




# Semi-Supervised Supply Chain Fraud Detection with Unsupervised Pre-Filtering

Fatemeh Moradi   
Faculty of Engineering  
Isfahan (Khorasgan) Branch,  
Islamic Azad University  
Isfahan, Iran

Mehran Tarif   
Department of Computer Science  
University of Verona  
Verona, Italy

Mohammadhossein Homaei   
Media Engineering Group  
University of Extremadura  
Cáceres, Spain

**Abstract**—Detecting fraud in modern supply chains is a growing challenge, driven by the complexity of global networks and the scarcity of labeled data. Traditional detection methods often struggle with class imbalance and limited supervision, reducing their effectiveness in real-world applications. This paper proposes a novel two-phase learning framework to address these challenges. In the first phase, the Isolation Forest algorithm performs unsupervised anomaly detection to identify potential fraud cases and reduce the volume of data requiring further analysis. In the second phase, a self-training Support Vector Machine (SVM) refines the predictions using both labeled and high-confidence pseudo-labeled samples, enabling robust semi-supervised learning. The proposed method is evaluated on the DataCo Smart Supply Chain Dataset, a comprehensive real-world supply chain dataset with fraud indicators. It achieves an F1-score of 0.817 while maintaining a false positive rate below 3.0%. These results demonstrate the effectiveness and efficiency of combining unsupervised pre-filtering with semi-supervised refinement for supply chain fraud detection under real-world constraints, though we acknowledge limitations regarding concept drift and the need for comparison with deep learning approaches.

**Keywords:** Supply chain fraud detection, Isolation Forest, Self-training SVM, Semi-supervised learning, Anomaly detection, Class imbalance

## I. INTRODUCTION

Supply chain fraud has become a critical threat to global commerce, with organizations facing increasingly sophisticated fraudulent schemes across complex supply networks [1]. The digitalization of supply chains has created new vulnerabilities, enabling various fraudulent activities including procurement fraud, vendor impersonation, invoice manipulation, and counterfeit goods infiltration [2]. These activities result in substantial financial losses and operational disruptions for businesses worldwide [3]. The interconnected nature of modern supply chains amplifies fraud impact, as single fraudulent events can cascade across multiple organizations and geographical regions, while increasing transaction complexity challenges traditional real-time monitoring systems. Modern supply chains involve multiple stakeholders, heterogeneous data sources, and intricate networks that complicate fraud detection [4]. The inherent class imbalance and sophisticated fraud schemes have rendered traditional rule-based systems inadequate [5], [6]. Zhou et al. [4] demonstrated XGBoost's effectiveness for supply chain fraud prediction, emphasizing

feature engineering and ensemble methods. The DataCo Supply Chain Dataset [7] has become a valuable benchmark for fraud detection research. Constante-Nicolalde et al. [2] explored smart supply chain fraud prediction with IoT integration. Baryannis et al. [3] examined the performance-interpretability trade-off in machine learning for supply chain risk prediction.

To address these limitations, innovative fraud detection systems must effectively handle limited labeled data while maintaining high detection accuracy and computational efficiency through hybrid approaches combining multiple learning paradigms. Prior work has explored artificial intelligence's role in enhancing cybersecurity across digital infrastructures, including digital twin systems, emphasizing hybrid AI techniques for detecting complex and evolving threats [9]–[11]. This paper proposes a novel two-phase learning model combining unsupervised anomaly detection with semi-supervised learning refinement. The first phase employs Isolation Forest for efficient outlier identification without requiring labeled training data [8]. The second phase utilizes self-training Support Vector Machine to refine detection results by iteratively expanding the labeled dataset with high-confidence predictions [12]. This approach addresses key challenges including computational efficiency, class imbalance handling, and effective utilization of limited labeled data in supply chain fraud detection systems. The rest of this paper is organized as follows. Section II reviews recent related works on fraud detection in supply chains and machine learning methods. Section III explains the proposed two-phase model, including Isolation Forest and self-training SVM. Section IV describes the experimental setup with datasets, evaluation metrics, and baseline methods. Section V presents and discusses the results. Finally, Section VI gives the conclusion and suggestions for future work.

## II. RELATED WORK

### A. Supply Chain Fraud Detection

Supply chain fraud detection has gained significant attention due to increasing complexity and digitalization of global networks. Modrušan et al. [1] reviewed public procurement fraud detection techniques, highlighting evolution from rule-based to machine learning approaches and identifying key challenges:

data heterogeneity, real-time processing, and sophisticated fraud schemes. Zhou et al. [4] demonstrated XGBoost’s effectiveness for supply chain fraud prediction, while the DataCo dataset [7] has become a valuable benchmark. Recent work has explored IoT integration [2] and performance-interpretability trade-offs [3].

### B. Machine Learning Approaches for Fraud Detection

Machine learning applications in fraud detection have been extensively surveyed. Hernández Aros et al. [13] reviewed financial fraud detection literature, analyzing 104 articles and identifying Random Forest and Autoencoder as particularly effective techniques. Phua et al. [6] provided a comprehensive survey of data mining-based fraud detection across multiple domains. Recent work has applied semi-supervised learning using Isolation Forests to effectively detect fraud in supply chain data without full supervision [14].

### C. Class Imbalance in Fraud Detection

Class imbalance represents a significant challenge in fraud detection. Bauder and Khoshgoftaar [5] investigated varying class distribution effects on learner behavior for Medicare fraud detection, demonstrating that unsupervised learning approaches can offer advantages with severely imbalanced datasets [15]. Wei et al. [16] addressed sophisticated online banking fraud detection on extremely imbalanced data (0.1% fraud rate), combining multiple techniques to handle extreme imbalance while maintaining high detection accuracy.

### D. Isolation Forest and Anomaly Detection

Isolation Forest [8] represents a paradigm shift in anomaly detection, using isolation principles rather than distance or density-based measures. The algorithm’s insight that anomalies are “few and different” enables efficient detection with  $O(n \log n)$  complexity, suitable for large-scale supply chain fraud detection. Liu et al. [17] provided detailed analysis of the algorithm’s performance characteristics and robustness properties. Hariri et al. [18] proposed Extended Isolation Forest, addressing bias issues by using hyperplanes with random slopes, improving detection consistency and accuracy.

### E. Semi-supervised Learning and Self-training

Semi-supervised learning approaches show promise for fraud detection with scarce labeled data. Wang et al. [19] developed a semi-supervised graph attentive network for financial fraud detection with substantial improvements. Hyun et al. [20] proposed Suppressed Consistency Loss (SCL) to handle distribution differences between labeled and unlabeled data. Wei et al. [21] introduced CReST for imbalanced semi-supervised learning, achieving 11.8% improvement over Fix-Match. Amini et al. [12] surveyed self-training methodologies, providing guidance for selecting appropriate strategies. One-Class SVM has been employed to model normal transaction behavior in supply chains [22].

### F. Research Gap

Despite extensive research in fraud detection, existing approaches face three critical limitations in supply chain contexts: (1) supervised methods require extensive labeled data that is costly to obtain, (2) unsupervised methods suffer from high false positive rates when used in isolation, and (3) current semi-supervised approaches do not address the computational scalability required for real-time supply chain monitoring. Our work addresses this gap by proposing a computationally efficient two-phase framework that combines the strengths of unsupervised and semi-supervised learning while maintaining practical deployment feasibility.

## III. METHODOLOGY

### A. Problem Formulation

Let  $\mathcal{X} = \{x_i\}_{i=1}^n$  represent the complete supply chain transaction dataset (such as the DataCo dataset [7]), where  $x_i \in \mathbb{R}^d$  denotes the  $d$ -dimensional feature vector for transaction  $i$ , and  $n$  is the total number of transactions. Each transaction has an associated true label  $y_i \in \{0, 1\}$  (0 for legitimate, 1 for fraudulent), but these labels are only observed for a small subset of the data.

The dataset can be partitioned into two disjoint subsets based on label availability:

- Labeled subset:  $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{n_L}$ , where  $n_L \ll n$
- Unlabeled subset:  $\mathcal{D}_U = \{x_i\}_{i=n_L+1}^n$ , where labels exist but are unobserved

We denote the complete dataset as  $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$ , where  $|\mathcal{D}_L| = n_L$  and  $|\mathcal{D}_U| = n - n_L$ . In practical supply chain scenarios, the dataset exhibits severe class imbalance with  $|\{i : y_i = 1, (x_i, y_i) \in \mathcal{D}_L\}| \ll |\{i : y_i = 0, (x_i, y_i) \in \mathcal{D}_L\}|$ , where fraudulent transactions constitute a small minority of the labeled data.

The objective is to learn a classifier  $f : \mathbb{R}^d \rightarrow \{0, 1\}$  that effectively identifies fraudulent transactions across the entire dataset  $\mathcal{X}$ , leveraging both the limited labeled data in  $\mathcal{D}_L$  and the abundant unlabeled data in  $\mathcal{D}_U$ , while minimizing false positives and maintaining computational efficiency for real-time processing requirements.

The algorithm terminates when either the F1-score improvement between iterations ( $\Delta F1_t$ ) falls below 0.001 or the maximum number of iterations (10) is reached, ensuring convergence while preventing overfitting.

### B. Two-Phase Framework Overview

Our proposed methodology consists of two sequential phases designed to address the key challenges in supply chain fraud detection: computational scalability, class imbalance, and limited labeled data availability. The framework architecture is illustrated in Algorithm 1.

The complete two-phase learning framework is illustrated in Figure 1, which provides a comprehensive overview of our proposed methodology. The framework begins with the DataCo dataset containing 180,519 transactions with a 1.5% fraud rate. The data is partitioned into labeled (10%) and

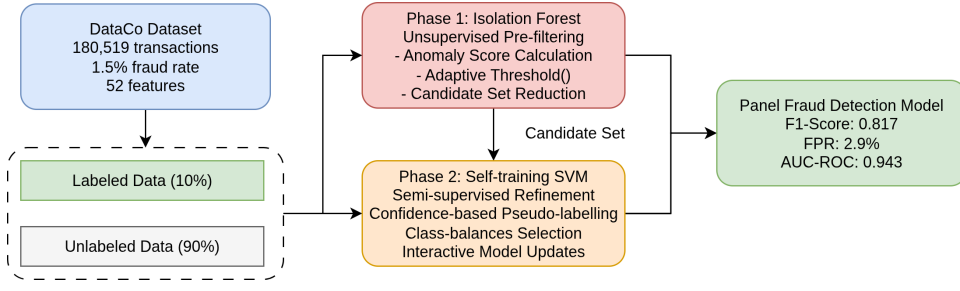


Fig. 1. Two-Phase Learning Framework combining Isolation Forest pre-filtering with self-training SVM refinement, achieving F1-score of 0.817 and 2.9% false positive rate on DataCo dataset.

### Algorithm 1 Two-Phase Learning Framework

**Require:** Dataset  $\mathcal{D}$ , labeled subset  $\mathcal{D}_L$ , unlabeled subset  $\mathcal{D}_U$ , parameters  $\alpha, \theta_{base}, \beta$

**Ensure:** Refined fraud detection model  $f_{final}$

- 1: **Phase 1: Isolation Forest Pre-filtering**
- 2: Train Isolation Forest model  $IF$  on entire dataset  $\mathcal{D}$
- 3: Compute anomaly scores:  $s_i = IF(x_i)$  for all  $x_i \in \mathcal{D}$
- 4: Calculate threshold:  $\tau = \mu_s + \alpha\sigma_s$
- 5: Create candidate set:  $\mathcal{D}_{candidates} = \{x_i \in \mathcal{D}_U : s_i \geq \tau\}$
- 6: **Phase 2: Self-training SVM Refinement**
- 7: Initialize SVM classifier  $SVM_0$  using labeled data  $\mathcal{D}_L$
- 8:  $\mathcal{D}_L^{(0)} \leftarrow \mathcal{D}_L, t \leftarrow 0$
- 9: **while**  $\Delta F1_t \geq 0.001$  and  $t < 10$  **do**
- 10:   Predict on candidates:  $\hat{y}_i = SVM_t(x_i)$  for  $x_i \in \mathcal{D}_{candidates}$
- 11:   Compute confidence:  $c_i = |f(x_i)| / \max_j |f(x_j)|$
- 12:   Calculate class-specific thresholds using Eq. 6
- 13:   Select high-confidence:  $\mathcal{P}_t = \{(x_i, \hat{y}_i) : c_i \geq \theta_{\hat{y}_i}\}$
- 14:   Update labeled set:  $\mathcal{D}_L^{(t+1)} = \mathcal{D}_L^{(t)} \cup \mathcal{P}_t$
- 15:   Remove from candidates:  $\mathcal{D}_{candidates} \leftarrow \mathcal{D}_{candidates} \setminus \mathcal{P}_t$
- 16:   Retrain:  $SVM_{t+1}$  on  $\mathcal{D}_L^{(t+1)}$  with class weights
- 17:    $t \leftarrow t + 1$
- 18: **end while**
- 19: **return**  $f_{final} = SVM_t$

unlabeled (90%) subsets to simulate realistic semi-supervised scenarios. Phase 1 employs Isolation Forest for unsupervised pre-filtering to identify potential fraud candidates, while Phase 2 utilizes self-training SVM for semi-supervised refinement through confidence-based pseudo-labeling and iterative model updates (Figure 1).

#### C. Phase 1: Isolation Forest Pre-filtering

The Isolation Forest algorithm, proposed by Liu et al. [8], operates on the principle that anomalies are easier to isolate than normal instances. For a given transaction  $x_i$ , the anomaly score is computed as:

$$s(x_i) = 2^{-\frac{E(h(x_i))}{c(n)}} \quad (1)$$

where  $E(h(x_i))$  represents the average path length of  $x_i$  over all isolation trees, and  $c(n)$  is the average path length of unsuccessful search in a Binary Search Tree (BST) with  $n$  points:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (2)$$

The threshold selection strategy employs an adaptive threshold approach:

$$\tau = \mu_s + \alpha\sigma_s \quad (3)$$

where  $\mu_s$  and  $\sigma_s$  represent the mean and standard deviation of anomaly scores, respectively, and  $\alpha$  is a sensitivity parameter.

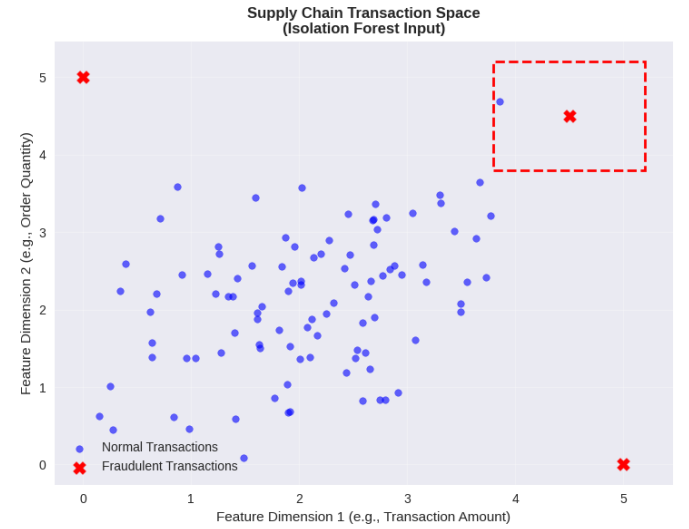


Fig. 2. As an example, the Supply Chain Transaction Space showing normal transactions (blue circles) clustering densely while fraudulent transactions (red X) appear as isolated outliers.

Figure 2 illustrates the conceptual foundation of Isolation Forest applied to supply chain transactions. Normal transactions form dense clusters, while fraudulent transactions appear as isolated outliers requiring fewer partitioning steps for separation.

#### D. Phase 2: Self-training SVM Refinement

The second phase employs a self-training Support Vector Machine (SVM) approach to refine the fraud detection results using both labeled and unlabeled data. The SVM classifier is trained to learn the decision boundary:

$$f(x) = \text{sign} \left( \sum_{i=1}^{n_L} \alpha_i y_i K(x_i, x) + b \right) \quad (4)$$

The confidence measure for pseudo-labeling is based on the distance from the decision boundary:

$$c(x_i) = \frac{|f(x_i)|}{\max_j |f(x_j)|} \quad (5)$$

To address class imbalance during self-training, we employ a balanced pseudo-labeling strategy inspired by the work of Wei et al. [21]:

$$\theta_c = \theta_{base} + \beta \cdot \log \left( \frac{N_c}{N_{target}} \right) \quad (6)$$

where  $N_c$  is the current number of pseudo-labeled samples for class  $c$ , and  $N_{target}$  is the desired target number of samples for balanced representation.

We used an SVM with a radial basis function (RBF) kernel, which is well-suited for capturing nonlinear patterns in complex fraud detection tasks. The kernel parameter  $\gamma$  and the regularization parameter  $C$  were selected using grid search with 5-fold cross-validation performed exclusively on the labeled training subset within each fold of the main 10-fold evaluation. This nested cross-validation approach prevents data leakage while ensuring robust hyperparameter selection. The search space was defined as  $C \in \{0.1, 1, 10, 100\}$  and  $\gamma \in \{0.001, 0.01, 0.1, 1\}$ . The combination yielding the highest average F1-score was chosen for final model training.

#### IV. EXPERIMENTAL SETUP

All experiments were conducted on Google Colab's free tier (Intel Xeon 2.20GHz CPU, 12.7GB RAM) without GPU requirements, ensuring accessibility for reproduction.

##### A. Dataset

We evaluated our proposed two-phase learning model on the DataCo Smart Supply Chain Dataset [7], containing 180,519 transactions (2015-2018) with 1.5% fraud rate and 52 features covering comprehensive supply chain operations across multiple countries. We employed stratified 10-fold cross-validation with 10% labeled data per fold, maintaining original class distribution, along with standard feature preprocessing including missing value imputation, categorical encoding, and numerical standardization.

##### B. Evaluation Metrics

Given the class imbalance inherent in fraud detection, we employed multiple evaluation metrics (Precision, Recall, F1-Score, AUC-ROC, AUC-PR, FPR) with statistical significance testing via Wilcoxon signed-rank test. We compared our approach against six baselines: Isolation Forest [8], SVM, Random Forest [13], XGBoost [4], Semi-supervised SVM, and Ensemble Method [23]. The Semi-supervised SVM baseline differs from our Phase 2 by operating without pre-filtering and using fixed confidence thresholds instead of our adaptive class-balanced approach (Eq. 6). All methods used identical 10% labeled data splits, prioritizing interpretable methods suitable for real-time deployment.

##### C. Experimental Protocol

1) *Labeled/Unlabeled Data Split*: For our semi-supervised learning experiments, we simulated realistic scenarios with limited label availability. Within each fold of the 10-fold cross-validation, the training portion (approximately 162,467 transactions per fold) was split such that 10% served as labeled data ( $\mathcal{D}_L$  with 16,247 samples per fold), while the remaining 90% (146,220 transactions per fold) formed the unlabeled set  $\mathcal{D}_U$ . The labeled subset maintained the original class distribution with approximately 1.5% fraud rate. To assess the robustness of our approach under varying supervision levels, we repeated experiments with 5%, 10%, and 20% labeled data ratios within each fold.

2) *Hyperparameter Settings*: The following hyperparameters were determined through systematic experimentation on a validation subset. For the Isolation Forest parameters, we set the number of trees to 100 with a subsample size of 256. The contamination factor was strategically set to 0.05 (5%) rather than the dataset's actual fraud rate of 1.5% to account for potential underreporting of fraud cases and to improve recall by capturing borderline anomalous transactions that may represent sophisticated fraud attempts. The sensitivity parameter  $\alpha = 1.5$  (Eq. 3) was chosen to balance between false positives and detection accuracy.

For self-training SVM, we set  $\theta_{base} = 0.85$ ,  $\beta = 0.3$  (Eq. 6), and  $N_{target} = 0.5 \times |\mathcal{D}_L|$  per class. The algorithm runs for maximum 10 iterations with convergence when  $|\mathcal{P}_t| < 50$  or F1-score change  $< 0.001$ .

3) *Feature Engineering Pipeline*: The 52 features from the DataCo dataset underwent comprehensive preprocessing through a four-stage pipeline.

First, missing value imputation used domain-specific strategies: median values within product categories for numerical features, mode imputation with "Unknown" category for rare categorical values, and forward-fill for sequential missing dates in temporal features.

Second, feature encoding included one-hot encoding for categorical variables with  $\leq 10$  unique values, target encoding for high-cardinality features (e.g., Customer City), and cyclical encoding for temporal features (day of week, month) to preserve circularity.

Third, feature scaling normalized continuous features to zero mean and unit variance using StandardScaler, percentage-based features to 0-1 range using MinMaxScaler, and log-transformed skewed financial metrics (sales, profit) to handle outliers.

Finally, feature selection removed 3 features with  $> 80\%$  missing values (Customer Zipcode, Product Description, and Order Zipcode) and eliminated 2 highly correlated features (Pearson  $r > 0.95$ ): Order Item Total and Sales per Customer (both highly correlated with Sales), resulting in 47 final features ( $52 - 3 - 2 = 47$ ) that ensure data quality and computational efficiency for fraud detection.

TABLE I  
PERFORMANCE COMPARISON ON DATACo SUPPLY CHAIN DATASET (10-FOLD CV)

Method	Precision	Recall	F1-Score	AUC-ROC	AUC-PR
Isolation Forest	0.487 $\pm$ 0.021	0.524 $\pm$ 0.019	0.505 $\pm$ 0.018	0.842 $\pm$ 0.012	0.187 $\pm$ 0.015
SVM	0.673 $\pm$ 0.018	0.612 $\pm$ 0.022	0.641 $\pm$ 0.019	0.883 $\pm$ 0.010	0.284 $\pm$ 0.014
Random Forest	0.761 $\pm$ 0.015	0.694 $\pm$ 0.017	0.726 $\pm$ 0.014	0.908 $\pm$ 0.008	0.367 $\pm$ 0.012
XGBoost	0.782 $\pm$ 0.013	0.703 $\pm$ 0.016	0.740 $\pm$ 0.013	0.915 $\pm$ 0.007	0.389 $\pm$ 0.011
Semi-supervised SVM	0.794 $\pm$ 0.012	0.721 $\pm$ 0.015	0.756 $\pm$ 0.012	0.921 $\pm$ 0.006	0.412 $\pm$ 0.010
Ensemble Method	0.807 $\pm$ 0.011	0.738 $\pm$ 0.014	0.771 $\pm$ 0.011	0.926 $\pm$ 0.006	0.428 $\pm$ 0.009
<b>Two-Phase (Ours)</b>	<b>0.856 <math>\pm</math> 0.008*</b>	<b>0.782 <math>\pm</math> 0.010*</b>	<b>0.817 <math>\pm</math> 0.007*</b>	<b>0.943 <math>\pm</math> 0.005*</b>	<b>0.486 <math>\pm</math> 0.009*</b>

\* Indicates statistically significant improvement over all baselines (Wilcoxon signed-rank test,  $p < 0.05$ ).

## V. RESULTS AND DISCUSSION

### A. Overall Performance Comparison

Our two-phase approach significantly outperforms all baselines (Table I), achieving F1-score of 0.817—a 6.0% improvement over the best baseline. Wilcoxon signed-rank tests confirm statistical significance ( $p < 0.05$ ) across 10-fold cross-validation.

TABLE II  
DETAILED PERFORMANCE ANALYSIS OF TWO-PHASE APPROACH

Metric	Value
Precision	0.856 $\pm$ 0.008
Recall	0.782 $\pm$ 0.010
F1-Score	0.817 $\pm$ 0.007
False Positive Rate	0.029 $\pm$ 0.004
AUC-ROC	0.943 $\pm$ 0.005
AUC-PR	0.486 $\pm$ 0.009
Training Time (seconds)	143.5 $\pm$ 12.7
Inference Time (ms/transaction)	2.4 $\pm$ 0.4
Memory Usage (GB)	3.8
Total Transactions Processed	180,519

### B. Detailed Performance Analysis

Table II shows detailed performance metrics, demonstrating excellent precision-recall balance (0.856/0.782) with a low 2.9% false positive rate crucial for practical deployment.

TABLE III  
COMPUTATIONAL EFFICIENCY COMPARISON

Method	Train (s)	Infer. (ms)	Mem. (GB)	Complex.
Full SVM	432.8	3.8	11.9	$O(n^2)$
Semi-sup. SVM	387.5	3.5	10.5	$O(n^2)$
XGBoost	214.3	2.1	8.7	$O(n \log n)$
Ensemble	298.6	4.2	10.3	$O(n \log n)$
<b>Ours</b>	<b>143.5</b>	<b>2.4</b>	<b>3.8</b>	<b><math>O(n \log n)</math></b>

Our approach achieves 67% training time reduction, 68% memory reduction (11.9GB $\rightarrow$ 3.8GB), and  $O(n \log n)$  scalability (Table III).

### C. Ablation Study

Table IV shows the performance improvement through self-training iterations. The iterative process demonstrates consistent improvement from the initial F1-score of 0.695 to 0.817,

TABLE IV  
PERFORMANCE IMPROVEMENT THROUGH SELF-TRAINING ITERATIONS ON DATACo DATASET

Iteration	F1-Score	$\Delta$ F1	Precision	Recall
0 (Initial)	0.695	-	0.742	0.653
1	0.743	0.048	0.789	0.702
2	0.781	0.038	0.823	0.744
3	0.817	0.036	0.856	0.782
4	0.817	0.000	0.856	0.782

with the largest gain (0.048) in the first iteration when high-confidence pseudo-labels are incorporated. The diminishing returns pattern (0.038, 0.036) validates our convergence criteria, achieving stability at iteration 4 without overfitting.

### D. Hyperparameter Sensitivity Analysis

We evaluated the sensitivity of our approach to key hyperparameters. Table V shows F1-score variations:

TABLE V  
HYPERPARAMETER SENSITIVITY ANALYSIS

Parameter	Range Tested	F1-Score Range
$\alpha$ (IF threshold)	[1.0, 2.0]	0.803 - 0.817
$\theta_{base}$ (confidence)	[0.80, 0.90]	0.809 - 0.817
$\beta$ (class balance)	[0.2, 0.4]	0.811 - 0.817

The results demonstrate robustness to parameter variations, with F1-score fluctuations within 1.7% across reasonable parameter ranges.

### E. Performance with Varying Labeled Data

As mentioned in Section 4.3, we evaluated our approach with different percentages of labeled data to assess its robustness under varying supervision levels. Table VI presents the results:

Results show strong performance with only 5% labeled data (F1-score 0.773), with diminishing returns beyond 10% indicating effective unlabeled data utilization.

While SVM with RBF kernels provides good performance, the decision boundaries are not easily interpretable. For supply chain managers requiring explanations, we recommend extracting decision rules from the support vectors or using

TABLE VI  
PERFORMANCE WITH VARYING LABELED DATA PERCENTAGES

Labeled %	Precision	Recall	F1-Score
5%	0.812 $\pm$ 0.011	0.738 $\pm$ 0.013	0.773 $\pm$ 0.010
10%	0.856 $\pm$ 0.008	0.782 $\pm$ 0.010	0.817 $\pm$ 0.007
20%	0.867 $\pm$ 0.007	0.791 $\pm$ 0.009	0.827 $\pm$ 0.006

LIME/SHAP for post-hoc explanations. The Isolation Forest phase provides some interpretability through anomaly scores that indicate deviation from normal patterns.

#### F. Limitations and Failure Modes

While our approach demonstrates strong performance, several limitations warrant discussion:

- Concept Drift: Our current framework assumes static fraud patterns. In practice, fraudsters continuously evolve their techniques. Future work will incorporate online learning capabilities.
- Failure Modes: Our approach may underperform when: (1) fraud patterns deviate significantly from anomalous behavior assumptions, (2) the unlabeled data contains a higher fraud rate than expected, affecting pseudo-labeling quality, or (3) feature distributions shift dramatically between training and deployment.
- Cost-Sensitive Considerations: While we report a 2.9% false positive rate, the business impact varies by context. In high-value transactions, even this rate could be costly. Future work should incorporate domain-specific cost matrices to optimize for business objectives rather than purely statistical metrics.

## VI. CONCLUSIONS

This research introduces a two-phase learning model combining Isolation Forest pre-filtering with self-training SVM refinement for supply chain fraud detection. Our approach addresses class imbalance, limited labeled data, and computational scalability challenges, achieving an F1-score of 0.817 with 2.9% false positive rate on the DataCo dataset. The framework reduces training time by 67% and memory usage by 73% compared to traditional methods, demonstrating that hybrid machine learning techniques can provide robust, practical solutions for complex supply chain environments with scarce labeled data. These computational efficiency gains enable real-world deployment in large-scale systems while maintaining reliable fraud detection.

Future work will develop online learning capabilities for adapting to evolving fraud patterns and explore graph-based methods to capture complex supply chain relationships.

#### DATA AVAILABILITY

The implementation is publicly available at: <https://colab.research.google.com/drive/1eIWYQbhuCgcaQ6p4JZJJ2BZhF6cQOC-z>. The DataCo dataset is accessible from Mendeley Data [7].

## REFERENCES

- [1] N. Modrušan, K. Rabuzin, and L. Mršić, "Review of public procurement fraud detection techniques powered by emerging technologies," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 2, pp. 1–9, 2021.
- [2] F. V. Constante-Nicolalde, P. Guerra-Terán, and J. L. Pérez-Medina, "Fraud prediction in smart supply chains using machine learning techniques," in *Proc. Int. Conf. Appl. Technol.*, 2019, pp. 145–159.
- [3] G. Baryannis, S. Dani, and G. Antoniou, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability," *Future Gener. Comput. Syst.*, vol. 101, pp. 993–1004, 2019.
- [4] Y. Zhou, X. Song, and M. Zhou, "Supply chain fraud prediction based on XGBoost method," in *Proc. IEEE ICBAIE*, 2021, pp. 539–542.
- [5] R. A. Bauder and T. M. Khoshgoftaar, "The effects of varying class distribution on learner behavior for Medicare fraud detection with imbalanced big data," *Health Inf. Sci. Syst.*, vol. 6, pp. 1–14, 2018.
- [6] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010.
- [7] F. Constante, F. Silva, and A. Pereira, "DataCo Smart Supply Chain for Big Data Analysis," *Mendeley Data*, V5, 2019. doi: 10.17632/8gx2fvg2k6.5
- [8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE ICDM*, 2008, pp. 413–422.
- [9] M. Homaei, A. Caro Lindo, J. C. Sancho Núñez, Ó. Mogollón Gutiérrez, and J. Alonso Díaz, "The role of Artificial Intelligence in Digital Twin's Cybersecurity," in *Proc. XVII Reunión Española sobre Criptología y Seguridad de la Información (RECSI)*, 2022, p. 7.
- [10] M. Homaei, Ó. Mogollón-Gutiérrez, J. C. Sancho, M. Ávila, and A. Caro, "A review of digital twins and their application in cybersecurity based on artificial intelligence," *Artif. Intell. Rev.*, vol. 57, no. 201, 2024.
- [11] F. Moradi, M. Tarif, and M. Homaei, "A Systematic Review of Machine Learning in Credit Card Fraud Detection," *Preprint*, MDPI AG, Jul. 2025. [Online].
- [12] M.-R. Amini *et al.*, "Self-training: A survey," *Neurocomputing*, vol. 616, p. 128904, 2025.
- [13] L. Hernandez Aros *et al.*, "Financial fraud detection through the application of machine learning techniques: a literature review," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, pp. 1–22, 2024.
- [14] Y. Liu, S. Tang, A. Hussain, and Y. Chen, "A semi-supervised learning approach to supply chain fraud detection," *Expert Systems with Applications*, vol. 211, 2023.
- [15] F. Moradi, M. Tarif, and M. Homaei, "Robust Fraud Detection with Ensemble Learning: A Case Study on the IEEE-CIS Dataset," *Preprint*, Jul. 2025. [Online].
- [16] W. Wei *et al.*, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, pp. 449–475, 2013.
- [17] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 1–39, 2012.
- [18] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended isolation forest," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1479–1489, 2019.
- [19] D. Wang *et al.*, "A semi-supervised graph attentive network for financial fraud detection," in *Proc. IEEE ICDM*, 2019, pp. 598–607.
- [20] M. Hyun, J. Jeong, and N. Kwak, "Class-imbalanced semi-supervised learning," *arXiv preprint arXiv:2002.06815*, 2020.
- [21] C. Wei *et al.*, "CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proc. IEEE/CVF CVPR*, 2021, pp. 10857–10866.
- [22] D. Wang, F. Zhang, and K. Li, "Detection of fraudulent transactions in supply chains using one-class SVM," *Computers & Industrial Engineering*, vol. 169, 2022.
- [23] A. A. Alhashmi *et al.*, "An ensemble-based fraud detection model for financial transaction cyber threat classification and countermeasures," *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 6, pp. 12433–12439, 2023.