Active Learning for Human-in-the-Loop Customs Inspection

Sundong Kim, Tung-Duong Mai, Sungwon Han, Sungwon Park, Thi Nguyen D.K., Jaechan So, Karandeep Singh, and Meeyoung Cha, *Member, IEEE*

Abstract—We study the human-in-the-loop customs inspection scenario, where an AI-assisted algorithm supports customs officers by recommending a set of imported goods to be inspected. If the inspected items are fraudulent, the officers can levy extra duties. The updated decisions are used as additional training data for successive iterations. Inspecting only the likely fraudulent items may lead to an immediate gain in revenue, yet it does not bring new insights for learning dynamic trade patterns. In contrast, including uncertain items in the inspection helps gradually acquire new knowledge that will be used as supplementary training resources to update the system. Based on multiyear customs declaration logs obtained from three countries, we demonstrate that some degree of exploration is necessary to cope with domain shifts in trade data. The results show that a hybrid strategy of jointly selecting likely fraudulent and uncertain items will eventually outperform the exploitation-only strategy.

Index Terms—Customs selection, fraud detection, active learning, online learning, human-in-the-loop, import control

1 Introduction

Suppose you are given the task of developing an AI-based selection system to assist customs officers working on-site who inspect goods based on recommendations. With the increasing prevalence of online retail, the kinds and amounts of trade traffic are growing astronomically, as are emerging fraudulent trades that try to deceive the system with sophisticated tactics. For example, during the COVID-19 pandemic, the World Customs Organization reported increased numbers of attempted fraud and tax evasion incidents [1]. A detection algorithm that is tailored to past logs will inevitably degrade over time. However, relearning the algorithm entirely may waste domain knowledge that has been accumulated over decades. How should the system balance between exploiting existing knowledge and exploring new trends?

As illustrated in this story, machine learning models in online prediction settings must adapt well to changes in the input, a challenge known as *concept drift* [2]. In the context of customs operations, the list of countries procuring a particular product will change over time, and products that are foreign to systems may be declared (e.g., new technology). Even a well-trained machine learning model can fall into the trap of confirmation bias and may not capture these changes. Particularly in situations where manual labeling is expensive, it can be challenging to make significant changes to the model's working logic.

To mitigate this problem, *active learning* techniques can help users decide how to query and interactively annotate data points in light of unknown concepts [3]. In the recent active learning setting, the model encourages querying uncertain samples while ensuring sample diversity. However, in customs risk management systems, queried data are often subject to evaluation. The system needs to be *profitable* while securing knowledge for the future. Therefore, the learning model cannot fully follow the exploration principles of active learning. We introduce a case study in which customs administrations maintain an AI-based selection model to support officers' collecting duties.

Fig. 1 depicts a customs clearance process. Importers need to specify the trade items' information in import declaration forms to trade goods across borders. We hypothesize that a trade selection model plays a role in prioritizing items for inspection. Officers follow the recommendation to manually inspect the authenticity of the chosen items and levy additional tariffs if there is fraud. In most customs offices, only a small subset of the import goods (say, 1–5%) are inspected due to the large trade volume. Once items are inspected, whether there is fraud or not, the performance of the customs offices is evaluated; at this time, the selection model's parameters can be updated with new knowledge. Many customs offices worldwide set aside a small set of random samples to be inspected to learn new fraud patterns [4]. This paper aims to innovate the sample selection strategy, together with the existing inspection process.

We propose a hybrid selection strategy to maximize long-term revenue from fraud detection while maintaining

- *Sundong Kim and Karandeep Singh are with Data Science Group, Institute for Basic Science, Daejeon 34126, Republic of Korea. E-mail: {sundong, ksingh}@ibs.re.kr.*
- *Tung-Duong Mai, Sungwon Han, Sungwon Park, Thi Nguyen D.K., and Jaechan So are with School of Computing, KAIST, Daejeon 34141, Republic of Korea. E-mail: {john_mai_2605, lion4151, psw0416, ndkthi, jcs0510}@kaist.ac.kr.*
- *Meeyoung Cha is with Data Science Group, Institute for Basic Science, Daejeon 34126, Republic of Korea, and also with KAIST, Daejeon 34141, Republic of Korea. E-mail: mcha@ibs.re.kr.*

Manuscript received 29 May 2021; revised 27 November 2021; accepted 15 January 2022. Date of publication 27 January 2022; date of current version 8 November 2023.
This work was supported in part by the Institute for Basic Science under Grants IBS-R029-C2 and IBS-R029-Y4, and in part by Basic Science Research Program through the National Research Foundation of Korea under Grant NRF-2017R1E1A1A01076400.
(Corresponding authors: Sundong Kim and Meeyoung Cha.)
Recommended for acceptance by J. Li, L. He, H. Peng, P. Cui, C.C. Aggarwal, and P.S. Yu.
Digital Object Identifier no. 10.1109/TKDE.2022.3144299

© 2022 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

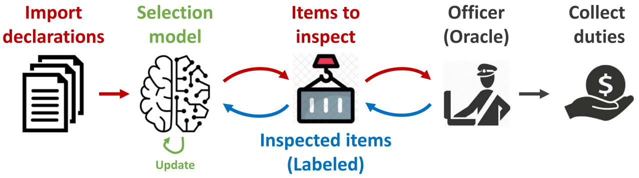


Fig. 1. Illustration of the customs clearance process.

a high income during a short-term inspection. By leveraging the concept of *exploration*, our model remains up to date against concept drifts. In contrast, the concept of *exploitation* maintains high-quality precision for current fraudulent transactions. As an alternative to the random exploration used by customs, we propose the gATE exploration methodology, built upon a *state-of-the-art* active learning approach [5]. gATE is designed to select the most informative samples in a diverse way to capture the dynamically changing traits of trade flows. Tested on actual multiyear actual trade data from *three* African countries, we empirically demonstrate that the proposed hybrid model outperforms state-of-the-art models in detecting fraud and securing revenue. Our key contributions are as follows:

1) We define the problem of concept drifts in the context of customs fraud detection, i.e., a dynamic trade selection setting that adaptively acknowledges new trends in data while securing revenue collected from fraud detection.
2) We propose a novel hybrid sampling strategy based on active learning that combines exploration and exploitation strategies.
3) The experiments demonstrate the long-term benefit of exploration strategies on real trade logs from three African countries.
4) We prepare the codes for simulating customs trade selection, considering the needs of customs administration. See the reproducibility section.

Since 2019, we have collaborated with customs communities represented by the World Customs Organization, and their partner countries, including the Nigeria Customs Service. Namely, our prior work DATE classifies and ranks illegal trade flows that contribute most to the overall customs revenue when they are identified [6]. DATE is open-source[1] and is being studied widely to advance data analytics capabilities in customs organizations. However, in the process of piloting DATE in a live environment, we found that concept drift could be fatal to the model performance. Considering that various fraud detection algorithms were tested in an offline setting [4], [6], [7], [8], it is very likely that the model will suffer from confirmation bias in a live setting. We also show that our DATE model suffers from confirmation bias in Fig. 4c. Avoiding bias in the model is the primary motivation for extending the research.

In contrast to our prior work, we propose a *hybrid* algorithm with a new exploration strategy gATE and refine our experiments from 80–20% data splitting to long-term simulation with consecutive inspections. We also tested our algorithm in datasets from multiple countries.

2 Related Work

2.1 Customs Fraud Detection

Earlier research on customs fraud detection focused on rule-based or random selection algorithms [4], [7]. While they are intuitive and widely used, these classical methods need to relearn patterns periodically, leading to high maintenance costs. The application of machine learning in customs administration has been a closed task primarily due to the proprietary nature of the data. Several recent studies have shown the use of off-the-shelf machine learning techniques such as XGBoost and the support vector machine (SVM) [8], [9]. Recently, dual attentive tree-aware embedding (DATE) was proposed, employing transaction-level embeddings in customs fraud detection [6]. This new model provides interpretable decisions that can be checked by customs officers and yield high revenue through the collected tax.

However, these algorithms are expected to face performance degradation over a long period due to their limited adaptability to uncertainty, diversity, and concept drifts in trade data. We introduce the concept of exploration to remain up to date against concept drifts to address this issue.

2.2 Concept Drift

Concept drift describes unexpected changes in the underlying distribution of streaming data over time [2]. Past research has studied three aspects of concept drift: drift detection, drift understanding, and drift adaptation [10]. Several methods are available for adapting existing learning models to concept drift. The most straightforward way is to retrain a model with the latest data and replace the obsolete model parameters when drift is observed [11]. In cases with recurring drift, ensemble methods are known to be effective. A classic example involves utilizing a tree-based classifier and replacing an obsolete tree with a new tree [12]. Voting schemes have also been applied to manage base classifiers by adding them to ensembles [13]. The requirement of maintaining a set of pre-defined classifiers is a major drawback of these methods.

For stream applications, where only a fraction of the given data is annotated by human effort, one can consider sample selection via active learning to maintain the optimal level of performance. This situation is often subject to the concept drift problem [14]. Since our customs trade selection problem also falls under this setting, we consider active learning as a potential solution.

2.3 Active Learning

Active learning enables an algorithm to elicit ground truth labels for uncertain data instances and enhance its performance [3], [15]. It has been utilized in training models to deal with high-dimensional data [16], to offer long-term benefits [17], [18], to select appropriate data instances to speed up the model training [19], and to train the model with a limited budget [20].

For example, one study proposed a way to measure the 'informativeness' of given samples [21]. Others have proposed collecting as much information as possible by prioritizing inspection of uncertain samples [22], [23], [24]. Another line of research has focused on improving diversity by

[1] http://bit.ly/kdd20-date

TABLE 1
Notation Used Throughout the Paper

Symbol	Definition
$\mathcal{B}$	Import flows.
$\mathbf{x}$	A transaction (item) from import flows $\mathcal{B}$.
y^{cls}	A binary label denoting that item $\mathbf{x}$ is fraud.
y^{rev}	A non-negative value (label) denoting item $\mathbf{x}$'s additional revenue upon its inspection.
$\mathcal{B}_t$	Import flows arriving during an unit interval before time t.
$\mathcal{B}_t^S(f)$	A set of items selected by strategy f at time t. These items are subject to inspection. After inspection, their labels $\boldsymbol{y}^{cls}$ and $\boldsymbol{y}^{rev}$ are obtained. For simplicity, we denote it as $\mathcal{B}_t^S$.
$\mathcal{B}_t^F$	A subset of $\mathcal{B}_t^S$ in which items are considered fraudulent.
$\mathcal{B}_t^U$	A subset of $\mathcal{B}_t^S$ in which items are considered uncertain.
f	A customs selection strategy.
X_t	Training data at time t. X_t is used to update the parameters of strategy f.
r_t	An inspection (selection) rate at time t. ($0\% \leq r_t \leq 100\%$)
m	An evaluation metric for $\mathcal{B}_t^S(f)$. (e.g., Revenue@k%).

$$\mathcal{B}_t^S = \{x \mid x \in \mathcal{B}_t; f\}, \quad (1) \qquad \mathcal{B}_t^S \subset \mathcal{B}_t, \quad (2)$$

$$|\mathcal{B}_t^S| = r_t \cdot |\mathcal{B}_t|, \quad (3) \qquad \mathcal{B}_t^U \subset \mathcal{B}_t^S, \quad (4)$$

$$\mathcal{B}_t^F \subset \mathcal{B}_t^S, \quad (5) \qquad \mathcal{B}_t^S = \mathcal{B}_t^F \cup \mathcal{B}_t^U, \quad (6)$$

$$\mathcal{B} = \bigcup_t \mathcal{B}_t, \quad (7) \qquad X_t = \bigcup_{t'<t} \mathcal{B}_{t'}^S(f). \quad (8)$$

strategically collecting samples to represent the overall data distribution. Diversity-based algorithms include region-based active learning [25] and core-set-based approaches [26]. Recent research has also focused on the concurrent inclusion of uncertainty and diversity aspects [5], [27].

Collectively, these approaches share common limitations in practical use. First, active learning research has assumed an offline setting [5], [28], [29], [30]. For example, HAL showed that including simple exploration helps margin sampling in a skewed dataset [30], and BADGE showed the effectiveness of sampling uncertain data points in a diverse way [5]. However, their evaluations are based on fixed test data, which cannot accommodate concept drifts in real active learning scenarios. Real-world logs exhibit substantial changes and dynamics over time, as we will demonstrate in this paper, making most static machine learning models obsolete.

Second, extant models separate the processes of exploitation (i.e., inspection) and exploration (i.e., annotation). In practice, every manual inspection or annotation is a cost in which the budget is often limited (for example, by inspection officers in customs). Given the shared budget, an exploration-oriented active learning algorithm is unlikely to succeed if it is learned separately. This constrained optimization setting has not been handled in conventional approaches. Our work addresses these two realistic settings.

3 Customs Trade Selection

The customs administration aims to detect fraudulent transactions and maximize the tax revenue from illicit trades — this is the customs fraud detection problem [6], [8]. Given an import trade flow $\mathcal{B}$, the main goal is to predict both the fraud score y^{cls} and the raised revenue y^{rev} obtainable by inspecting each transaction $\mathbf{x}$. Given the limited budget of inspection and annotation, we address the problem of devising an efficient selection strategy to identify suspicious trades and increase revenue as follows:

> Customs Trade Selection Problem.
> *Given trade flows $\mathcal{B}$, construct a sample selection strategy f that maximizes the detection of fraudulent transactions and the associated tax revenue.*

The trade flows $\mathcal{B}$ consist of the online stream of trade records[2], including the importer ID, commodity ID (such as the Harmonized System Codes), and declared price of goods. The characteristic distributions of illicit transactions are assumed to change over time (see Sec. 5.1.1).

3.1 Active Selection for the Online Setting

Extant research on customs fraud detection mainly concentrates on the static setting, in which a model is trained on large training batches and deployed for fraud detection without further updates [6], [8]. We consider a practical online setting where the characteristic distribution of trade flows $\mathcal{B}$ and the traits of illicit trades change over time. This is the case for the *active customs trade selection problem*. The active customs selection problem requires the selection strategy to help the model update and adapt for new fraud types. All inspected items can bring additional information, and strategically choosing the right items to maximize the model performance is handled in this problem.

We formally define the active customs trade selection problem as follows: At each time t, given a batch of items $\mathcal{B}_t$ from trade flows $\mathcal{B}$, based on a strategy f trained with X_t, customs officers select a batch of items $\mathcal{B}_t^S$ to inspect physically. After inspection, the annotated results are used to update the strategy f. We evaluate the model from timestamp t_k onward. The goal is to devise a *strategy* f^* that maximizes the precision and revenue in the long-term:

$$f^* = \underset{f}{\operatorname{argmax}} \sum_{t \geq t_k} m(\mathcal{B}_t^s(f)), \tag{1}$$

where m is the evaluation metric, which is the precision or revenue from fraud detection. Table 1 lists the notation used frequently throughout the paper, and the main training process for fraud detection with active customs selection is described in Algorithm 1.

4 Hybrid Selection Strategy

The quality of the active customs trade selection problem depends on having a good selection strategy f. We propose a new strategy that combines two approaches: *exploitation* and *exploration*. The exploitation approach selects the most likely fraudulent and highly profitable items to secure short-term revenue for customs administration. The exploration approach, in contrast, selects uncertain items at the risk of temporary revenue regret, yet potentially detects novel fraud patterns in the future. Our algorithm mixes these two components to gain long-term benefits and secure immediate revenue from imbalanced customs datasets.

[2] These terms are used interchangeably: transactions, items, goods.

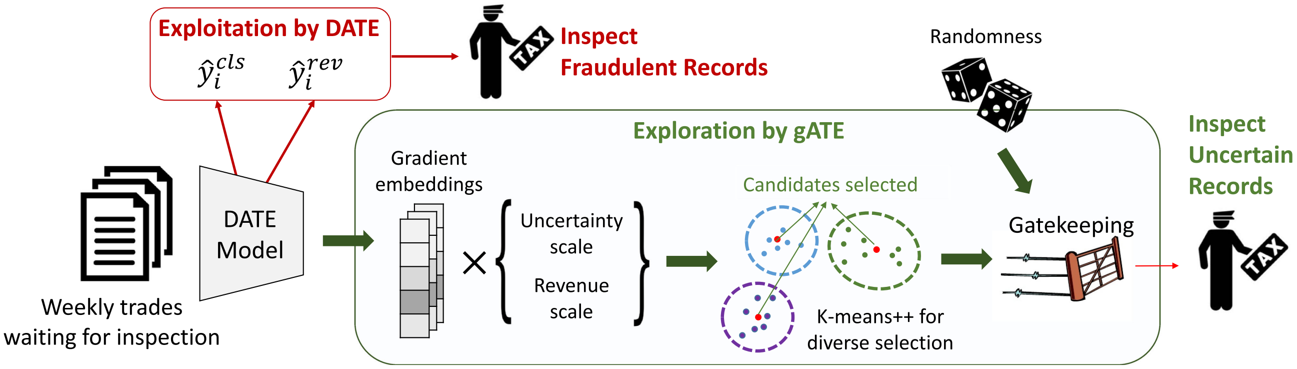

Exploitation by DATE
$\hat{y}_i^{cls}$ $\hat{y}_i^{rev}$
Inspect Fraudulent Records
Randomness
Exploration by gATE
Inspect Uncertain Records
Gradient embeddings
Uncertainty scale
Revenue scale
Candidates selected
Gatekeeping
K-means++ for diverse selection
DATE Model
Weekly trades waiting for inspection
TAX

Fig. 2. Illustration of the hybrid selection framework. The state-of-the-art exploitation model DATE first computes whether the input trade records are fraudulent. Then, we backpropagate it with pseudo-labels to generate a gradient embedding and rescale it with its uncertainty and revenue. Then, the k-means++ algorithm selects diverse yet uncertain samples for inspection. A gating unit decides which items to explore.

Fig. 2 illustrates the overall framework of the proposed model.

4.1 Exploitation Strategy: Customs Fraud Detection Tailored to Maximize the Short-Term Revenue

We employ the current state-of-the-art algorithm in illicit trade detection, DATE [6], as a baseline for this research. It is a tree-enhanced dual-attentive model that optimizes the dual objectives of (1) illicit transaction classification and (2) revenue prediction. We leverage the predicted fraud score of DATE for our exploitation strategy. We update the DATE model at each timestamp and select the most suspicious items as per the inspection budget (see Algorithm 2).

4.2 Exploration Strategy: Customs Fraud Detection Adapt to Concept Drift and Aiming at Long-Term Revenue

The exploitation strategy selects the more familiar and highly suspicious transactions for inspection; therefore, it tends to underperform over time as trade patterns gradually change. In contrast, our hybrid strategy chooses to add a small portion of new and uncertain trades as a learning sample in the training data, which gradually affects the model's future prediction performance. Since fraud types are constantly evolving, the model performance might drop over time. We propose an exploration strategy to select uncertain trade items, with additional consideration of diversity and revenue, to resolve these issues.

4.2.1 Exploration in Light of Uncertainty and Diversity

One approach to detecting new fraud types is to utilize uncertainty in the query strategy. Selecting items for which the model is least confident can provide more information on similar new observations. However, this strategy can create an unfavorable scenario where newer labeled data do not include diverse transaction types and labels for identical transactions continue to accumulate. Considering this, we include the concept of diversity along with uncertainty in our selection strategy; i.e., we choose the most diverse samples possible for stable and fast exploration [5], [28]. We take in gradient embedding and k-means++ initialization from BADGE [5] in our exploitation model DATE to determine which trades should be queried for inspection by considering uncertainty and diversity concepts. The detailed implementation of each concept is described below.

Algorithm 1. Active Customs Trade Selection

```
Input: Previous inspection histories $\mathcal{H}$, initial inspection rate $r_0$, target inspection rate $r$, unlabeled datastream of new items in each timestamp $\mathcal{B}_t$
Output: Items for inspection $\mathcal{B}_t^S$ in each timestamp t
/ * Considering a weekly inspection is made.* /
Initialize the training data $X_1$ from inspection histories $\mathcal{H}$;
for $t = t_1, \cdots,$ do
    Obtain the batch of new items $\mathcal{B}_t$;
    Determine the weekly inspection budget $r_t$, using $r_0$ and $r$;
    / * Selection by the algorithm * /
    Train the selection strategy $f$ with $X_t$;
    Based on $f$, select a set $\mathcal{B}_t^S$ of $r_t|\mathcal{B}_t|$ items for inspection;
    / * Inspection by officers * /
    Obtain the ground-truth annotation $(\mathbf{x}_i, y_i^{cls}, y_i^{rev})$ for each item $\mathbf{x}_i \in \mathcal{B}_t^S$ after manual inspection;
    Evaluate the results by precision and revenue;
    Add the newly annotated items into the training data: $X_{t+1} = X_t \cup \mathcal{B}_t^S$;
end
```

Algorithm 2. Exploiting suspicious items by DATE

```
Input: Training set $X_t$, items received $\mathcal{B}_t$, inspection rate $r_t$
Output: A batch of selected items $\mathcal{B}_t^S$
  / * Corresponds to the selection part in Alg. 1.* /
Train the DATE model using training set $X_t$;
Perform prediction on $\mathcal{B}_t$, obtain the predicted annotation $(\mathbf{x}_i, \hat{y}_i^{cls}, \hat{y}_i^{rev})$ for each item $\mathbf{x}_i \in \mathcal{B}_t$;
Obtain the set $\mathcal{B}_t^S$ of $r_t|\mathcal{B}_t|$ items with the highest fraud score $\hat{y}_i^{cls}$;
```

Uncertainty. If a sample generates a large gradient loss and, consequently, a large parameter update, the item potentially contains useful information. This means that the magnitude of *gradient embedding* reflects the uncertainty of the model on samples. We aim to choose trade flows with uncertainty using the magnitude of gradient embedding with this motivation. At time t, for each trade item $\mathbf{x}_i$ in $\mathcal{B}_t$, the illicitness classifier h_θ from the DATE model returns its fraud score $\hat{y}_i^{cls}$, which indicates the illicit class of y_i^{cls}.

$$h_\theta(\mathbf{x}_i) = \sigma(W \cdot z_\phi(\mathbf{x}_i)), \tag{2}$$

where W is a weight matrix that projects the transaction embedding z_ϕ to the DATE illicitness class space.

The gradient embedding g_{x_i} is the gradient of the loss function with respect to W and sample x_i. Since the received data points are unlabeled (not yet inspected), we predict the pseudo label $\hat{c}_i$ by the fraud score with a threshold of 0.5 (i.e., $\hat{c}_i = \mathbb{1}(\hat{y}_i^{cls} \geq 0.5)$). This pseudo label is used to calculate the loss, resulting in the gradient embedding described as:

$$g_{x_i}^c = (p_i^c - \mathbb{1}(\hat{c}_i = c)) \cdot z_\phi(x_i), \tag{3}$$

where $c \in \{0, 1\}$ corresponds to the two classes and p_i^c is the predicted probability for class c; $p_i^{c=0} = 1 - \hat{y}_i^{cls}$ and $p_i^{c=1} = \hat{y}_i^{cls}$.

Diversity. We inspect items that bring large changes and diverse views to the model for effective exploration. Considering diversity, we can avoid a situation where similar items are inspected redundantly. We suggest a batch construction method by selecting the most representative samples through the k-means++ algorithm [31], which produces a good initial clustering situation. K-means++ obtains a set $\mathcal{B}_t^s$ of k centroids sampled in proportion to the nearest set's centroids. Samples with small gradients are also unlikely to be chosen, as their distances are small. Gradient embedding with k-means++ seeding tends to result in the selection of a batch of large and diverse gradient samples. By doing so, our selection strategy can consider both sample uncertainty and batch diversity.

4.2.2 Scale Uncertainty and Revenue Effect

To induce the algorithm to select more uncertain and high-revenue items, we introduce extra weights to amplify the effect of uncertainty and revenue. These weights, called the uncertainty scale and revenue scale, adjust the probability of chosen samples by resizing their gradient embedding vectors.

Uncertainty Scale. We magnify the impact of uncertain items by quantifying the model's ability to calibrate an item. We give each item an *uncertainty score* (Eq. (4)) such that the score indicates the magnitude of the model's uncertainty about the item. The uncertainty score unc_i is defined as follows:

$$unc_i = -1.8 \times |\hat{y}_i^{cls} - 0.5| + 1. \tag{4}$$

This concave function maximizes the uncertainty score when the system cannot determine whether an item is fraudulent or not (i.e., the uncertainty score is the largest when the predicted fraud score $\hat{y}_i^{cls}$ is 0.5). We adjust the uncertainty score using a multiplier -1.8 and set its range between 0.1 and 1, leading the exploration algorithm to select every item with some chance.[3] Our intention here is to leave some degree of uncertainty even when the base model is overconfident about its prediction results (i.e., $\hat{y}_i^{cls}$ is close to 0 or 1) [32].

Revenue Scale. Active learning in customs operation requires additional consideration, as revenue needs to be collected as the customs duty. Maximizing the customs duty is one of the top priorities of customs authorities. Therefore, we further amplify the gradient embedding by the DATE model's predicted revenue $\hat{y}_i^{rev}$. The distribution of the amount of customs duty is right-skewed, so we take the *log* of the predicted revenue (Eq. (5)). We can define the final scale factor S_i of $\mathbf{x}_i$ as

$$S_i = unc_i \cdot log(\hat{y}_i^{rev} + \epsilon). \tag{5}$$

As a result, the gradient embedding $g_{x_i}^c$ becomes

$$g_{x_i}^c = S_i \cdot (p_i^c - \mathbb{1}(\hat{c} = c)) \cdot z_\phi(\mathbf{x}_i). \tag{6}$$

k is a constant for computational stability. The algorithm covered in Sections 4.2.1 and 4.2.2 is named bATE, inspired by BADGE [5] and DATE [6].

4.2.3 Gatekeeping

In practice, some importers might commit fraud by analyzing and reverse engineering the model's prediction patterns. We can call them adaptive adversaries of the model. In this situation, randomness is known to improve the robustness and competitiveness of the online algorithm [33]. With this motivation, we introduce randomness to our sampling strategy. Using the validation performance of the DATE model, we establish a gatekeeper. If Rev@n% is higher than the predefined value of θ, the bATE exploration algorithm is used. Otherwise, if the DATE models' outputs are highly unreliable, these inputs can be considered an attack, thereby facilitating the random selecting of items for inspection. To address the issues above, we propose the final exploration strategy, gATE, which is formally written as Algorithm 3:

Algorithm 3. Exploring Unknown Items by gATE

```
Input: Training set $X_t$, items received $\mathcal{B}_t$, inspection rate $r_t$
Output: A batch of selected items $\mathcal{B}_t^S$
/* Corresponds to the selection part in Alg. 1. */
Train the DATE model using training set $X_t$;
Obtain Rev@n% from validation set;
if Rev@n% $> \theta$ then
    Perform prediction on $\mathcal{B}_t$, obtain the predicted annotation $(\mathbf{x}_i, \hat{y}_i^{cls}, \hat{y}_i^{rev})$ for each item $\mathbf{x}_i \in \mathcal{B}_t$;
    Calculate the gradient embedding $g_{x_i}$ (Eq. (6));
    Obtain the set $\mathcal{B}_t^S$ of $r_t|\mathcal{B}_t|$ items by $k$-means++ initialization;
else
    Obtain the set $\mathcal{B}_t^S$ of $r_t|\mathcal{B}_t|$ items by random sampling;
end
```

4.3 Hybrid Strategy

The exploitation-only model can lead to confirmation bias. With a model trained only on historical data and considering the concept drift in customs datasets, the model tends to be unreliable because of outliers. However, a pure exploration strategy cannot secure customs revenue and is unrealistic in the customs setting. Hence, we consider a balance between the two to achieve both short-term and long-term performance. We propose a *hybrid selection strategy* under the online

[3] This setting shows the best results on our datasets. For practical use, the best parameters can be found using the validation set.

TABLE 2
Overview of the Item-Level Import Data, in Which the Description and Example of Each Variable are Provided

Type	Variable	Description	Example
Features	*sgd.id*	An individual numeric identifier for Single Goods Declaration (SGD).	SGD347276
	sgd.date	The year, month and day on which the transaction occurred.	13-11-28
	importer.id	An individual identifier by importer based on the tax identifier number (TIN) system.	IMP364856
	declarant.id	An individual identification number issued by Customs to brokers.	DEC795367
	country	Three-digit country ISO code corresponding to transaction.	USA
	office.id	The customs office where the transaction was processed.	OFFICE91
	tariff.code	A 10-digit code indicating the applicable tariff of the item based on the harmonised system (HS).	8703232926
	quantity	The specified number of items.	1
	gross.weight	The physical weight of the goods.	150kg
	fob.value	The value of the transaction excluding, insurance and freight costs.	$350
	cif.value	The value of the transaction including the insurance and freight costs.	$400
	total.taxes	Tariffs calculated by initial declaration.	$50
Target	***illicit***	Binary target variable that indicates whether the object has fraud.	1
	revenue	Amount of tariff raised after the inspection, only available on some illicit cases.	$20

active learning setting that includes two main approaches, *exploitation* and *exploration*, by utilizing DATE and gATE.

To select items that will potentially enhance the model's performance, we design a gATE strategy for exploration (Sections 4.2.1, 4.2.2, and 4.2.3). We use the DATE strategy that exploits historical knowledge to generate the highest possible revenue [6] and guarantees short-term revenue. The final selection is made by the hybrid approach, which balances the gATE and DATE strategies. The hybrid selection logic can be formally represented by Algorithm 4.

Algorithm 4. Hybrid Selection Using DATE and gATE

```
Input: Training set $X_t$, items received $\mathcal{B}_t$, inspection rate: $r_t$, predefined ratio between two strategies $p_1, p_2$ $(p_1 + p_2 = 1)$
Output: A batch of selected items $\mathcal{B}_t^S = \mathcal{B}_t^F \cup \mathcal{B}_t^U$
/ * Corresponds to the selection part in Alg. 1.* /
Train the DATE model using training set $X_t$;
Obtain the set $\mathcal{B}_t^F$ of $p_1 r_t |\mathcal{B}_t|$ items by the DATE strategy;
$\mathcal{B}_t = \mathcal{B}_t \setminus \mathcal{B}_t^F$;
Obtain the set $\mathcal{B}_t^U$ of $p_2 r_t |\mathcal{B}_t|$ items by the gATE strategy;
```

5 Experiments

5.1 Evaluation Settings

5.1.1 Datasets

We employed item-level import declarations for three countries in Africa. The import data fields included numeric variables such as the item price, weight, and quantity and categorical variables such as the commodity code (HS code), importer ID, country code, and received office. After matching the data format for each country, we preprocessed the variables by following the approach used in a previous study [6]. For categorical variables, we quantified the risk indicators of the importers, declarants, HS code, and countries of origin from their non-compliance records. For example, importers were ranked by their past fraud rates. The importers, whose ranks were above the 90th percentile, were regarded as high-risk importers, and their risk indicators were given values of 1; otherwise, the values are set to 0. This is called risk profiling, and it is more efficient than one-hot encoding those variables. We also add frequently-used cross features such as *unit.value* (=cif.value/quantity), *value/kg* (=cif. value/gross.weight), *tax.ratio* (=total.taxes/cif.value), *unit.tax* (=total.taxes/quantity), and *face.ratio* (=fob.value/cif.value). Table 2 illustrates the import declaration data.

TABLE 3
Statistics of the Datasets

Datasets	Country M	Country N	Country T
Periods	Jan 13–Dec 16	Jan 13–Dec 17	Jan 15–Dec 19
# imports	0.42M	1.93M	4.17M
# importers	41K	165K	133K
# tariff codes	1.9K	6.0K	13.4K
GDP per capita	$412	$2,230	$3,317
Avg. illicit rate	1.64%	4.12%	8.16%

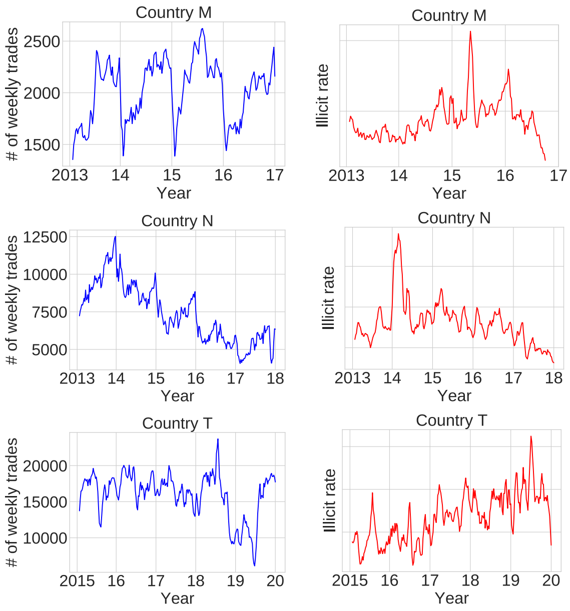
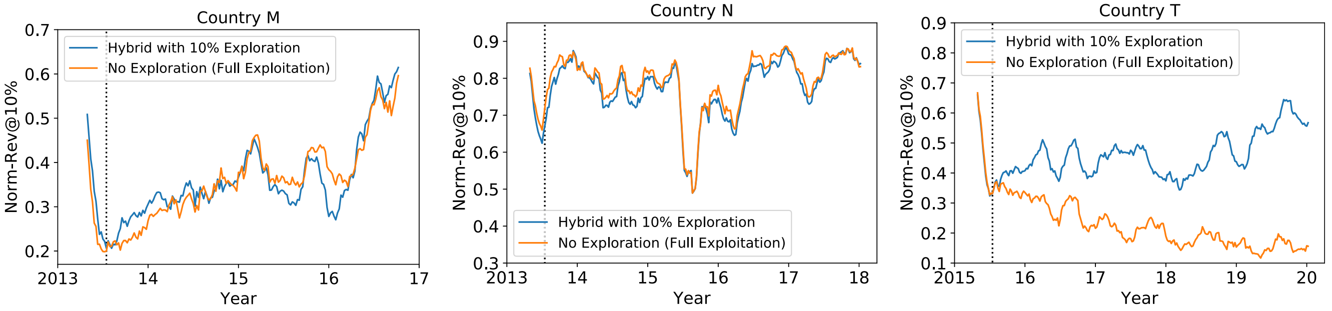


Fig. 3. Number of weekly trades and illicit rate trend over time. The actual values of the illicit trends are hidden due to nondisclosure agreements.



(a) In country M, the performances of both strategies increase over time.

(b) In country N, the performances of both strategies are high and stable.

(c) In country T, the exploitation strategy failed, unlike a hybrid model.

Fig. 4. For some cases, the performance of the exploitation strategy DATE drops over time, but the performance of hybrid strategies remains stable even for cases in which the exploitation strategy fails. This shows that exploration is necessary for maintaining a selection system in the long run.

The three customs were subjected to detailed inspection (i.e., achieving a nearly 100% inspection rate). However, this practice is not sustainable, and the customs offices of these countries plan to reduce the inspection rate in the future. Due to the manual inspection policy, the item labels and tariffs charged are accurately labeled in these logs at the single-goods level. Table 3 and Fig. 3 depict the statistics of the data we utilized.

5.1.2 Long-Term Simulation Setting

The experiment aims to find the best selection strategy to maintain the customs trade selection model in the *long run*. Therefore, we simulated an environment in which a selection model is deployed and maintained for multiple years[4]. Given that one month of training data is available, the system receives import declarations and selects a batch of items to inspect during the week. A selection model is trained based on a predefined strategy, and the most recent four weeks of data are used to validate the model. By using the inspection results, the model is updated every week.

To simulate a scenario of data providers who are willing to reduce the inspection rate gradually, we implemented several methods to decay the inspection rate over time. In this experiment, we set the target inspection rate to 10%. Starting with 100% inspection, we used a linear decaying policy by reducing the inspection rate by 10% each week. Once the target inspection rate is reached, the system maintains this inspection rate for the remaining period. In Figs. 4, 6, 7, and 13, we use a vertical dashed line to indicate when the decay ends, and the target rate is maintained.[5]

5.1.3 Evaluation Metrics

We evaluate the selection strategy performance by referring to two metrics used in previous work [6]: Precision@n% and Revenue@n%. Since the underlying data distribution changes each week in an online setting, these value metrics also fluctuate significantly. In other words, unless the illicit rates or item prices are fixed, these two value metrics will be difficult to interpret directly. We used normalized performances by dividing each value metric by the maximum achievable value, namely, the oracle value.

- Norm – Precision@n%: In a situation where n% of all declared goods are inspected, Pre@n% indicates how many actual instances of fraud exist among the inspected items. The Norm – Pre@n% value of the corresponding algorithm is defined as the value obtained by dividing the Pre@n% of the algorithm by the Pre@n% of the oracle.
- Norm – Revenue@n%: Rev@n% indicates how much revenue (extra tax) can be secured by examining the set of items. The oracle will select the items with the highest revenue. The Norm – Rev@n% value of the corresponding algorithm is defined as the value obtained by dividing the Rev@n% of the algorithm by the Rev@n% of the oracle.

Example. For example, if a system with a 10% inspection rate is operating in an environment with a 2% illicit rate, the Pre@10% and Rev@10% of the oracle would be 0.2 and 1, respectively. Let us consider that the deployed selection strategy achieves a Pre@10% value of 0.18. To prevent any potential interpretation bias caused by the illicit rate that varies from country to country, we divide 0.18 by the performance upper bound of 0.2, which results in 0.9 for Norm – Pre@10%.

Note. We employ a fully labeled dataset and these metrics as the ground truth information. For countries already maintaining a low inspection rate, these metrics can be modified by conditioning on their observable goods. Securing tax revenue was

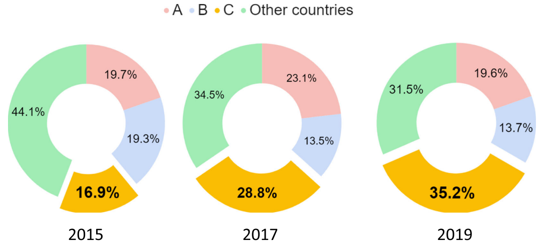
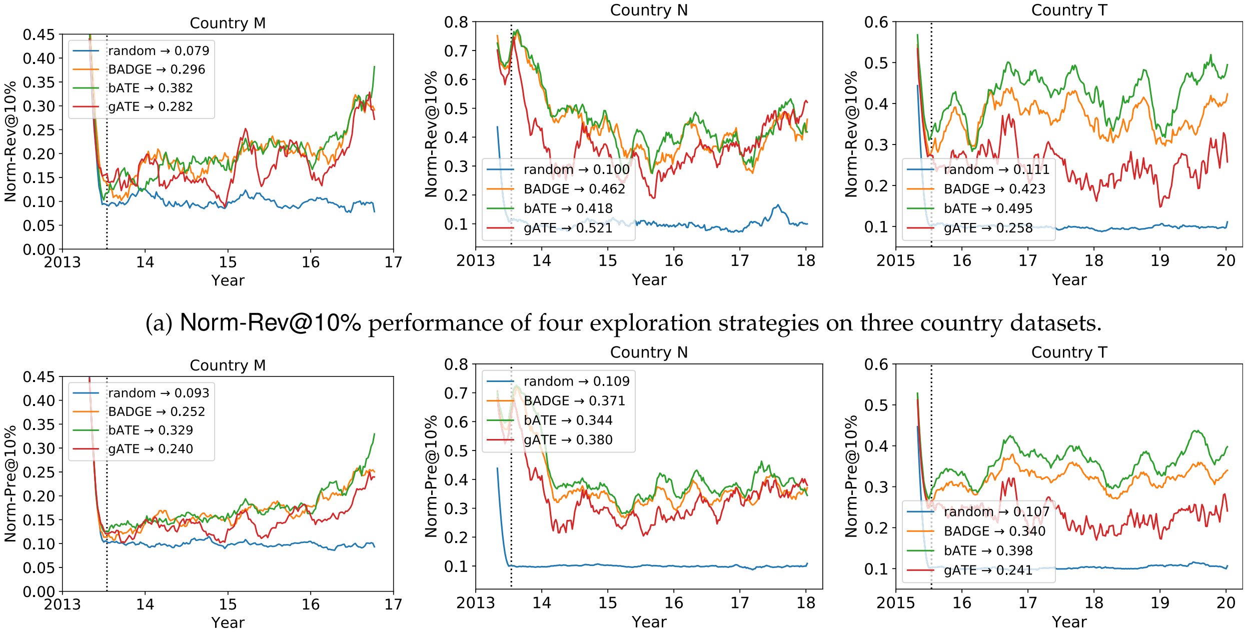


Fig. 5. An example of concept drift in country T: The source country for commodity X (HS-code starting with 620) rapidly changes over the years.

[4] Previous works split the data into training and testing sets on a temporal basis and compared the performance of diverse machine learning models [6], [8]. However, the algorithm's performance in a static prediction state cannot indicate the model's performance in a real setting when the model is deployed.

[5] In countries where the daily import declaration is larger, it would be possible to update the selection strategy every day, and more reliable results could be obtained even with a shorter period.



(a) Norm-Rev@10% performance of four exploration strategies on three country datasets.

(b) Norm-Pre@10% performance of four exploration strategies on three country datasets.

Fig. 6. The performance of the advanced exploration strategy outperforms random selection when the customs trade selection system is operated by the exploration strategy. Note that random exploration is widely used in many customs offices. In addition, the performance of bATE outperforms BADGE, suggesting that the introduced scaling components are practical on active customs trade selection settings.

the most critical screening factor for the developing countries we interacted with since their fiscal income depends on customs services [34]. Therefore, we mainly used $\text{Norm}-\text{Rev@n}\%$ in reporting the results in the following sections.

5.2 When the Exploitation Strategy Fails

To explore this possibility, we first compare the performance of the pure exploitation strategy DATE and a partial-exploitation strategy with some *random* exploration, called a naive hybrid strategy. The naive hybrid strategy uses DATE and random exploration at a ratio of *nine* to *one*. The data reveal that a pure exploitation strategy can lead to a substantial degree of malfunctioning.

Fig. 4c shows that for country T, the performance of the *state-of-the-art* DATE exploitation strategy drops unexpectedly from time to time, yet the hybrid strategy remains stable. The degradation continues despite the increasing training data size, confirming that items chosen for inspection are uninformative and indicating a concept drift in the country's trade pattern. Hence, we conclude that the exploration strategy items significantly boost the performance of the exploitation strategy. Considering that the randomly selected items may affect only 1% of the total revenue on average, the performance boost arises from inspecting unknown items.

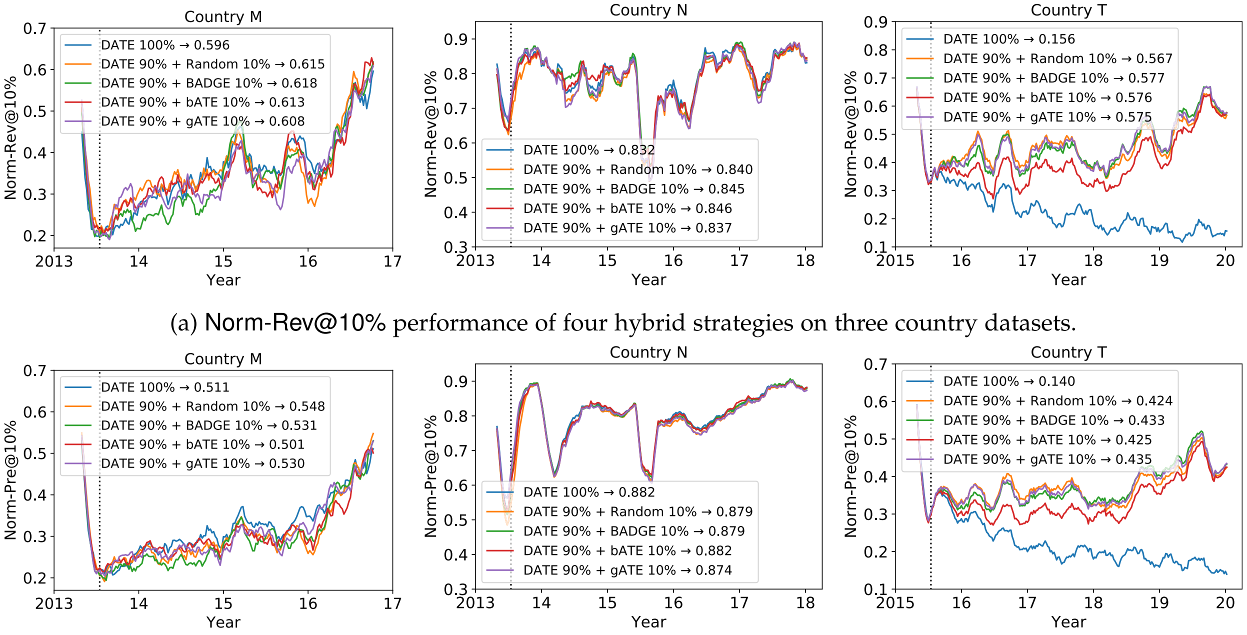
The longitudinal data also allow us to examine how frequently concept drifts occurred in the trading pattern of country T. Fig. 5 shows the ratio of each import country for an item with a commodity code starting with 620 in 2015, 2017, and 2019, indicating a significant level of concept drifts in trade rates for the top imported items. Countries A and B used to be where the item was imported the most, but starting in 2017, the shift in import countries sharply changed, and the country C became the dominant source country for imported goods.

5.3 When the Exploitation Strategy Does Not Fail

Is it common for the performance of the exploitation strategy to decrease over time? We check again to see if these behaviors are common across all countries. Reassuringly, we also observe that the full-exploitation strategy does not always fail. In Figs. 4(a) and 4b, we can see the results obtained from country M and country N. For these countries, maintaining the strategy of screening the most fraudulent items is still valid. However, when we compare the average performance of the exploitation strategy and the hybrid strategy, we can also find that the former does not outperform the latter ($\text{Norm}-\text{Rev@}10\%$, Exploitation versus Hybrid: 59.6% versus 61.5% for country M, 83.2% versus 84.0% for country N; moving average over the previous 13 weeks). It is interesting to see that inspecting a set of random items is even better than inspecting reasonably fraudulent items with high $\hat{y}^{cls}$ values (top 9-10%) for maintaining a customs trade selection system in the long run. Therefore, how much will the performance improve if the better exploration strategy is used rather than the random strategy? We measured the performance of the proposed exploration strategies.

5.4 Finding the Best Exploration Strategy

A natural question arises regarding which strategies would be best for exploration. We first compared the performance of pure exploration strategies, assuming a need to build a system that tends to explore. This experimental setting is necessary for customs administration where there are not enough import histories available, so customs want to construct the working selection system as quickly as possible. This experimental setting is also widely used in the active learning community [5], [28] to compare performances between pool-based active learning algorithms. We performed experiments with four exploration strategies, including our proposed model designed in Section 4.2.



(a) Norm-Rev@10% performance of four hybrid strategies on three country datasets.

(b) Norm-Pre@10% performance of four hybrid strategies on three country datasets.

Fig. 7. Hybrid strategies outperforms the *state-of-the-art* fully exploitation strategy DATE. We confirm that a robust hybrid model can be made even with a simple exploration strategy.

TABLE 4
Summary of the Overall Simulation Results

Model	Country M		Country N		Country T		Synthetic†	
	Without DATE (Fully exploration)	With DATE (Hybrid)	Without DATE (Fully exploration)	With DATE (Hybrid)	Without DATE (Fully exploration)	With DATE (Hybrid)	Without DATE (Fully exploration)	With DATE (Hybrid)
gATE	0.282 / 0.240	0.608 / 0.530	0.521 / 0.380	0.837 / 0.874	0.258 / 0.241	0.575 / 0.435	0.103 / 0.103	0.292 / 0.264
bATE	0.382 / 0.329	0.613 / 0.501	0.418 / 0.344	0.846 / 0.882	0.495 / 0.398	0.576 / 0.425	0.169 / 0.163	0.292 / 0.273
BADGE	0.296 / 0.252	0.618 / 0.531	0.462 / 0.371	0.845 / 0.879	0.423 / 0.340	0.577 / 0.433	0.166 / 0.161	0.287 / 0.268
Random	0.079 / 0.093	0.615 / 0.548	0.100 / 0.109	0.840 / 0.879	0.111 / 0.107	0.567 / 0.424	0.095 / 0.096	0.291 / 0.256
DATE only	-	0.596 / 0.511	-	0.832 / 0.882	-	0.156 / 0.146	-	0.303 / 0.294

The first number denotes Norm – Rev@10%*, and the second number denotes* Norm – Pre@10% *of the model. († Along with three countries dataset, we tested our approach on synthetic data. See the reproducibility section for detail.)*

- Random [4]: Known to be used as an exploration strategy to detect novel fraud in the production systems of some countries.
- BADGE [5]: State-of-the-art active learning approach that selects items considering uncertainty and diversity.
- bATE: Explores by considering predicted revenue as well as item uncertainty and item diversity.
- gATE: Strategically determines whether the exploration strategy is random or bATE, depending on the performance of the base model.

Fig. 6 shows the 13-week moving average performance of the exploration strategies. The results show that the three advanced strategies, BADGE, bATE, and gATE, outperform the random strategy by a large margin. bATE is the top-performing strategy in countries M and T, and gATE performs the best in country N. In the point of view of active learners, these results suggest that the introduced scaling components in our method play a role in constructing the working customs trade selection system more rapidly.

However, it turns out that the performance of the full-exploration strategy is not comparable to the full-exploitation strategy. In our experiments, the performance of the full-exploitation strategy reaches 0.844 in country N (Table 4), while the performance of the full-exploration strategy is 0.52 in the same country, which is not impressive in itself. Although a set of explored samples consists of items with uncommon HS codes or under-invoiced item near the decision boundary, they are not always frauds. Including these items is helpful to the model training process to some extent. However, a model solely trained on these items (i.e., full exploration) is susceptible to noise and hence does not exhibit the best performance. It can be seen that the exploration strategy and exploitation strategy need to be used *together* to guarantee the reliable performance of the customs trade selection system.

5.5 Best Exploration Strategy for the Hybrid Model

Next, we compare the performance of these exploration strategies by applying them with an exploitation strategy. Following Sections 5.2 and 5.3, each hybrid strategy selects 90% of the items by DATE, and the four exploration strategies select the remaining 10% of the items. We also compare the strategies with DATE to show the long-term sustainability of the hybrid strategies.

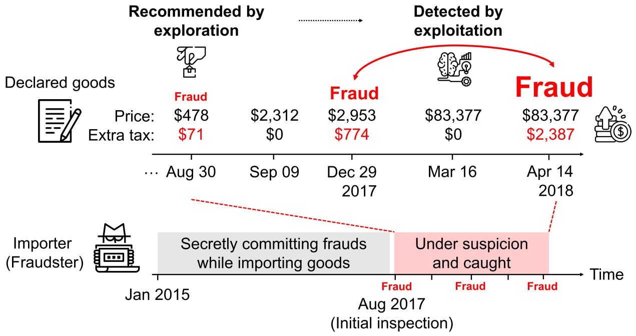
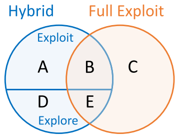

Recommended by exploration
Detected by exploitation
Declared goods
Fraud
Fraud
Fraud
Price: $478 $2,312 $2,953 $83,377 $83,377
Extra tax: $71 $0 $774 $0 $2,387
… Aug 30
Sep 09
Dec 29 2017
Mar 16
Apr 14 2018
Importer (Fraudster)
Secretly committing frauds while importing goods
Under suspicion and caught
Time
Jan 2015
Fraud
Fraud
Fraud
Aug 2017 (Initial inspection)

Fig. 8. Trade log of a fraudster: Thanks to the exploration strategy which noticed the importer's fraud action, the hybrid targeting system is later able to detect major frauds he committed. The system that only operates the exploitation strategy could not detect the frauds until the end.



(a) Breakdown

Inspected items	A	B	C	D	E
# imports	216,935	152,914	255,221	38,286	2,814
# importers	7,547	1,216	785	8,509	235
Fraud rate	36.4%	21.7%	5.9%	5.8%	5.5%
Avg. revenue	$1,882	$782	$138	$180	$129

(b) Statistics of each component

Fig. 9. Statistics of inspected items chosen by each algorithm in country T. Group D represents the items explored and inspected by the hybrid model, but not selected by the full exploitation model. Note the gap between groups A and C.

Fig. 7 and Table 4 summarize the performance of the hybrid models with different exploration strategies. First, we can see that all hybrid strategies outperform a full-exploitation strategy DATE by some margin. For country T, where a staggering decline in the exploitation strategy performance is recorded, our hybrid strategy performs exceptionally stably, and the model ultimately improves. Even though the DATE model for exploitation remains effective for the other two countries, the 10% trade-off for exploration does not hurt the overall performance; rather, this method slightly outperforms the exploitation algorithm. This proves our initial claim that even if we inspect suspicious items, we can guarantee similar performance by learning new patterns from the unknown items.

Second, advanced exploration strategies can help to improve the performance of the whole hybrid model as much as possible. This is shown by the result that DATE+bATE achieves 1.6% higher revenue than DATE+random in country T. It is noteworthy that the best exploration algorithm contributes the most to the hybrid strategy when it is used in the country with the largest trade volume and the highest illicit rate.

Third, the hybrid model's performance with a random exploration strategy is still comparable to the hybrid model in general. In practice, we encourage customs administration to start with a simple exploration strategy without using additional computing power. In contrast to relying on single exploitation or exploration model, the customs trade selection model will be improved even more robustly with both strategies.

5.6 Case Study

Timely exploration allows customs to inspect goods from unknown importers and extend their knowledge. Based on this input, the updated model can prevent potential frauds. Fig. 8 shows a successful case of detecting sequential frauds by our hybrid strategy. This example introduces a trade log of an importer who has imported goods since 2015. After 2.5 years, one of the transactions is subjected to physical inspection by exploration and was labeled a fraud.[6] The importer mixed normal and fraudulent transactions to avoid further inspection. Yet, the newly updated exploitation strategy was able to catch his sequential frauds. Without being triggered by exploration (i.e., fully-exploitation), the targeting system would not have detected frauds from unidentified importers. In our experiment, 1,652 importers and their 170,683 items followed the same pattern (i.e., sequential and sporadic frauds) and were subjected to inspection by hybrid strategy. Among them, only 74 importers and their trades are inspected by the full-exploitation model.

5.7 Detailed Analysis

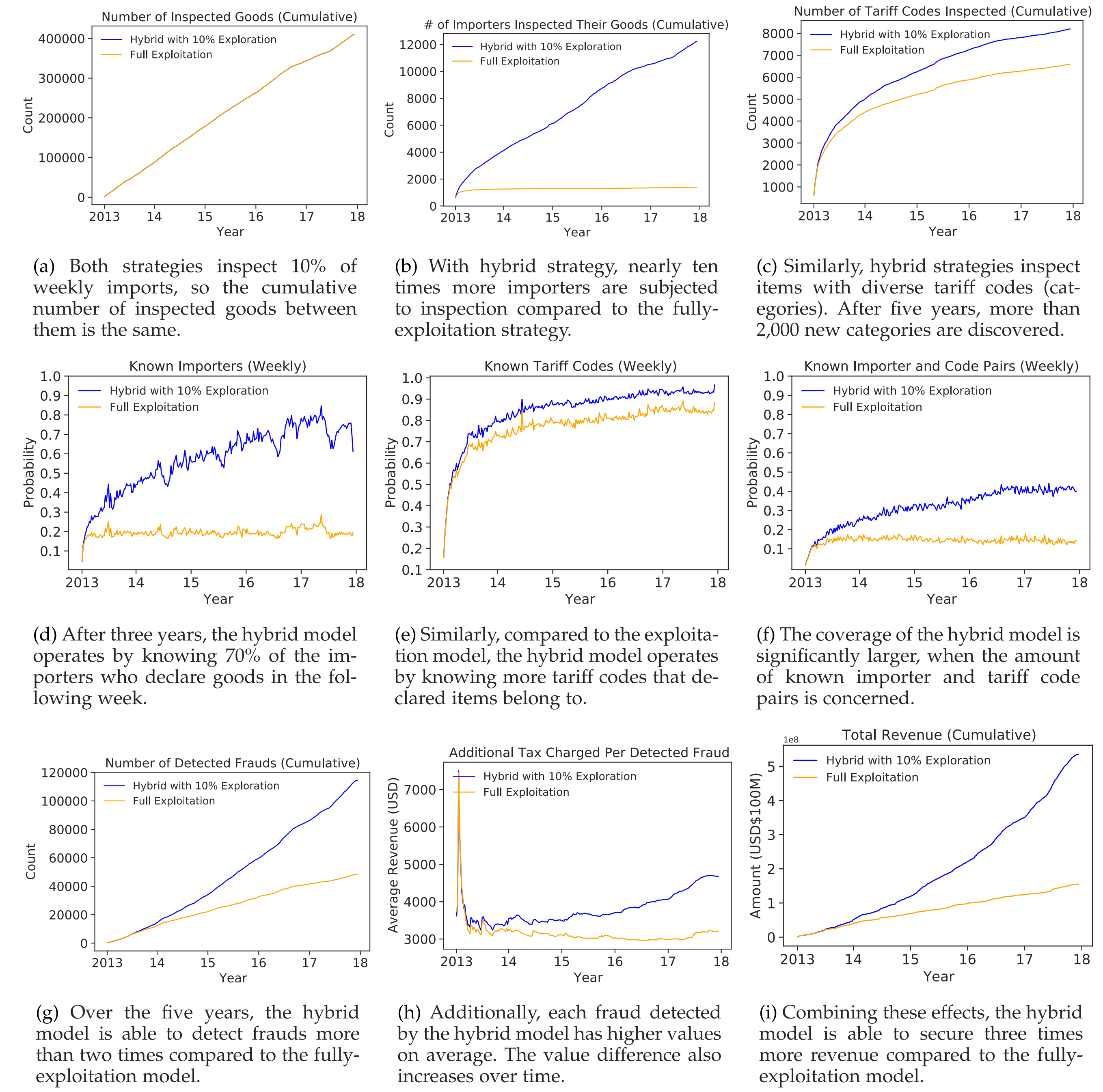
This case study demonstrates that a timely exploration triggers targeting systems to cope better with frauds from new importers. We further compared the statistics of the selected items between two targeting systems (i.e., hybrid and full exploitation). Fig. 9 illustrates how we break down the results into five components and shows their statistics. In the left figure, $A \cup B$ is a set of items selected by hybrid exploitation, and $D \cup E$ is a set of items selected by hybrid exploration. Likewise, $B \cup C \cup E$ is a set of items selected by the full-exploitation model.

The hybrid targeting system makes better trade selection based on the inputs it receives through exploration. Since the exploration strategy inspected 41,100 items from 8,744 importers ($D \cup E$), the exploitation module selected 369,849 suspicious items from 7,944 importers ($A \cup B$). In contrast, the full-exploitation model operated from a limited importer pool. Total 410,949 items were selected from merely 1,392 importers ($B \cup C \cup E$). In addition, the detection rate of the hybrid model and the corresponding revenue per item are higher than those of the full-exploitation model. For comprehensive understanding, we also compared the performance of the hybrid model and the full-exploitation model on various criteria over time. Detailed results are summarized in Fig. 10 with explanations.

6 Concluding Remarks

This paper investigates the human-in-the-loop online active learning problem, where the indicators of the annotated samples are the key criteria for evaluation. One such example can be found in customs inspection, where customs officers need to decide which new cargo to examine (i.e., an exploration strategy) while retaining the history of existing illicit trades (i.e., an exploitation strategy). We present a selection strategy that efficiently combines exploration and exploitation strategies. Our numerical evaluation, based on multiyear transaction logs, provides insights for practical guidelines for setting model parameters in the context of customs screening systems.

[6] From Jan 2015 to Aug 2017, the importer processed 59 transactions, including eight frauds, but none of these transactions was inspected yet.



(a) Both strategies inspect 10% of weekly imports, so the cumulative number of inspected goods between them is the same.

(b) With hybrid strategy, nearly ten times more importers are subjected to inspection compared to the fully-exploitation strategy.

(c) Similarly, hybrid strategies inspect items with diverse tariff codes (categories). After five years, more than 2,000 new categories are discovered.

(d) After three years, the hybrid model operates by knowing 70% of the importers who declare goods in the following week.

(e) Similarly, compared to the exploitation model, the hybrid model operates by knowing more tariff codes that declared items belong to.

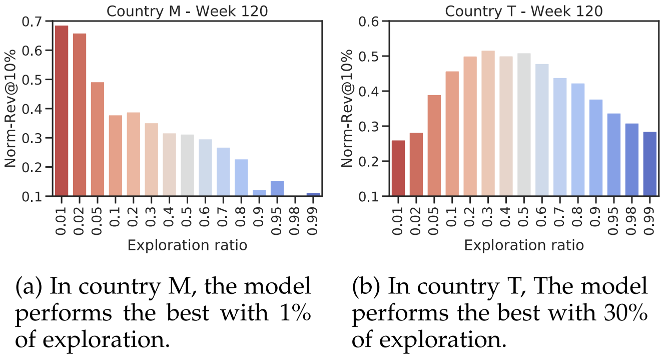
(f) The coverage of the hybrid model is significantly larger, when the amount of known importer and tariff code pairs is concerned.

(g) Over the five years, the hybrid model is able to detect frauds more than two times compared to the fully-exploitation model.

(h) Additionally, each fraud detected by the hybrid model has higher values on average. The value difference also increases over time.

(i) Combining these effects, the hybrid model is able to secure three times more revenue compared to the fully-exploitation model.

Fig. 10. Detailed comparison between the hybrid and fully-exploitation-based targeting systems in country T.



(a) In country M, the model performs the best with 1% of exploration.

(b) In country T, The model performs the best with 30% of exploration.

Fig. 11. Best performing exploration ratio differs by data. In the case which the exploitation strategy does not work well, increasing an exploration ratio helps (Country T).

To facilitate the proposed approach in customs administrations, we make the code and implementation details open to the public. It currently supports diverse exploitation and exploration strategies with various tunable parameters ranging from models to simulation settings so that users can confirm whether our proposed work is well suited for their data. With minor adjustments, our code can also support various decision-making problems with constrained resources. Refer to the supplementary material for the code and data availability. Our forthcoming work will include the following two areas:

- *Determining right balance for hybrid strategies [35]*: In this paper, the ratio between exploration and exploitation is set empirically. The model performance is

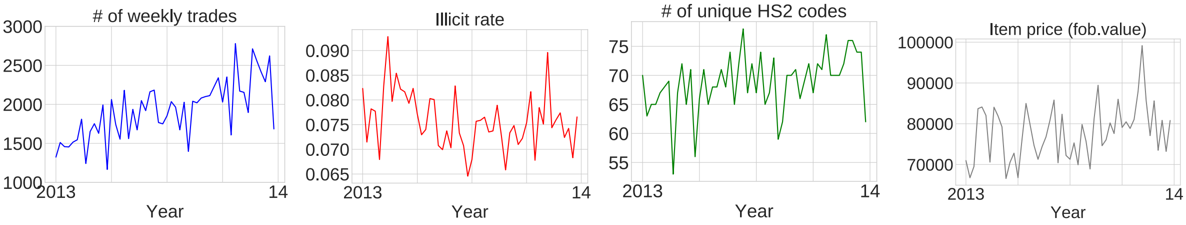


Fig. 12. Weekly statistics of the synthetic data.

sensitive to this ratio, and the performance numbers vary depending on the dataset (Fig. 11). An adaptive algorithm for selecting this ratio will manage this trade-off better. The RP1 algorithm [36] leverages an online learning mechanism with an exponential weight framework [37] to dynamically tune this ratio, which could be applicable in our model.

- *Using richer information*: Higher performance can be achieved by using richer information from a set of uninspected imports, or by using reliable trade data from other countries [41]. Building a set of augmented customs data and learning from it would be a key challenge for learning from richer information, with a semi-supervised learning or transfer learning model.

Reproducibility

Code and Data Availability

In line with this study, we prepared a GitHub repository for simulating customs selection considering the needs of customs administration. Our code is released at https://github.com/Seondong/Customs-Fraud-Detection.

The import transaction data used in the paper cannot be made public due to nondisclosure agreements. Nevertheless, the source code runs compatibly with the synthetic data we included in the repository. In the next section, we will share a step-by-step guide for running our code with synthetic data.

How to Run the Code

Instructions for running the code and reproducing our experiments are as follows:

1) Setup the Python environment: e.g., Anaconda Python 3.7[7]

```
$ source activate py37
```

2) Install the requirements:

```
(py37) $ pip install -r requirements.txt
```

3) Run the simulation: Run `main.py` by selecting the query strategies defined in `./query_strategies/`. The command below runs on synthetic data with a hybrid strategy consisting of 90% DATE and 10% bATE. By running the command, the orange line in Fig. 13c can be reproduced.

```
(py37) $ export CUDA_VISIBLE_DEVICES = 3 &&
↪python main.py –data synthetic
↪ –semi_supervised 0 –batch_size 512
↪ –sampling hybrid –subsamplings
↪ DATE/bATE –weights 0.9/0.1 –mode
↪ –scratch –train_from 20130101
↪ –test_from 20130201 –test_length 7
↪ –valid_length 28
↪ –initial_inspection_rate 100
↪ –final_inspection_rate 10 –epoch 10
↪ –closs bce –rloss full –save 0
↪ –numweeks 100 –inspection_plan
↪ fast_linear_decay
```

4) Check the results: The simulation summaries are saved in .csv format in `./results/performances/`. The figures in this paper can be drawn by running Jupyter Notebooks in the `./analysis/` directory.

5) For further usage: the `.sh` files in the `./bash` directory will give you some ideas for running repeated experiments. See `main.py` for hyperparameter descriptions. Customs officers can simulate our strategies using their data by plugging them into the `./data` directory and adding an argument in `main.py`. The framework can support new selection strategies; The simple XGBoost selection method is found in `./query_strategies/xgb.py`).

Testing on the Synthetic Dataset

Dataset

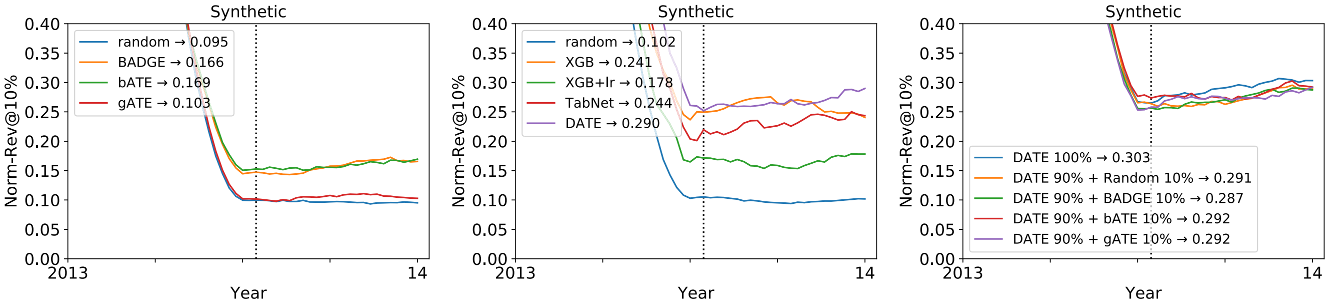
For reproducibility, we provide the experimental results using synthetic import declarations. The dataset is generated by CTGAN [38] and shares similar data fields with real datasets. It consists of 100,000 artificial imports collected from Jan 2013 to Dec 2013. The number of unique importers is 8,653, and the average illicit rate is 7.6%. Fig. 12 depicts the weekly statistics of the dataset.

Results

Fig. 13 shows the experimental results on synthetic data. We confirm that the synthetic data we introduce can help simulate customs selection. According to Fig. 13, the advanced model showed higher performances, supporting our statement that 'Synthetic data also has its fraudulent patterns'. Looking into the details further, we can re-establish our findings from Sections 5.4 and 5.5 with the synthetic data. We run all the experiments five times and report their averages.

Figs. 13(a) and 13b compares the performance between exploration strategies and exploitation strategies. Among the exploration strategies, the state-of-the-art active learning approaches—BADGE and bATE—outperform random learning by a large margin. Additionally, gATE performs nearly randomly because its default hyperparameter $\theta = 0.3$ (Algorithm 3) is set too high for synthetic data.

[7] http://bit.ly/conda-managing-environments



(a) Pure exploration. (b) Pure exploitation. (c) Hybrid strategies.

Fig. 13. Experimental results on the synthetic dataset.

However, pure exploration is not comparable to exploitation due to the nature of our problem. Note the large performance gap between the performance of bATE in Fig. 13a and the performance of simple XGBoost [9] in Fig. 13b. Customs administration should secure short-term revenue by inspecting the most likely fraudulent and highly profitable items and inspecting uncertain items that bring new insights for changing traffic. DATE is well designed for that purpose, showing its effectiveness compared to the *state-of-the-art* classification models for tabular data (e.g., XGB, XGB with logistic regression [39], and TabNet [40]).

Since the data length is relatively short, it is difficult to say that mixing exploration boosts the customs selection performance—Fig. 13c—and yet simple exploration is effective enough to be used as a component of the hybrid strategy. Moreover, the benefit of inspecting 1% of the uncertain items is meaningful enough to compensate for the loss of not inspecting 1% of the fraudulent items.

Acknowledgments

The authors would like to thank the World Customs Organization (WCO) and their partner countries to support their datasets. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of WCO.

References

[1] World Customs Organization, "COVID-19 urgent notice: Counterfeit medical supplies and introduction of export controls on personal protective equipment," 2020. [Online]. Available: http://bit.ly/wco-covid19-frauds

[2] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2346–2363, Dec. 2019.

[3] B. Settles, "Active learning literature survey," Univ. Wisconsin–Madison, Madison, WI, USA, Comput. Sci., Tech. Rep. 1648, 2009.

[4] C. Han and R. Ireland, "Performance measurement of the KCS customs selectivity system," *Risk Manage.*, vol. 16, no. 8, pp. 25–43, 2014.

[5] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–26.

[6] S. Kim *et al.*, "DATE: Dual attentive tree-aware embedding for customs fraud detection," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2880–2890.

[7] Y. Kültür and M. U. Çağlayan, "Hybrid approaches for detecting credit card fraud," *Expert Syst.*, vol. 34, no. 2, 2017, Art. no. e12191.

[8] J. Vanhoeyveld, D. Martens, and B. Peeters, "Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue," *Pattern Anal. Appl.*, vol. 23, pp. 1457–1477, 2020.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[10] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Advances in Artificial Intelligence – SBIA 2004*, Berlin, Germany: Springer, 2004, pp. 286–295.

[11] S. H. Bach and M. A. Maloof, "Paired learners for concept drift," in *Proc. Int. Conf. Data Mining*, 2008, pp. 23–32.

[12] H. M. Gomes *et al.*, "Adaptive random forests for evolving data stream classification," *Mach. Learn.*, vol. 106, pp. 1469–1495, 2017.

[13] Y. Xu, R. Xu, W. Yan, and P. A. Ardis, "Concept drift learning with alternating learners," in *Proc. Int. Joint Conf. Neural Netw.*, 2017, pp. 2104–2111.

[14] I. Žliobaitė, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with drifting streaming data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 27–39, Jan. 2014.

[15] P. Ren *et al.*, "A survey of deep active learning," *ACM Comput. Surveys*, vol. 54, no. 9, pp. 1–40, Dec. 2022.

[16] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1183–1192.

[17] J. Moon, D. Das, and C.-S. G. Lee, "Multi-step online unsupervised domain adaptation," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, 2020, pp. 41 172–41 576.

[18] Y. Chen, H. Luo, T. Ma, and C. Zhang, "Active online domain adaptation," *ICML Workshop Real World Experiment Design Active Learn.*, pp. 1–25, 2020.

[19] H. Song, M. Kim, S. Kim, and J.-G. Lee, "Carpe diem, seize the samples uncertain "at the moment" for adaptive batch selection," in *Proc. Conf. Inf. Knowl. Manage.*, 2020, pp. 1385–1394.

[20] Y. Zhang *et al.*, "Online adaptive asymmetric active learning for budgeted imbalanced data," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 2768–2777.

[21] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.

[22] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," 2011, *arXiv:1112.5745*.

[23] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2016.

[24] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE Conf. Vis. Pattern Recognit.*, 2019, pp. 93–102.

[25] C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang, "Adaptive region-based active learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2144–2153.

[26] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–13.

[27] F. Zhdanov, "Diverse mini-batch active learning," 2019, *arXiv: 1901.05954*.

[28] A. Kirsch, J. van Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[29] H. Song, S. Kim, M. Kim, and J.-G. Lee, "Ada-boundary: Accelerating DNN training via adaptive boundary batch selection," *Mach. Learn.*, vol. 109, pp. 1837–1853, 2020.

[30] A. Kazerouni, Q. Zhao, J. Xie, S. Tata, and M. Najork, "Active learning for skewed data sets," 2020, *arXiv:2005.11442*.

[31] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.

[32] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[33] N. Buchbinder, K. Jain, and J. S. Naor, "Online primal-dual algorithms for maximizing ad-auctions revenue," in *Proc. Eur. Symp. Algorithms*, 2007, pp. 253–264.

[34] World Customs Organization, "WCO annual report 2019–2020," 2020. [Online]. Available: https://tinyurl.com/y4957n9v

[35] T.-D. Mai, K. Hoang, A. Baigutanova, G. Alina, and S. Kim, "Customs fraud detection in the presence of concept drift," in *Proc. Int. Conf. Data Mining IncrLearn Workshop*, 2021, pp. 370–379.

[36] P. Ball, J. Parker-Holder, A. Pacchiano, K. Choromanski, and S. Roberts, "Ready policy one: World building through active learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 591–601.

[37] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, pp. 212–261, 1994.

[38] L. Xu, M. Skoularido, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional Gan," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[39] X. He *et al.*, "Practical lessons from predicting clicks on ads at facebook," in *Proc. Data Mining Audience Intell. Advertising*, 2014, pp. 1–9.

[40] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. Assoc. Adv. Artif. Intell.*, 2021, pp. 6679–6687.

[41] S. Park, S. Kim, and M. Cha, "Knowledge Sharing via Domain Adaptation in Customs Fraud Detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022

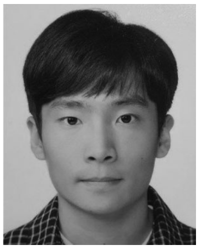
Sundong Kim received the PhD degree from KAIST. He is currently a senior researcher with Data Science Group, Institute for Basic Science and a leading expert with BACUDA initiative, developing fraud detection and category classification algorithms with the World Customs Organization. He has authored or coauthored more than 10 peer-reviewed articles in leading conferences and journals. His research focuses on predictive analytics with real-world data with temporal and imbalanced in nature.

Tung-Duong Mai is currently working toward the undergraduate degree with the School of Computing, KAIST. He has participated in the BACUDA project, developing customs fraud detection algorithms with the World Customs Organization. He worked on developing an algorithm to mitigate concept drift. His research focuses on predictive analytics with machine learning techniques.

Sungwon Han is currently working toward the PhD degree with the School of Computing, KAIST. He has worked on unsupervised learning algorithms for unstructured data and anomaly detection to discriminate out-of-distribution samples from data. He also developed a semi-supervised anomaly detection algorithm to discriminate illicit trades. His research focuses on developing robust representation learning algorithms to deal with data deficiency and corruption, which are common in publicly available datasets.

Sungwon Park is currently working toward the master's degree with the School of Computing, KAIST. He worked on a cross-national customs fraud detection model using domain generalization to support developing countries' customs administration. His research interests include general machine learning theory and machine learning application for social goods.

Thi Nguyen D.K. is currently working toward the undergraduate degree with the School of Computing, KAIST. Her research interests include data science and prediction analytics. She worked on concept drift analysis, experimented with active customs trade selection algorithms.



Jaechan So is currently working toward the undergraduate degree in electrical engineering and computer science with KAIST. His research interests include active learning, efficient and accurate online learning strategy, and data analysis in terms of uncertainty.

Karandeep Singh received the PhD degree from ETRI, Daejeon, in 2019. He is currently a senior researcher with Data Science Group, Institute for Basic Science. Specific areas of application include graph-structured systems at all scales, including interpersonal relationships in a social network, (mis)information flows on the web, and customs flows across different countries. His research interests include the application of computational techniques for networked systems and their information flows.

Meeyoung Cha (Member, IEEE) is currently an associate professor with the School of Computing, KAIST and the chief investigator with the Institute for Basic Science. Her research interests include network and data science with an emphasis on modeling, analyzing complex information propagation processes, machine learning-based computational social science, and deep learning. She has served as the co-editor-in-chief of ICWSM.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.