further analyses as a balanced option with consistent performance across scales.

5. DISCUSSION OF RESULTS

The comparative evaluation demonstrates that the KNN-inspired method and *k*-means clustering can identify a subset of anomalies in smaller datasets but are inadequate for larger, more complex data due to scalability limitations and reduced accuracy. Density-based clustering methods (DBSCAN and HDBSCAN) exhibit strong performance in smaller datasets and successfully detect all anomalies, but these methods become impractical as data size increases due to excessive computation times. While theoretically promising for high-dimensional data, OCSVM is hampered by its sensitivity to parameter selection and the absence of robust, automated tuning mechanisms.

IF emerges as the most effective classical approach, consistently detecting all manually inserted anomalies in both datasets with minimal computation time. However, its relatively low average anomaly score suggests limited confidence in its predictions and a tendency to over-detect, potentially increasing false positives.

Deep learning-based ensembles, particularly those based on AEs, outperform all classical methods in both accuracy and robustness. The ensemble strategy mitigates the risk of overfitting or missing subtle patterns, and the models are well suited to the complex, high-dimensional nature of insurance data. The VAE ensemble also performs strongly but, in smaller datasets, may over-classify anomalies, indicating the need for careful thresholding or further refinement. Across both datasets, the AE ensemble delivers highly consistent results with shorter runtimes than the VAE ensemble and is a strong candidate for scalable anomaly detection in practice.

Robust decision-making within autoencoder ensembles

A central aspect of this study is the definition of criteria for identifying data points as anomalies within an ensemble of 10 AE models featuring different architectures. This study employed 10 models, while the Groll life insurance study only used three models. In the Groll life insurance study, the mean reconstruction error was computed for each data point across all AE and VAE models, and data points with an average reconstruction error exceeding a manually defined threshold were flagged as anomalies.

However, this approach is prone to bias. Models that produce generally higher reconstruction errors for a larger number of points can dominate the decision-making process and potentially lead to incorrect classifications.

To address this issue, we drew on insights from the literature and implemented a modified ensemble-based detection strategy. The approach begins with training several separate AE models on the same dataset. For every data point, each model produces a reconstruction error - a measure of how well the model can reproduce the original input.

For each individual model, the distribution of reconstruction errors across all data points is examined, and the 99th percentile of this distribution is calculated. This percentile serves as a model-specific anomaly threshold, representing the error level that only the highest 1% of reconstruction errors exceed. Any data point whose reconstruction error is greater than this threshold is considered anomalous by that specific model. The threshold can be adjusted depending on the desired confidence level. For example, one can use boxplots and classify as anomalies all points lying outside the whiskers, or alter the percentile value from 99% to 95% or 99.9% to make the detection more or less conservative.

Once this process has been carried out for all models, the anomaly decisions are combined into a final classification. This can be done by requiring absolute consensus, meaning a data point is labeled anomalous only if all models agree, or by applying a majority rule, where a point is considered anomalous if more than half of the models classify it as such. In either case, the approach allows the ensemble to adapt to differences in model sensitivity and reduces the impact of outlier decisions from individual models, while leveraging agreement to improve robustness.

Application area: Migration

Data migration, a relatively recent phenomenon in the insurance sector involving the transfer of complete datasets from legacy systems to modern platforms, presents a heightened risk when outliers are present and may compromise data integrity, consistency, and system performance. Outliers may originate from legacy system errors, manual entry mistakes, or data corruption, and can lead to issues such as failed migrations, schema mismatches, or inaccurate analytical outcomes in the target system. These anomalous data points can disrupt automated validation checks, trigger unnecessary error handling processes, or propagate flawed information, ultimately compromising the reliability of the migrated database.

By employing automated outlier detection methods prior to migration, organizations can systematically identify and address problematic records by correcting, excluding, or flagging them for further review. This proactive approach ensures a cleaner, more reliable transfer of data, reduces the workload associated with manual inspection, and minimizes the risk of costly post-migration issues. In doing so, it safeguards operational continuity and preserves the quality of information in the target environment. Furthermore, the early detection of errors within systems can reduce the risk of attracting unwanted attention from regulatory authorities - an especially relevant factor in industries such as insurance, which are subject to strict supervision and compliance requirements.

This seems to represent one of the most promising application areas for unsupervised anomaly detection, with the potential to generate substantial value. In the context of data migration, successful anomaly detection can save both time and money, while enhancing the overall quality of organizational data. Further research and analysis in this field are highly warranted as advancements could significantly improve migration processes and strengthen data governance practices.

Fraud detection

1. PROJECT IDEA

After conducting extensive theoretical analyses, we aimed to assess whether our models could perform effectively on real-world data. To achieve this, we needed two key elements: First, a partner willing to provide data for analysis, and second, a relevant analytical task suited to the available data. As previously discussed, we identified two domains where anomaly detection could add significant value: data migration and fraud detection. We then began searching for partners interested in collaborating with us to bring this research into practice. Ultimately, we partnered with a middle sized insurance company and launched a pilot project.

This project investigates the applicability of unsupervised anomaly detection methods - specifically ensembles of AEs - for fraud detection in tabular automobile insurance data. In collaboration with the insurer, we obtained access to their contracts and claims datasets for the last five years, including a list of fraud cases identified through their internal investigations. Our initial objective was to employ ensembles of unsupervised AE models to detect potential fraudulent schemes beyond those already recognized by the insurer. As the research progressed, we incorporated supervised techniques to enhance analytical rigor and develop a comprehensive fraud detection pipeline.

Upon further analysis, we realized that investigating only the already known fraudulent claims would not generate sufficient new insights. A closer examination revealed that the insurer’s provided list of fraudulent claims included only those with certain special structures. Consequently, we split the data into two segments based on structural differences in confirmed fraud cases. The first segment included a supervised analysis for claims exhibiting these structures and leveraged the labeled data. The second segment included an unsupervised analysis using ensembles of AEs for the remaining claims, where we aimed to uncover novel patterns potentially indicative of smaller-scale fraud.

2. DATA RECEIVED

The insurer provided two principal datasets:

- **Customer and contract dataset:** This dataset contains detailed information on customers and their contracts, including contract structures, insured risks, coverage amounts, and relevant dates (such as date of birth, contract start date, and last claim date). The data is available in semi-annual snapshots, allowing us to track changes over time and to contextualize each claim within its contractual framework. The sample size is over two million policies.
- **Claims dataset:** This dataset comprises nearly all claims from the past five years, documenting information available at claim submission (a priori data), as well as subsequent actions, payments, involved third parties, and regulatory activities throughout the claims process (a posteriori data). The breadth of this dataset supports both temporal and process-oriented analyses. The sample size is over one hundred thousand claims.

We also received a list of claims confirmed to be fraudulent by the insurer. Notably, the imbalance level was really high, which reflects the rarity and complexity of confirmed fraud cases in the dataset.

A comprehensive data description accompanied the datasets and clarified the meaning and possible values for each attribute. Where relevant, we requested and received additional information, such as police involvement in claims, although some details remained unavailable due to privacy constraints.

3. DATA PREPARATION

Extensive data preparation was conducted to ensure compatibility with both unsupervised and supervised modeling approaches:

- **Attribute understanding and selection:** Each variable was assessed for relevance and informational value. Features deemed noninformative or redundant were excluded from further analysis to improve model performance and interpretability.
- **Temporal transformation:** All date fields were converted into numerical measures by calculating their difference in days from the claim report date. This transformation resulted in positive values for events occurring after the claim and negative values for events occurring before, facilitating chronological feature engineering.
- **Data integration:** Claims data was enriched with contract details via a left join, ensuring each claim could be analyzed in the context of its associated contract attributes. This step was crucial for capturing relationships between customer profiles and claim characteristics.
- **A priori and a posteriori separation:** Features were systematically divided into those available immediately upon claim reporting (a priori) and those accessible only during or after claims processing (a posteriori). This separation enabled nuanced modeling strategies reflecting different stages of the claims life cycle.
- **Categorical encoding:** Categorical variables were transformed via one-hot encoding, converting each category into a binary indicator column suitable for machine learning algorithms. This approach ensures that categorical information is effectively utilized without introducing ordinal bias.
- **Handling missing values:** The final dataset contained relatively few missing values, represented as NAs, 999999s, or similar placeholders. To minimize information loss, missing values were imputed using the column mean. While this approach may introduce some bias, it was chosen to retain as much data as possible for modeling. For the supervised models, NA values were kept to avoid a bias.
- **Duplicate removal:** A small number of claims appeared twice in the delivered dataset; these duplicates were removed to ensure data integrity.
- **Normalization:** For unsupervised modeling with AEs, all feature values were normalized to the range [0, 1]. This step prevents features with extreme values from disproportionately influencing the model and ensures stable training dynamics.
- **Feature engineering:** Certain columns provided limited information on their own but could be combined with other attributes to capture more meaningful patterns. Accordingly, several new features were constructed, leveraging domain knowledge to enhance the dataset’s informational richness.
- **Final dataset assembly:** The prepared dataset was structured to enable consistent use by both unsupervised and supervised models, facilitating reproducible and comparable analyses across modeling approaches.

4. PROCESS STRUCTURE

Based on insights from our pilot project, we identified a best-practice framework for structuring the fraud detection process into sequential steps. This approach, outlined below, provides a general guideline. In our specific case, several elements had to be adapted due to practical challenges encountered during implementation.

1. **Initial anomaly detection (unsupervised):** An AE is to be applied to the raw, preprocessed data to identify anomalous structures. Claims exhibiting high reconstruction errors, which indicate unfamiliar or atypical patterns, are flagged as potential anomalies and submitted to the client for expert assessment regarding their fraudulent nature.
2. **Expert review and labeling:** After the initial anomaly detection, the insurer reviews the flagged claims and provides feedback, categorizing them as fraudulent or nonfraudulent. Combined with the insurer's original list of confirmed frauds, this step yields a labeled dataset suitable for supervised modeling.
3. **Supervised fraud modeling:** With labeled data available, supervised models are trained to learn patterns associated with fraudulent claims. These models can then evaluate new claims, indicating which are likely to be fraudulent based on their feature profiles. Additionally, anomaly score based on unsupervised models can be used as an additional feature of supervised models.

5. METHODOLOGY

Ensembles of autoencoders

As described in the results section above, we employed an ensemble of 10 AE models with varying architectures, which differed in the dimensions of their two hidden layers and the size of the latent space for each model. For each model individually, the threshold for anomalies was set at the 99th percentile of that model's reconstruction error, meaning that any data point whose reconstruction error exceeded this model-specific threshold was flagged as anomalous. Finally, we applied a majority voting scheme across the 10 models to determine the overall anomaly labels.

Supervised methods

As a key method for the supervised analysis, we selected XGBoost due to its quick implementation, built-in variable selection, mechanisms to avoid overfitting, and strong performance when dealing with unbalanced data. Moreover, XGBoost offers extensive possibilities for model interpretation, which is particularly valuable in business applications where transparency and explainability are essential.

XGBoost is an efficient and scalable implementation of gradient boosting decision trees.[2] The method builds an ensemble of decision trees in a sequential manner, where each new tree is trained to correct the errors of the previous ensemble by minimizing a differentiable loss function. XGBoost incorporates regularization terms in its objective function to prevent overfitting, as well as supports handling of missing values and allows for parallel computation. Its flexibility in defining custom loss functions and evaluation metrics makes it suitable for a wide range of predictive modeling tasks, including fraud detection.

To properly address the challenge of unbalanced data, we applied a weighted approach and assigned higher weights to the minority class to ensure that fraudulent cases were adequately represented in the training process. To reduce the risk of overfitting, the dataset was first split into separate training and test sets. The training set was further divided into cross-validation folds to enable robust parameter tuning and calibration.

The choice of the classification threshold was kept intentionally flexible, allowing adjustment according to the specific business case. For example, the threshold could be lowered to increase recall and minimize the likelihood of undetected fraud or raised to reduce the costs associated with additional investigations triggered by false positives.

A critical step in the modeling process was the careful selection of predictor variables to prevent target leakage. Certain variables, such as information on detective or lawyer involvement, can significantly boost model performance but are inherently biased because they reflect actions taken by the insurance company in response to suspected fraud in the past. Including such features would therefore compromise the validity of the predictive model.

2. Chen, T., & Guestrin, C. (August 13, 2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA. Retrieved December 1, 2025, from https://doi.org/10.1145/2939672.293978.

Additionally, such models should be tested from the perspective of potential bias in input data, possible unfair treatment or even discrimination. Unfortunately, as in the data set which we obtained we did not have any variables available such as gender, disability, race or religion, we could not assess the model from this perspective, and it was kept as out of scope of this project. Nonetheless, such procedure would be highly advisable.

For model interpretation, we employed several explainable AI (xAI) techniques, including partial dependence plots (PDP), individual conditional expectation (ICE) plots, variable importance analysis, and both global and local Shapley value analyses. These methods provided insights into the relationships between predictors and the target variable, as well as explanations for individual predictions, thereby enhancing the transparency and trustworthiness of the model's outputs.

6. PRELIMINARY ANALYSIS RESULTS

It is important to note that we cannot disclose the full and detailed results of the analysis. This limitation arises both from the need to prevent malicious actors from exploiting specific fraud detection patterns and from the fact that the insurer has not granted permission to release such sensitive information. Revealing exact variable importance rankings or concrete patterns could enable individuals to adapt their behavior in ways that avoid detection. Nevertheless, the XGBoost approach proved effective in rapidly identifying a relevant subset from several hundred available predictors that can support the detection of suspicious claims.

In practice, insurance companies often rely on known indicators to flag potentially risky cases. Our analysis confirmed that some of these commonly used indicators are also reflected in the data, alongside other less obvious factors identified by the model.

An important practical outcome of this pilot project was that the insurer received a list of claims identified by the model as potentially fraudulent. These insights allowed the client to focus investigative resources on the most suspicious cases and potentially identify instances of fraud that might have otherwise gone unnoticed.

To evaluate the performance of the supervised model, we used a confusion matrix. As discussed earlier, the decision threshold can be flexibly adjusted to meet different business objectives, such as prioritizing higher recall to capture more potential fraud cases or increasing precision to reduce the number of false positives. In our setting, the supervised model was able to identify the majority of historically confirmed fraud cases while maintaining a manageable number of false positives. While some false positives stem from prediction errors, they can also serve as valuable leads for further investigation and potentially uncover previously undetected fraud.

Finally, the use of xAI methods offers practical value for operational integration. For example, local Shapley values can provide claim handlers with case-specific insights, indicating why a claim appears suspicious and guiding their attention to the most relevant aspects. These interpretability tools can be used to refine existing manual rules or to directly embed the model into the claims handling process, thereby improving detection efficiency and decision-making.

At the final stage of the analysis, the dataset was split into two parts because the confirmed fraud cases provided by the insurer followed a specific pattern. The first part was used for the supervised modeling approach described above. On this subset, the supervised models achieved improved performance compared to earlier experiments, which suggests that the differing composition of the two subsets has a significant influence on model accuracy. The second part of the data is currently being analyzed using an ensemble of 10 AEs in an unsupervised learning setting.

Preliminary unsupervised findings show recurring structural patterns that merit deeper investigation. Additionally, the pre-analysis conducted with the unsupervised models was used to generate new features, which were then incorporated into the supervised models. This integration led to a slight improvement in predictive performance; however, the gain was not substantial enough to conclude that such an approach would consistently yield significant benefits across different datasets. The deeper analysis of the unsupervised results will be conducted after the completion of the present paper and will therefore not be included in the preliminary results section.

Outlook

FURTHER USE CASES

This research suggests promising opportunities for applying advanced anomaly detection across several areas of insurance operations beyond fraud detection. The system could be leveraged to help identify unusual customer behaviors, such as frequent claims or sudden changes in policy details, which may support proactive risk management. In addition, the hybrid approach may also be applicable for detecting potential data entry errors within policy and claims records, thereby supporting high data quality and reducing the likelihood of issues affecting business processes. Insurers might also explore its use for monitoring emerging risks or trends within their datasets, allowing timely adjustments to products and policies. Finally, the model's capability to prioritize claims based on anomaly scores could help investigation teams focus on the most suspicious or high-risk cases, potentially improving the efficiency and effectiveness of their workflows.

Furthermore, the bad actors are still coming up with new ideas which are not present in the old data set which might lead do a data drift. Ongoing monitoring, model recalibration are essential for production-ready anomaly detection systems.

In sum, this research has practical applications in fraud detection, risk assessment, data quality assurance, trend monitoring, and claims investigation, although further validation is needed to confirm the effectiveness of these approaches in each domain. Finally, due to very limited labeled data and the fact the project is not yet fully finalized the paper results should be treated with appropriate level of caution.

MORE COMPLEX PIPELINES

Integrating supervised and unsupervised learning models represents only the first step in exploring the potential synergy between different approaches to anomaly detection. A promising enhancement would be a governed architecture that integrates LLMs for unstructured inputs, automated data-quality monitors, and human-in-the-loop checkpoints. Such architecture should be designed in line with transparent and responsible-AI principles, explicitly defining governance expectations, ensuring decisions can be audited, and maintaining interpretability of model outputs for stakeholders. Image-based pattern recognition could further support the analysis of accident scenes, while clear documentation and version control would strengthen operational reliability in production settings.

A robust process could combine preprocessing with an ensemble of AEs to detect anomalous patterns, validation by clients or domain experts to confirm true anomalies, and supervised learning methods such as XGBoost for final classification. Embedding human oversight throughout the workflow not only ensures accountability but also provides targeted guidance to clients on which cases warrant closer investigation. This combination of technical rigor, interpretability, and governance alignment would help build trust and deliver sustained value in real-world anomaly detection applications.

PILOT PROJECT STATUS

Although some preliminary results are presented in this paper, the pilot project is still ongoing. The upcoming analysis of the second subset of claims using unsupervised methods is expected to provide additional insights into fraud detection. These findings may help refine the approach and improve its overall effectiveness. The continuation of the project offers the opportunity to validate the methodology on a broader dataset and to better understand its strengths and limitations in real-world conditions.

DATA MIGRATION

As described above, an unsupervised approach or a combination of unsupervised and supervised models has not yet been applied to a real-world dataset within the context of data migration. We see significant potential in this area and can envision integrating such methods into daily operations to improve the quality of support services and to showcase a modern, innovative way of working.

Solutions for a world at risk™

Milliman leverages deep expertise, actuarial rigor, and advanced technology to develop solutions for a world at risk. We help clients in the public and private sectors navigate urgent, complex challenges—from extreme weather and market volatility to financial insecurity and rising health costs—so they can meet their business, financial, and social objectives. Our solutions encompass insurance, financial services, healthcare, life sciences, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Leonid Zeldin
leonid.zeldin@milliman.com

Bartosz Gaweda
bartosz.gaweda@milliman.com

Dmitrii Ivanitckii
dmitrii.ivanitckii@milliman.com