An Empirical Study of Simulated Litigation Detection in Brazil: A Machine Learning Approach

Amilcar A. Netto[2] **, Péricles B. C. Miranda,**[1,2]**, André C. A. Nascimento** [1]

[1] Departamento de Computação, Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Brazil

[2]Cesar School, Recife, Brazil
aan@cesar.school
{pericles.miranda, andre.camara}@ufrpe.br

Abstract. *Simulated lawsuits (lides simuladas), in which parties collude to obtain a favorable ruling, impose a substantial and under-examined burden on Brazil's judiciary. This paper asks whether such cases can be flagged automatically from routinely available institutional data. We benchmark a broad range of machine learning methods on 6,412 labor-court settlements from Recife, of which only 1.75% are confirmed simulations: supervised classifiers (KNN, decision trees, random forests, gradient boosting) paired with over- and under-sampling; four anomaly detectors (Local Outlier Factor, Isolation Forest, One-Class SVM, Robust Mahalanobis); and an AutoML pipeline. The central finding is negative but informative: every method performs near chance, the best reaching only ROC AUC* $= 0.5976$ *and AUCPR* $= 0.0242$*. The available features (claim value, share capital, CNAE code, firm size, settlement value) lack the discriminative signal the task requires, a conclusion reinforced by boxplot and t-SNE analyses showing heavy class overlap. The paper contributes a transparent, reproducible benchmark for a difficult and rarely studied problem, as well as evidence-based recommendations for enriching judicial data pipelines with textual, temporal, and network-based features. We conclude that current models are not operationally ready and, until richer data become available, should at most assist judges under strict human oversight and ethical safeguards.*

Resumo. *As lides simuladas (simulated lawsuits), nas quais as partes agem em conluio para obter uma decisão favorável, impõem um ônus substancial e pouco examinado ao Judiciário brasileiro. Este artigo investiga se tais casos podem ser sinalizados automaticamente a partir de dados institucionais rotineiramente disponíveis. Avaliamos comparativamente uma ampla gama de métodos de aprendizado de máquina em 6,412 acordos da Justiça do Trabalho do Recife, dos quais apenas 1,75% são simulações confirmadas: classificadores supervisionados (KNN, árvores de decisão, florestas aleatórias, gradient boosting) combinados com sobreamostragem e subamostragem; quatro detectores de anomalias (Local Outlier Factor, Isolation Forest, One-Class SVM, Mahalanobis Robusto); e um pipeline de AutoML. A principal constatação é negativa, porém informativa: todos os métodos têm desempenho próximo ao acaso, com o melhor alcançando apenas ROC AUC* $= 0{,}5976$ *e AUCPR* $= 0{,}0242$*. As características disponíveis (valor da causa, capital social, código CNAE, porte da empresa, valor do acordo) carecem do sinal discriminativo que a tarefa exige, conclusão reforçada por análises de boxplot e t-SNE que evidenciam forte*

sobreposição entre as classes. O artigo contribui com um benchmark transparente e reprodutível para um problema difícil e raramente estudado, bem como com recomendações baseadas em evidências para enriquecer os fluxos de dados judiciais com características textuais, temporais e baseadas em redes. Concluímos que os modelos atuais não estão operacionalmente prontos e, até que dados mais ricos se tornem disponíveis, devem no máximo auxiliar os juízes sob rigorosa supervisão humana e salvaguardas éticas.

1. Introduction

A simulated lawsuit, known in Brazil as a *lide simulada*, is a fraudulent legal action in which two nominally opposing parties, typically an employer and a former employee, collude to have a court ratify a settlement they have arranged in advance. By passing a private agreement through the judiciary, the parties lend it a semblance of legality it would not earn under genuine adversarial review, circumventing labor-law guarantees while burdening the courts with cases that are not, in substance, disputes at all.

The institutional cost of this practice is substantial. Estimates compiled from judicial-press sources suggest that simulated lawsuits account for a sizeable share of all suits in major urban centers, contributing to the R$132.8 billion spent by the Brazilian judiciary in 2023 (Ferreira and de Matos 2024). The damage extends beyond direct expenditure: these cases distort official caseload statistics, harm legitimate creditors, and consume judicial capacity that should be reserved for genuine conflicts. Labor courts are especially vulnerable, because their settlement-driven workflow can be repurposed to convert informal arrangements between employers and ex-employees into court-sanctioned outcomes.

Detecting such cases is intrinsically difficult. Their fraudulent character resides in the *intent* of the parties rather than in any overt property of the petition or the settlement value, and whatever signal exists is diluted within very large case volumes. This combination of high volume, severe class imbalance, and a low signal-to-noise ratio makes the problem a natural candidate for machine learning, yet also a hostile environment for standard supervised baselines.

This paper reports an empirical investigation of that difficulty. We draw on a real dataset of 6,412 labor-court settlements from Recife (January 2019 to June 2022), labeled by court agents and enriched with public data from the Federal Revenue. With only 1.75% of cases labeled as simulated, we benchmark three families of methods that the literature recommends for imbalanced detection: (a) supervised classification, with and without resampling; (b) anomaly detection; and (c) AutoML, including deep learning. A single protocol and metric set is applied uniformly across all three, so that their relative behavior can be compared on equal footing.

The study is organized around two research questions. **RQ1:** Can simulated and legitimate settlements be distinguished using the institutional features currently recorded (primary economic activity, claim value, firm size, registered share capital, and settlement value)? **RQ2:** Among interpretable and AutoML-discovered models, which configurations perform best, and how large is the gap to a naive majority-class baseline?

The answer to RQ1, foreshadowed by the strong class overlap evident in our ex-

ploratory analysis, is largely negative. Across every model family, performance hovers near chance, and the best AUC obtained (0.5976) represents only a marginal gain over random guessing. Our contribution is therefore an *honest empirical map* of the difficulty rather than a positive detection result. Concretely, the paper offers:

1. The first empirical comparison, in this domain, of resampled supervised classification, anomaly detection, and AutoML/deep learning, all evaluated on the same dataset, under the same protocol, and against the same metric set (Accuracy, Precision, Recall, F1, ROC AUC, AUCPR).
2. Evidence that the currently available institutional feature set is insufficient for the task, supported by exploratory analysis (boxplots, t-SNE) and confirmed by consistent near-chance performance across all method families.
3. A reproducible benchmark and a roadmap for future detection systems, identifying which classes of features (the textual content of the petition, temporal patterns, and party–lawyer–firm network structure) are most likely to unlock useful signal, and which changes to the judicial data pipeline (e.g. richer labeling within the PJe system) would be required to support them.

Together, these contributions reframe the work as a *negative-result, lessons-learned* study rather than a performance-driven one, an academic genre of particular value in domains where deployment carries non-trivial ethical risk, as judicial decision support clearly does.

The remainder of the paper is structured as follows. Section 2 surveys related work on machine learning for fraud detection and judicial applications, with emphasis on imbalanced settings. Section 3 presents the consolidated methodology: dataset, models, resampling strategies, anomaly detectors, the AutoML configuration, the experimental protocol, and the evaluation metrics. Section 4 reports and discusses the results and limitations. Section 5 concludes and outlines the data and modeling work most likely to move the problem forward.

2. Related Work

Recent Brazilian academic and governmental efforts have applied machine learning (ML) to detect fraud in public procurement and tax evasion. Xavier (Xavier 2022) uses graph neural networks (GNNs) over transactional networks to flag shell companies in ICMS tax-credit schemes, while Gomes (Gomes 2024) shows that supervised models such as Random Forests and Neural Networks can identify suspicious patterns in tax documentation. Analyses of federal procurement data from 1990 to 2022 (Faria 2022; Gomes 2024) further indicate that non-competitive bidding processes and unusually high contract values are strong predictors of irregularity. These studies share two properties relevant to our setting: they exploit open-government data, and they cast fraud as a tabular classification problem.

Within the judiciary, ML systems have been deployed against predatory litigation and procedural fraud. The Pernambuco Court of Justice's *Bastião* project applies natural language processing to identify repetitive lawsuits and fraudulent document patterns, chiefly in consumer disputes (TJPE 2023). At the national level, the *Conselho Nacional de Justiça* (CNJ) has issued ethical guidelines for AI adoption that emphasize human oversight and bias mitigation (Barroso 2024; Conselho Nacional de Justiça (CNJ) 2025).

What these efforts have in common is a focus on petition text and case-flow metadata. *Collusive* litigation in labor courts, where the parties cooperate rather than oppose, has received considerably less attention, and it is precisely this gap that our work addresses.

The international literature offers a substantial methodological toolkit for problems of this type. Resampling techniques such as SMOTE (Chawla et al. 2002), Borderline-SMOTE, SVM-SMOTE, ADASYN, NearMiss, Tomek Links, and Edited Nearest Neighbors have been extensively benchmarked in surveys of imbalanced learning (Branco et al. 2023). Cost-sensitive learning provides an algorithmic alternative that encodes asymmetric error costs directly into the loss function (Baesens et al. 2021). For anomaly detection over tabular data, recent surveys (Blázquez-García et al. 2020) catalog density-based methods such as the Local Outlier Factor, isolation-based methods, kernel methods such as One-Class SVM, and statistical methods such as the Robust Mahalanobis distance (Hawkins 1980). We draw on all three of these strands in our benchmark.

A separate line of research warns that, under extreme imbalance, ROC AUC can remain visually unchanged even as performance varies sharply in the practically relevant region of the operating curve. Several authors therefore recommend the Precision–Recall curve and its area (AUCPR) as a more informative summary in this regime (Chawla et al. 2002; Branco et al. 2023). We adopt this recommendation in Section 4.

To the best of our knowledge, no published study has empirically compared supervised classification with resampling, anomaly detection, and AutoML/Deep Learning on a real, court-supplied dataset of *simulated labor settlements*. Existing Brazilian work targets either procurement and tax fraud (Xavier 2022; Gomes 2024; Faria 2022) or *predatory* consumer litigation (TJPE 2023), in which the adversarial structure of the case differs fundamentally from that of collusive labor settlements.

Our work departs from this prior research in three respects. First, we evaluate three method families on the *same* dataset and under the *same* protocol, enabling a direct comparison that single-method studies cannot offer. Second, we report a uniform set of metrics (including AUCPR and F1) across all families, rather than the per-family metrics typical in the literature. Third, and unlike most prior work, we report a partly negative empirical outcome and turn it into actionable guidance: instead of claiming detection success, we provide concrete evidence about which features the current judicial pipeline fails to capture and which classes of features future systems should prioritize.

3. Materials and Methods

This section describes the dataset and features, the exploratory analysis that motivates our methodological choices, and the modeling pipelines, experimental protocol, and evaluation metrics used throughout the study. A single dataset, cross-validation protocol, and metric set are applied across all pipelines to ensure that performance comparisons are not confounded by methodological differences.

3.1. Dataset

The dataset was compiled from labor-court settlements ratified by the Pernambuco regional labor court between January 2019 and June 2022. The original extraction contained 26,604 cases. After data restructuring, record-level cleaning, and enrichment with public information from the Brazilian Federal Revenue Service (e.g. registered share capital,

primary economic activity, and firm size), the modeling dataset consists of 6,412 cases from the Recife judicial district.

Each case is labeled by court agents as either *legitimate* (98.25%; 6,300 cases) or *simulated* (1.75%; 112 cases). The label is thus the outcome of human classification by court personnel and may itself be subject to underreporting: cases that an agent never flags as suspicious are recorded as legitimate by default.

Five features serve as input to the models, all derived either from court metadata or from the Federal Revenue enrichment; they are summarized in Table 1. We highlight one subtlety: the *Settlement Value* is used as an *input feature* describing the magnitude of the agreed amount, whereas the *target variable* is the binary label `simulated / legitimate` assigned by court agents.

Table 1. Features used for model training.

Feature	Description
Initial LC Value	Initial monetary value claimed in the petition (numerical).
Share Capital	Registered share capital of the defendant firm (numerical).
PEA (CNAE)	Primary economic activity of the defendant firm (categorical).
Company Size	Brazilian Federal Revenue size class of the defendant (categorical).
Settlement Value	Monetary value of the ratified settlement (numerical).

The exploratory analysis already anticipates the task's difficulty. Box plots of the numerical features (Figure 1) show that both the claim value and the share capital overlap substantially across the two classes: their interquartile ranges and medians are nearly indistinguishable. No simple threshold on these variables, taken in isolation, can therefore separate simulated from legitimate cases.

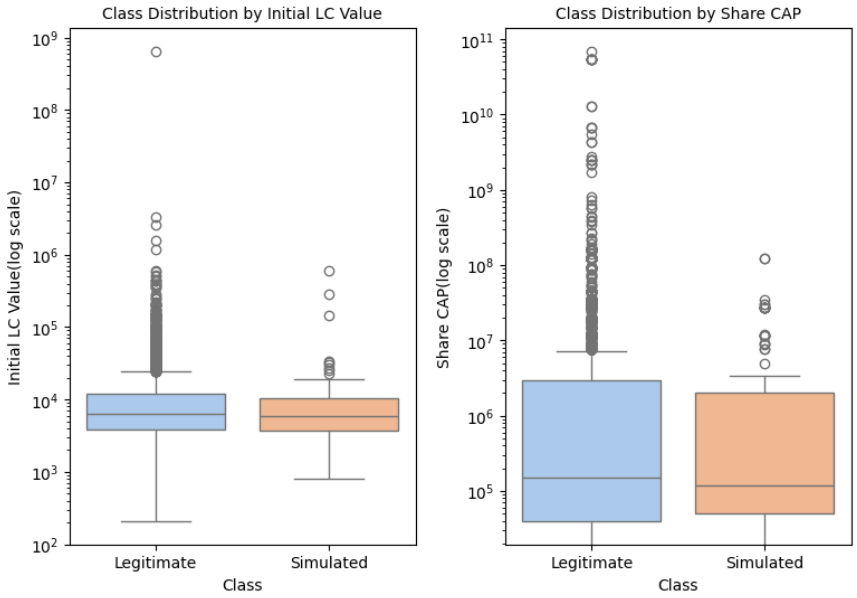


Figure 1. Box plots of *Initial LC Value* and *Share Capital* for legitimate (blue) and simulated (red) settlements.

The t-SNE projection of the full feature space (Figure 2) reinforces this observation. The minority class does not form a distinct cluster; its instances are diffusely scattered *inside* the dense cluster of legitimate cases. This high-dimensional overlap prefigures the limited performance later reported in Section 4 and motivates our framing of the work as an empirical study of feature-set limitations.

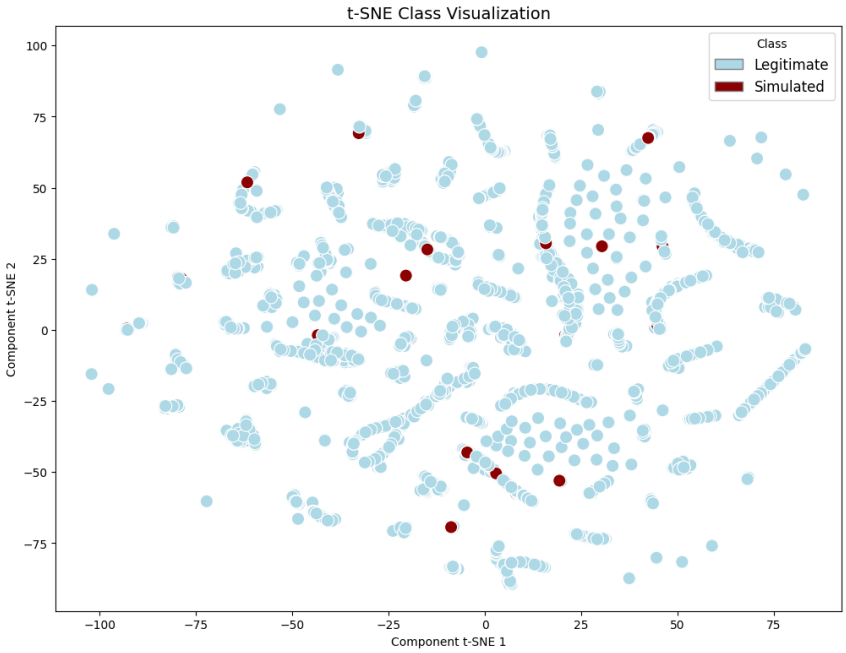


Figure 2. t-SNE projection of the 5-feature space. Categorical features (PEA, Company Size) were one-hot encoded and concatenated with the standardized numerical features. Perplexity = 30, learning rate = 200, 1,000 iterations, random seed fixed. Simulated cases (red) overlap heavily with legitimate cases (blue).

3.2. Models

Guided by the overlap revealed above, we adopt three families of methods routinely recommended for imbalanced binary detection, plus an AutoML baseline, and evaluate them all under an identical protocol.

Supervised classifiers. We use four classifiers chosen to balance interpretability and predictive power: K-Nearest Neighbors (KNN) (Aggarwal 2015), Decision Trees (Faceli et al. 2021), Random Forests (Wade 2020), and Gradient Boosting (with XGBoost as a strong representative (Wade 2020)). KNN and Decision Trees serve as proximity- and rule-based baselines, while Random Forests and Gradient Boosting add ensembling and provide the strong defaults widely reported for tabular fraud problems.

Resampling. To isolate the effect of rebalancing, we combine the four classifiers with four oversamplers (SMOTE (Chawla et al. 2002), Borderline-SMOTE, SVM-SMOTE, and ADASYN) and three undersamplers (NearMiss, Tomek Links, and Edited Nearest Neighbors (Branco et al. 2023; Zhang and Mani 2003)). Each (model, resampler) pair is trained and evaluated independently to observe the marginal effect of resampling.

Anomaly detection. Treating simulated cases as anomalies, we benchmark four detectors spanning density-, isolation-, kernel-, and statistical-based families: Local Outlier Factor (LOF), Isolation Forest, One-Class SVM (Blázquez-García et al. 2020), and the Robust Mahalanobis distance (Hawkins 1980). These methods are fitted with no access to the binary label and then scored on the test fold.

AutoML. Finally, to check whether the choices above leave performance on the table, we run an AutoML pipeline using `H2O.ai` with its default search space, which includes Deep Learning, Gradient Boosting, Generalized Linear Models, and stacked ensembles. AutoML serves as a strong upper-bound reference rather than as the proposed solution.

A compact summary of all hyperparameter choices appears in Table 2. We use default scikit-learn, imbalanced-learn, and H2O.ai parameters unless explicitly noted to preserve comparability, and we report those defaults explicitly so that the protocol is reproducible.

Table 2. Hyperparameters of all pipelines. "†" = library default at the version listed in Section 3.

Family	Method	Settings
Classifiers	KNN	$k = 5$ (def.); Euclidean distance; uniform weights.
	Decision Tree	Gini criterion †; no max depth †.
	Random Forest	100 trees †; Gini; bootstrap †.
	Gradient Boosting/XGB	100 estimators (def.); learning rate 0.1 †.
Oversamplers	SMOTE / Borderl.-SMOTE / SVM-SMOTE	$k_{\text{neigh}} = 5$ †.
	ADASYN	$n_{\text{neigh}} = 5$ †.
Undersamplers	NearMiss	Version 1 †.
	Tomek Links	†
	Edited NN	$k = 3$ †.
Anomaly	LOF	$n_{\text{neigh}} = 20$ †; contamination = prior rate.
	Isolation Forest	100 trees †; contamination = prior rate.
	One-Class SVM	RBF kernel; ν = prior rate.
	Robust Mahalanobis	Minimum Covariance Determinant †.
AutoML	H2O.ai	Default search, max_models=20, seed=1.

Experiments are run in Python 3.11 with `scikit-learn` 1.4, `imbalanced-learn` 0.12, and `H2O.ai` 3.46. The random seed is fixed at `seed = 42` for all scikit-learn and imbalanced-learn pipelines and at `seed = 1` for H2O.ai.

For the classification and resampling pipelines, we use stratified 5-fold cross-validation, which preserves the class ratio within each fold; each fold serves once as the test set while the remaining four are used for training. Every preprocessing step with a data-driven component (standardization, encoding, resampling) is fitted only to the training fold and then applied to the test fold. In particular, oversampling and undersampling are applied *after* the fold split, never on the full dataset, which avoids the well-known leakage that arises when synthetic minority points are generated from neighbors that include test instances.

For LOF, Isolation Forest, One-Class SVM, and Robust Mahalanobis, the contamination parameter is fixed to the empirical prior rate (1.75%). The label is not used during fitting and serves only to evaluate the resulting anomaly scores against the binary ground truth. The H2O.ai AutoML experiments use the same 5-fold cross-validation configuration, the same data splits, and a maximum of 20 models, with AUC as the internal selection metric.

3.3. Evaluation metrics

We evaluate test performance using the standard binary classification metrics: Accuracy, Precision, Recall, F1-Score, ROC AUC (area under the receiver operating characteristic curve), and AUCPR (area under the Precision–Recall curve).

We pay particular attention to Recall, the fraction of simulated cases actually retrieved, and to AUCPR, which quantifies whether a model can rank simulated cases above legitimate ones across operating points. Accuracy is reported only as a baseline reference: under a 98.25%/1.75% prior, a trivial `always-legitimate` classifier already attains 0.9825. We make this naive baseline explicit in Section 4.

4. Results and Discussion

We report the empirical results in four blocks (baseline classification, classification with resampling, anomaly detection, and AutoML), all obtained under the unified protocol of Section 3.2. Each block is presented as a table of averaged cross-validation metrics with standard deviations, followed by a discussion of the overall picture relative to the naive majority-class baseline.

4.1. Naive baseline

Under a 98.25%/1.75% prior, the trivial classifier that always predicts the majority class attains Accuracy ≈ 0.9825 with Precision $=$ Recall $= F_1 = 0$ and ROC AUC $= 0.5$. To improve on this baseline, a model must raise at least one of Precision, Recall, F_1, ROC AUC, or AUCPR *without* driving the others to non-operational levels. This baseline is the reference against which every subsequent table should be read, and it makes clear that high Accuracy is, on its own, *not informative*.

4.2. Baseline classification (no resampling)

The pattern in Table 3 is uniform: the classifiers reproduce the naive baseline almost exactly. Accuracy hovers around 0.9825 because the models effectively predict `legitimate` for every instance, and Precision, Recall, and F_1 collapse to zero. ROC AUC values slightly above 0.5 indicate marginal ranking ability, but the standard deviations are large enough (0.07–0.12) that the ordering among models is not robust.

4.3. Classification with resampling

Table 4 reports the most informative (model, resampler) pairs, which fall into two operating regimes. *Borderline-SMOTE* variants preserve accuracy at the cost of very low recall (≤ 0.15), since they barely shift the operating point. *NearMiss* undersampling does the opposite: it pushes recall above 0.92 but collapses accuracy below 0.10, meaning the classifier now predicts *simulated* for almost every test instance.

Table 3. Baseline classification results (no resampling), 5-fold cross-validation averages and standard deviations.

Model	Accur.	Prec.	Recall	F_1	ROC AUC	σ AUC
Gradient Boosting	0.9824	0.0000	0.0000	0.0000	0.5809	0.1226
Random Forest	0.9825	0.0000	0.0000	0.0000	0.5778	0.1103
XGBoost	0.9825	0.0000	0.0000	0.0000	0.5578	0.1074
Decision Tree	0.9825	0.0000	0.0000	0.0000	0.5528	0.1101
KNN	0.9810	0.0000	0.0000	0.0000	0.5225	0.0758

Neither regime is operationally useful: the first misses nearly every fraudulent case, while the second floods the court with false positives. The intermediate SMOTE variants yield only a small, non-significant change in ROC AUC relative to the no-resampling baseline. Because the per-fold standard deviations (around 0.10 on AUC) exceed the gap between any two configurations, paired statistical tests (e.g., Wilcoxon) on the fold-level scores would not, on the available evidence, support any specific ranking.

Table 4. Classification with resampling, 5-fold cross-validation averages and standard deviations.

Resampler / Model	Accur.	Recall	ROC AUC	σ AUC
Borderl.-SMOTE / Random Forest	0.8648	0.1150	0.5596	0.1195
SMOTE / Random Forest	0.7909	0.2450	0.5552	0.1030
SMOTE / KNN	0.8275	0.1767	0.5308	0.1048
Borderl.-SMOTE / KNN	0.8601	0.1500	0.5185	0.0956
SMOTE / Decision Tree	0.7533	0.2767	0.5065	0.1073
NearMiss / Decision Tree	0.0906	0.9283	0.5004	0.0500
NearMiss / KNN	0.1840	0.8600	0.4870	0.1047
Borderl.-SMOTE / Decision Tree	0.8759	0.1417	0.4817	0.1061
NearMiss / Random Forest	0.0834	0.9833	0.4737	0.1164

4.4. Anomaly detection

All four detectors sit at or below the random-ranking line (Table 5). The strongest, Isolation Forest, reaches a ROC AUC of 0.5057, which is statistically indistinguishable from chance given the reported standard deviations. The hypothesis that simulated settlements behave as low-density anomalies in this feature space is therefore not supported by the data.

Table 5. Anomaly detection results, 5-fold cross-validation averages and standard deviations. F_1 computed from the reported Precision and Recall.

Model	Accur.	Prec.	Recall	F_1	ROC AUC	σ AUC
LOF	0.9649	0.0000	0.0000	0.0000	0.4911	0.0022
Isolation Forest	0.9651	0.0274	0.0295	0.0284	0.5057	0.0132
One-Class SVM	0.9596	0.0201	0.0195	0.0198	0.4980	0.0184
Robust Mahalanobis	0.9658	0.0095	0.0095	0.0095	0.4962	0.0105

4.5. AutoML

Table 6 reports the top-6 models found by the H2O.ai AutoML pipeline, ranked by ROC AUC. The best model is a Deep Learning configuration with ROC AUC = 0.5976 and AUCPR = 0.0242. This is the highest ROC AUC across the entire study, yet the absolute margin to the second-best AutoML model is only 0.0126, and to the best classical classifier (Gradient Boosting, at 0.5809) only 0.0167; both gaps fall well below the per-fold standard deviation observed in the other experiments. The AUCPR tells the same story: at 0.0242, it is barely above the prior of 0.0175, the AUCPR of a uniformly random ranker on this class distribution, and orders of magnitude below the level at which the model would be operationally trustworthy. We therefore present the AutoML result as the *best obtainable upper bound under the current feature set*, not as a deployable detector.

Table 6. Top-6 AutoML models. Threshold-dependent metrics (Precision, Recall, F_1) are omitted because the AutoML interface does not expose a single operating point.

Model	ROC AUC	AUCPR	MPCE
DeepLearning_grid_1_AutoML_1_model_2	0.5976	0.0242	0.4087
StackedEnsemble_AllModels_5_AutoML_1	0.5850	0.0257	0.4756
StackedEnsemble_BestOfFamily_7_AutoML_1	0.5724	0.0218	0.4349
DeepLearning_grid_3_AutoML_1_model_5	0.5694	0.0224	0.4600
StackedEnsemble_BestOfFamily_5_AutoML_1	0.5686	0.0240	0.4713
GBM_grid_1_AutoML_1_model_8	0.5646	0.0210	0.4438

4.6. Discussion

The four blocks of results converge on a single message: with the five institutional features available, no method family separates simulated from legitimate cases meaningfully better than chance. Resampling can shift the operating point but adds no new information; anomaly detection finds nothing distinctive about simulated cases in this feature space; and AutoML, despite a far larger search budget, yields only a marginal AUC gain that lies within the noise of the cross-validation. This outcome is fully consistent with the exploratory analysis of Section 3.1, where the strong class overlap in the box plots and t-SNE projection already *anticipated* the ceiling we now observe. Three conclusions follow.

1. **High accuracy is misleading in this regime.** Almost every Accuracy figure above 0.96 reflects the prior distribution rather than any modeling skill. Accuracy should be de-emphasized in this domain in favor of AUCPR and operational-utility metrics.
2. **The bottleneck is the feature set, not the model.** The AutoML upper bound is essentially identical to the best classical configuration despite Deep Learning's much larger capacity, which is the clearest empirical signal that further modeling effort on the same features will yield diminishing returns. The features most likely to break this ceiling (textual, temporal, and network-based) are discussed below as threats to validity.
3. **Operational utility is low even at the best operating point.** Any deployed model would route to judicial review a number of legitimate cases several orders of mag-

nitude larger than the number of simulated cases caught. A Top-k Precision analysis, asking "among the k most suspicious cases, how many are actually simulated?", is the natural way to quantify this cost asymmetry, and we recommend that future versions of this study report it as the primary deployment-oriented metric.

On the basis of these results, the present model is not ready for operational deployment and should be used, if at all, only as a single signal among several in a human-led review pipeline.

Because this paper's contribution is empirical, the limitations of the evidence shape what we can claim. **Label noise.** The label "simulated" was assigned by court agents during conciliation hearings, so uninspected cases are recorded as "legitimate" by default. The 1.75% positive rate is therefore a *lower bound* on true prevalence, and since ROC AUC and AUCPR assume a clean negative class, the real performance ceiling is likely *lower* than our modest values. **Scope of the feature set.** Section 3.1 and the results above show that the five institutional features do not separate the two classes, the paper's main finding. Our conclusions thus apply to detection *from current institutional metadata only*, and do not rule out detection from richer data, such as petition text, case-filing timing, or the lawyer–plaintiff–firm network. **Evaluation protocol.** We use random stratified cross-validation, whereas the forward-in-time deployment scenario calls for a temporal split (e.g., train on 2019–2021, test on 2022); we list this as a priority for future work. **Statistical power.** With only five folds and per-fold standard deviations around 0.10 on ROC AUC, the confidence intervals overlap for many configurations, so the data cannot support strong claims of one being better than another; paired Wilcoxon or McNemar tests would be appropriate once the saved predictions are recompiled.

5. Conclusion

We presented an empirical study of simulated-litigation detection in Brazilian labor courts, benchmarking four supervised classifiers (with and without resampling), four anomaly detectors, and a Deep-Learning-enabled AutoML pipeline on a real dataset of 6,412 settlements from Recife, all under a unified protocol.

Across all method families, performance remains close to chance: the baseline classifiers reproduce the naive majority-class predictor; resampling only shifts the operating point; the anomaly detectors yield ROC AUC values indistinguishable from 0.5; and the best AutoML model reaches just ROC AUC = 0.5976 and AUCPR = 0.0242. This convergence is itself the central finding: the bottleneck is the feature set, not the model. At any practical operating point, such a model would flag far more legitimate cases than simulated ones, which is incompatible with the due-process guarantees of conciliation hearings. The study nonetheless identifies precisely which institutional data gaps must be closed before automated assistance can be embedded responsibly.

Three directions follow. First, textual features from the initial petition (e.g., TF-IDF or fine-tuned BERTimbau encodings) may carry a far richer signal than institutional metadata. Second, network features from the co-occurrence graph among plaintiffs, lawyers, and firms could capture collusive recurrence that is invisible to single-case features. Third, temporal features of the case lifecycle (time-to-settlement, law-firm recurrence patterns) may reveal dynamics a static snapshot misses. Together, these constitute

the most promising path beyond the ceiling documented here.

References

[Aggarwal 2015] Aggarwal, C. C. (2015). *Data Classification: Algorithms and Applications*. CRC Press, Taylor & Francis Group, Boca Raton.

[Baesens et al. 2021] Baesens, B., Höppner, S., and Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems*, 150:113492.

[Barroso 2024] Barroso, L. R. (2024). Inteligência artificial: Promessas e riscos. *Revista Direito Público*.

[Blázquez-García et al. 2020] Blázquez-García, A. et al. (2020). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*. arXiv preprint, Accessed: 15 December 2024.

[Branco et al. 2023] Branco, P., Ribeiro, R., and Torgo, L. (2023). Ennclassif: Edited nearest neighbor for multiclass imbalanced problems. `https://rdrr.io/cran/UBL/man/ENNClassif.html`. Accessed: 23 November 2024.

[Chawla et al. 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Conselho Nacional de Justiça (CNJ) 2025] Conselho Nacional de Justiça (CNJ) (2025). Resolução sobre uso ético de ia no judiciário. Available at: `https://www.cnj.jus.br/`.

[Faceli et al. 2021] Faceli, K. et al. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, Rio de Janeiro.

[Faria 2022] Faria, G. A. (2022). Detecção de fraudes em compras governamentais via aprendizado de máquina.

[Ferreira and de Matos 2024] Ferreira, V. and de Matos, V. P. (2024). Litigância abusiva no brasil: Criação de fraudes e erosão da confiança. Available at: `https://www.migalhas.com.br/depeso/420645/litigancia-abusiva-no-brasil-criacao-de-fraudes\-e-erosao-da-confianca` Accessed: 2025-05-13.

[Gomes 2024] Gomes, G. S. L. (2024). Machine learning para créditos ilegais de icms. *Revista de Administração Pública (RAP)*.

[Hawkins 1980] Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall, London.

[TJPE 2023] TJPE (2023). Projeto bastião: Relatório técnico. Technical report, TJPE. Available at: `https://www.tjpe.jus.br/`.

[Wade 2020] Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn*. Packt Publishing, Birmingham.

[Xavier 2022] Xavier, O. C. (2022). Identificação de evasão fiscal com redes neurais. *SciELO*.

[Zhang and Mani 2003] Zhang, J. and Mani, I. (2003). knn approach to unbalanced data distributions: A case study involving information extraction. In *Proceedings of the ICML Workshop on Learning from Imbalanced Datasets II*, pages 42–48, Washington DC. AAAI Press.