WT-Shapley: Efficient and Effective Incentive Mechanism in Federated Learning for Intelligent Safety Inspection

Chengyi Yang*, Jia Liu*, Hao Sun, Tongzhi Li, Zengxiang Li*

[1]Digital Research Institute, ENN Group
yangchengyia, liujiaam, sunhaot, litongzhi, lizengxiang @enn.cn

Abstract

Artificial Intelligence (AI) has been widely applied for Safety Inspection in a number of industrial domains. However, individual company usually could not provide sufficient data to support well-trained AI models. Federated Learning, as a new AI paradigm, enables a number of participants to contribute training data to co-create high-performance models without compromising data privacy. However, an effective incentive mechanism is essential to encourage participants to contribute high-quality data, and the fundamental of the incentive mechanism is to evaluate participants' contribution fairly. Shapley Value (SV) is a well-known approach to evaluate individual's marginal contribution in a coalition, but the canonical SV calculation and its available variants are very costly. In this paper, we proposed an FL framework to enable a number of natural gas companies from different cities to jointly train an object detection computer vision deep learning model for the purpose of identifying potential hazards, without sharing their confidential inspection photos directly. We improve state-of-the-art SV calculation algorithms further by proposing weighted truncation (WT) for unnecessary computations, achieving better trade-off between efficiency and accuracy. Based on the proposed WT-Shapley participant contribution evaluation approach, an effective end-to-end incentive mechanism is designed by leveraging knowledge of both data scientists and domain experts. According to our experiments, it could encourage participants to contribute scarce photos with potential hazards, and thus co-create a high-performance AI model to identify various hazards accurately for residential natural gas installation safety inspection.

1 Introduction

Over the past decade, we have witnessed the rapid development of Artificial Intelligence (AI) technology in both academia research and industry deployment. According to recent report released by iResearch Consulting Group, Computer Vision (CV) takes 57% of the entire AI market in China in 2020 (iResearch 2020). CV algorithms, e.g., photo classification, object detection, OCR, face, body and behavior recognition, have been widely applied in various industry domains, including public safety, finance, healthcare, manufacture and energy. In this paper, we focus on safety inspection in energy sector or natural gas supply chain more specifically. An object detection AI model could identify potential hazards in photos, to release the burden of manual inspection and reduce the operation cost of an energy company significantly.

However, it is non-trivial to train an effective deep learning-based AI model, as it has become more and more complex with millions to billions of parameters, requiring huge amount of images and videos with high cost manual annotation. Moreover, the photos with potential hazard, which are essential for model training, is very scarce in real-world applications. As a result, small and medium enterprises usually do not have sufficient resources to train a high performance AI model.

Traditionally, an AI technology provider with sufficient centralized GPU computation power, collects photos from a number of energy companies, trains a super deep learning model, and then provides Cloud services to them or deploys the models in their local data center of edge computing devices. As regulations on data privacy, such as the General Data Protection Regulation (GDPR) in EU (GDPR 2018) and Personal Information Protection Law of the PRC (PIPL 2021), were put into effect. To meet the regulation compliance, energy companies are more and more reluctant to directly share their inspection photos to the technology provider, especially those photos which were taken at their customers' workplace and household. Moreover, energy companies has spent human resources and domain knowledge to collect photos and label the potential hazards, but losing control on how the data would be used after uploading them to the centralized data center. On the another hand, the AI technology provider takes advantages of the data to feed their AI model training, and gets high revenue by applying the well-trained model in the market without notifying the energy companies. Therefore, the traditional solution could not be continued due to more and more stringent regulations and the unfairness in the ecosystem, resulting in 'Data Silos'among energy companies.

In order to break the data silos, Federated Learning (FL), a new paradigm of AI model training, emerges and has received widespread attention in past few years (Yang et al. 2019). FL enables energy companies to train a super AI model in a collaborative manner, while keeping their private data within their own IT infrastructure. The brief steps of a typical horizontal-FL (with overlapping features among

*These authors contributed equally.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

participants) are: 1. Initialization, 2. Participant Selection, 3. Local Training, 4. Secure Aggregation, and 5. Reward Distribution as listed in (Huang et al. 2020).

In addition, Incentive Mechanism using monetary or non-monetary schemes could be adopted to encourage energy companies with different data resources to collaborate in a fair manner. The fundamental of the incentive mechanism is to measure participants' contribution efficiently and accurately. Shapley Value (SV) proposed by Lloyd Shapley who won the Nobel Prize in Economics in 2012 is a well-known solution in Cooperative Game Theory (Shapley 1953). It provides a classic and elegant method for fair distribution of benefits in the case of multiplayer cooperation and has been used in FL to measure participants' contribution to the global model. Although SV is fair and authoritative, the classical SV calculation method needs to evaluate the utilities of all possible coalitions among participants, which is computationally expensive and impractical in FL.

In this paper, we propose weighted truncation (WT) - Shapley algorithm to calculate the approximate SV of FL participants by truncating those coalitions with limited effect on SV of participants, during the iterative AI model training. Hence, it could reduce the costly computation for constructing the sub-models corresponding to participant coalitions and evaluate their accuracy. According to the experimental results on public dataset and real-world application evaluation, our proposed WT-Shapley has advantages over the state-of-the-art algorithms, achieving better trade-off between accuracy and efficiency. It could obtain lower error rate with the same computation cost, and vice versa. Based on WT-Shapley, we propose an end-to-end incentive mechanism, which could encourage FL participants to contribute scarce safety inspection photos with potential hazards. Since they are beneficial for enhancing the performance AI models, our incentive mechanism providing them higher rewards could form a positive cycle for the development of the ecosystem, and thus, boost intelligent hazard identification accuracy and speed up the AI model adoption in real-world applications.

The rest of the paper is organized as follows. Section 2 provides literature review on Shapley Value based participant contribution evaluation approaches and incentive mechanism in Federated Learning. Section 3 introduces our Federated Learning application scenario, i.e, intelligent hazard identification for safety inspection, and its requirement on efficient and effective incentive mechanism. The proposed WT-Shapley and its advantages over state-of-the-art SV algorithm are illustrated in Section 4. The details of our end-to-end incentive mechanism for encouraging the collaborations among natural gas companies are described in Section 5. Experiments and application evaluation results are reported and analyzed in Section 6. Section 7 concludes the paper and outline directions of future work.

2 Related Work

The subject of this paper is related to the incentive mechanism in FL, specifically contribution evaluation.

As in FL training, participants are required to provide computation power and network bandwidth, not to mention the collection and annotation of local data, rewards are indispensable to inspire more participants and maintain a flourish ecosystem.

There are several comprehensive surveys on incentive mechanism in FL. Focused on clients, (Zhan et al. 2021) presents incentive mechanism designs driven by clients' data contribution (including data quality and quantity), reputation, and resource allocation. From the perspective of functional differentiation, (Zeng et al. 2021) summarizes a framework of IM, including node selection, contribution evaluation and payment allocation (not only monetary but also reputation, well-trained model, etc.). Another review (Ali et al. 2021) compares incentive mechanism in terms of design principles and techniques, including contract theory, game theory, auction theory, etc. and highlights the security challenges involved. Also in IEEE guide framework of FL (IEEE 2021), incentive mechanism is regarded as a standalone module and the constraints that need to be considered are listed, and fairness is the most essential. A review of contribution evaluation (Huang et al. 2020) summarizes three major taxonomy contribution measurement strategies, they are test/self-reported based, marginal loss based, and similarity based contribution evaluation. In summary, participant contribution evaluation is a fundamental step.

As FL involves multi-party cooperation, Game Theory naturally becomes a fertile source for reference, and various approaches have been proposed. The profit-sharing scheme can be categorized in three: egalitarian, marginal gain and marginal loss, (Yu et al. 2020).

Also a marginal contribution-based scheme, Shapley Value (SV) is a classic and elegant solution in Cooperative Game Theory (Shapley 1953). SV is proved to be the only solution that satisfies all 4 properties, Efficiency, Symmetry, Dummy and Additivity, which together can be considered a definition of a fair contribution evaluation method (Molnar 2019).

However, the time complexity of canonical SV algorithm is $O(2^n)$, when adopting in FL, a huge amount of re-training and evaluating is introduced, thus the canonical SV is impractical. Recently, a collection of works have been devoted to improving the efficiency of SV calculation. Typical techniques are as follows. 1) **Gradient-based approaches** use gradient aggregation instead of sub-model re-training, e.g., MR (Song, Tong, and Wei 2019). 2) **Sampling-based approaches** apply Monte-Carlo sampling to truncate unnecessary sub-models re-construction and evaluation, e.g., (Castro, G´omez, and Tejada 2008), (Štrumbelj and Kononenko 2014), (Wang, Dang, and Zhou 2019), (Okhrati and Lipani 2020). 3) **Truncation-based approaches** truncates entire unnecessary round. And others like Group Testing (Jia et al. 2019) etc. The complete list can be found in (Liu et al. 2021).

In fact, advanced algorithms utilize a combination of the above techniques, such as TMC (Ghorbani and Zou 2019), TMR (Wei et al. 2020), and the state-of-the-art approach GTG-Shapley (Liu et al. 2021).

As noted in survey (Zeng et al. 2021), incentive mechanism for cross-silo FL are neglected in current studies, however cross-silo is a common setting of FL between enterprises. It should be emphasized that the study reported in

this paper focuses on cross-silo FL and has real-world application verification.

As far as our knowledge goes, the end-to-end incentive mechanism framework proposed is the first that can encourage FL participants to provide scarce data and form a positive cycle.

3 Application Description: Intelligent Hazard Identification for Safety Inspection

In energy sector, safety management has the highest priority and is the most important service quality KPI for company management, as an accident may cause damage to people's lives and properties, affecting the reputation and revenue of the involved companies seriously. Take recent accidents related to residential gas associated to gas pipeline corrosion leakage as an example, on Jun-2021, a gas explosion happened in Shiyan, Hubei, causing 12 deaths and 138 injuries (ChinaDaily 2021). In order to prevent most accidents, energy companies (or nature gas suppliers) conduct safety inspection regularly to identify potential hazards along the entire supply chain, such as gas stations, pipelines, workplaces and households.

This paper focuses on household inspection on natural gas pipeline and device installation as shown in Figure 1. Based on long-term operation experience, domain knowledge and safety management standards, tens of the potential hazards should be considered and checked regularly, including exhaust gas water heater without chimney flue, stove without flame-out protection, pipeline corrosion, un-certificated extension using improper soft tube, and etc.

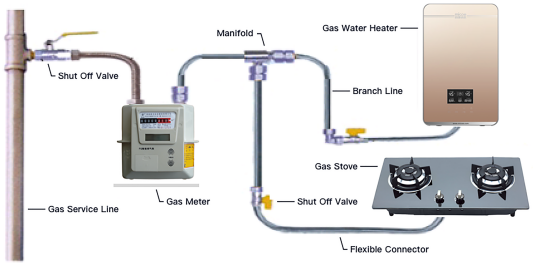

Figure 1: Residential gas piping diagram

Nowadays, many safety inspectors are regularly sent to visit household located in every corner of the city to identify potential hazards in-time, which takes high portion of human resource and operation cost of city-gas companies. The inspectors are required to take photos of pipeline connections and gas devices for auditing and post-event investigation. In order to guarantee the inspection were done properly, another group of backstage inspectors are hired to double-check tons of uploaded photos manually.

A high-performance AI model is desired to assist the inspection procedure in multi-fold manners.

- Release the burden of backstage inspectors to identify potential hazards on the uploaded photos, highlight those were not detected by household inspectors and then send out warning to them and the corresponding residential users.
- Recommend household inspectors the potential hazards once a photo was taken by their smartphones, to improve their work efficiency and quality.
- Encourage residential users to take photos by themselves and provide AI model results immediately, and thus achieve high inspection coverage in-time with much lower operation cost.

The success of an intelligent hazard identification model relies on large amount photos with professional annotation, especially those with potential hazards which are scarce within a small and medium city-gas company. FL enables a number of gas companies to train a high-quality model cooperatively, without compromising data privacy. An incentive mechanism is required to ensure the fairness of the collaboration among companies with different data and computational resources by measuring their contributions and returning monetary rewards accordingly. As mentioned in Section 2, contribution evaluation is fundamental that needs to be well addressed first, and we propose an efficient and accurate algorithm WT-Shapley. Then, the end-to-end incentive mechanism will be presented in Section 5.

4 WT-Shapley: Efficient and Accurate Participant Contribution Evaluation

In this section, we illustrate Weighted Truncation (WT)-Shapley algorithm for efficient participant contribution evaluation in FL.

Suppose there are $N = \{1, \ldots, n\}$ participants, each has its private local dataset $D_i, \forall i \in N$, and participated in T rounds of FL training process. For each round $\forall t \in \{1, \ldots, T\}$, participant i downloads the initial global model $M^{(t-1)}$, and trains several local epochs with its own dataset D_i, gets a local model $M_i^{(t)}$. Then, all participants upload their gradient updates $\{\Delta_i^{(t)} = M_i^{(t)} - M^{(t-1)}\}, \forall i \in N$ to the coordinator server. The coordinator server executes an aggregation algorithm to generate a new global model $M^{(t)}$, it will also be the initial model for the next round. The aggregation algorithm $Agg(\cdot)$ may vary, such as $FedAvg(\cdot)$ (McMahan et al. 2017):

$$M^{(t)} = M^{(t-1)} + \sum_{i=1}^{n} \frac{|D_i|}{\sum_{i=1}^{n} |D_i|} \Delta_i^{(t)}. \tag{1}$$

As mentioned in Section 2, SV is fair and authoritative. In FL of coalition N, a participant i's SV is defined as:

$$
\begin{aligned}
&\phi_i(N,V) = \\
&\frac{1}{|N|!}\sum_{S\subseteq N\setminus\{i\}} |S|!(|N|-|S|-1)!(V(S\cup\{i\})-V(S)) \\
&= \sum_{S\subseteq N\setminus\{i\}} \frac{1}{|N| * \binom{|N|-1}{|S|}} * (V(S\cup\{i\})-V(S)) \\
&= \sum_{S\subseteq N\setminus\{i\}} \quad w_{|S|} \quad * \quad \Delta v_i,
\end{aligned}
\tag{2}
$$

where $V(\cdot)$ is the utility function. $w_{|S|}$ denotes the weight of the sub-coalition S and can be considered as its probability of occurrence, which is only related to the cardinality of S. Δv_i is the marginal utility when i joins S. Usually the utility function $V(\cdot)$ in FL is obtained by measuring the sub-model's performance on the standard validation dataset:

$$
V(S) = V(M_S) = V(Agg(M^{(0)}, \{\Delta_i\})), \forall i \in S. \tag{3}
$$

The state-of-the-art GTG-Shapley has achieved a good balance between computation efficiency and accuracy (Liu et al. 2021). The key steps are summarized as follows. 1) **between-round truncation**, if one round's marginal gain is significantly tiny, the entire round can be omitted. 2) **within-round truncation**, as the larger the size of sub-coalition, the less significant the marginal gain of new entrant i, therefore a considerable number of the sub-coalition's evaluation can be omitted. 3) **guided Monte-Carlo sampling policy**, which emphasizes the importance of the head permutations and improves convergence.

However, we found a flaw in the within-round truncation criteria, which leads to the proposed WT-Shapley.

An alternative definition of SV is the average of marginal utility of participant i in all possible order of joining the coalition N. Let $\pi(N)$ be the set of all permutations of coalition N, the cardinality of $\pi(N)$ is $|N|!$. Given a permutation $O = \{O[1], \ldots, O[n]\} \in \pi(N)$, let $Pre^i(O) = \{O[1], \ldots, O[j-1]\}$ be the set of predecessors of participant i, if $i = O[j]$. Thus, the Eq.2 can be re-written by the following way (Castro, G´omez, and Tejada 2008):

$$
\phi_i(N,V) = \sum_{O\in\pi(N)} \frac{1}{|N|!}[V(Pre^i(O)\cup\{i\}) - V(Pre^i(O))]. \tag{4}
$$

Obviously, given a participant i and a sub-combination $S \subseteq N\setminus\{i\}$, the following equations holds.

$$
\begin{aligned}
\Delta v_i &= V(S\cup\{i\}) - V(S) \\
&= V(Pre^i(O)\cup\{i\}) - V(Pre^i(O)), \\
&\forall O \in \pi_{S,i}(N),
\end{aligned}
\tag{5}
$$

$$
\begin{aligned}
&w_{|S|} * \Delta v_i = \\
&\sum_{O\in\pi_{S,i}(N)} \frac{1}{|N|!}[V(Pre^i(O)\cup\{i\}) - V(Pre^i(O))],
\end{aligned}
\tag{6}
$$

Algorithm	Sub-model reconstruction	Between-round truncation	Within-round truncation
MR	gradient based	-	-
GTG-Shapley	gradient based	yes	Δv_i
WT-Shapley	gradient based	yes	$w_{\|S\|} * \Delta v_i$

Table 1: Comparison of SV approximation algorithms

where $\pi_{S,i}(N)$'s first $|S|$ bits filled by S randomly, the $(|S|+1)$-th bit is i, and the succeeding $(|N|-|S|-1)$ bits are filled by $N\setminus S\setminus\{i\}$ randomly.

The insights are two, 1) regardless of the order of the participants in the combination S, its utility $V(\cdot)$ is equal. There are $p_{|S|} = |S|!(|N|-|S|-1)!$ permutations in total. 2) Superficially, the probability of occurrence of each Δv_i is equal (right part of Eq.6), which is $1/|N|!$, but the total permutations $p_{|S|}$ varies with different size of S. Resulting in the contribution to SV is Δv_i multiplied by the weight $w_{|S|}$ that varies significantly.

In particular, the denominator of $w_{|S|}$ is $|N|$ multiplies $\binom{|N|-1}{|S|}$, a variant of Pascal's Triangle (Yang Hui's Triangle), causing $w_{|S|}$ to decrease sharply as $|S|$ increases when $|S| <= |N|/2$.

However, only marginal gain is considered in the within-round truncation of GTG-Shapley (Liu et al. 2021). Let $\Delta\overline{v}_i = V(N)$–$V(Pre^i(O))$:

$$
\Delta v_i = \begin{cases} 0, & if \Delta\overline{v}_i < \epsilon, \\ V(Pre^i(O)\cup\{i\}) - V(Pre^i(O)), & if \Delta\overline{v}_i >= \epsilon, \end{cases}
\tag{7}
$$

where ϵ denotes truncation threshold. In essence, the truncation criteria in GTG-Shapley treats each $\Delta\overline{v}_i$ equally, which is contrary to the facts.

Derived from the second insight, if considering weight, a more accurate and refined truncation criteria will be obtained. Based on GTG-Shapley (Liu et al. 2021), the complete pseudo code of WT-Shapley is listed in Algorithm 1.

Line 1-3 show parameters initialization, weights calculation is added. Line 5-9 show the same between-round truncation and guided sampling policy as GTG-Shapley (Liu et al. 2021). Line 13-23 show weighted within-round truncation operation, which are also the major difference from GTG-Shapley.

In summary, the main features of WT-Shapley and two typical methods MR and GTG-Shapley are compared in table 1.

As a variant of sampling-based algorithm, WT-Shapley, GTG-Shapley and (Castro, G´omez, and Tejada 2008) shares similar complexity, which is polynomial time if sub-model's constructing and evaluating cost would be zero (Štrumbelj and Kononenko 2014).

Algorithm 1: WT: Weighted Truncation Shapley

Input: For communication round (t), the initial FL model $M^{(t-1)}$, final FL model $M^{(t)}$, evaluation function $V(\cdot)$, participants' gradient updates $\{\Delta_i\}$, aggregation algorithm $Agg(\cdot)$

Parameter: Between-round truncation threshold λ, within-round truncation threshold η

Output: SVs $\phi_i^{(t)}$, for all participants $i \in N$, for communication round (t)

Let $w_{|S|} = 1/(|N| * \binom{|N|-1}{|S|}), \forall\, |S| \in [0, |N|)$
Let $v_0 = V(M^{(t-1)})$, $v_N = V(M^{(t)})$, $k = 0$
Let $\phi_i^{(t)} = 0, \forall i \in N$
 # between-round truncation
if $|v_N - v_0| > \lambda$ **then**
  while convergence criteria not met **do**
    $k = k + 1$
    # guided sampling
    $O^k = \{O^k[1], \ldots, O^k[n]\}$: partial $(n - m)$ permutation of participants $i \in N$
    $v_0^k = v_0$
    for $j = 1, \ldots, |N|$ **do**
      # weighted within-round truncation
      Let $C = \{O^k[1], \ldots, O^k[j]\}$, $S = C \backslash \{O^k[j]\}$
      if $w_{|S|} >= \eta$ **then**
        if $w_{|S|} * \left|v_N - v_{j-1}^k\right| >= \eta * |v_N - v_0|$ **then**
          $\tilde{M}_C^{(t)} = Agg(M^{(t-1)}, \{\Delta_C\})$
          $v_j^k = V(\tilde{M}_C^{(t)})$
        else
          $v_j^k = v_{j-1}^k$ # $v_j^k - v_{j-1}^k = \Delta v_{O^k[j]} \Rightarrow 0$
        end if
      else
        $v_j^k = v_{j-1}^k$
      end if
      $\phi_{O^k[j]}^{(t)} = \frac{k-1}{k} \phi_{O^k[j]}^{(t)} + \frac{1}{k}(v_j^k - v_{j-1}^k)$
    end for
  end while
end if
return $\{\phi_1^{(t)}, \ldots, \phi_n^{(t)}\}$

5 Effective End-to-End Incentive Mechanism

Based on our proposed WT-Shapley, an effective end-to-end incentive mechanism is designed for natural gas household inspection application. As illustrated in Figure 2, the incentive mechanism is seamlessly integrated with the FL framework including AI model training and deployment, as well as the contribution evaluation and monetary reward distribution.

The FL framework contains three parts, which are 1) an open ecosystem with many data contributors, 2) an FL platform operator and 3) model users in the marketplace. It is worth pointing out that a city-gas companies can play the roles of data contributors and model users simultaneously.

The ecosystem of data contributors is open to any city-gas company that intends to use FL AI models to improve its safety inspection service. Participants of the ecosystem collect inspection photos with professional annotation of the listed potential hazards. During the interactive FL model training, they provide computation power for local model training and communication bandwidth for uploading model updates and downloading global model through the interactions with the platform operator.

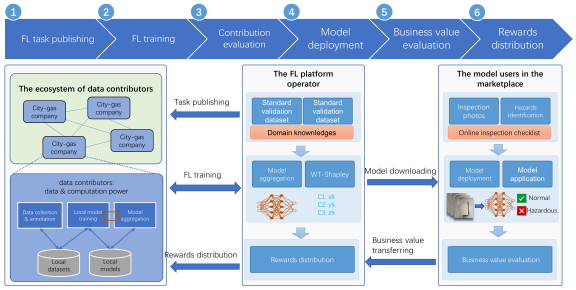

Figure 2: The end-to-end Incentive Mechanism framework

The FL platform operator plays a key role for managing the FL tasks and cultivating the ecosystem. It defines the hazard identification application scenario according to domain knowledge and safety management standards. Meanwhile, a standard validation dataset is created, which contains almost all typical hazards under different situations. The AI model can only be deployed in real-world applications if it achieves the desired accuracy on the validation set. During the iterative FL model training, the operator acts as the coordinator to aggregate the model updates uploaded by the participants and then broadcast the aggregated global model. SV-based algorithms are adopted by the platform operator to calculate the participants' contribution based on how their uploaded model updates affect the performance of the aggregated model over the standard validation dataset. In addition, the participant's computation and communication costs are counted as another dimension of its' contribution to the FL model. Last but not least, the platform operator deploys the AI model to users with payments based on real-time performance, and then distributes the monetary rewards to participants according to their contributions.

The model users in the marketplace can be any city-gas company in the ecosystem or an external company. They trace the AI model's inference results and corresponding business values. Besides rewards to correct inferences, different amount of penalties are applied to false alarms and missed identifications. The rewards and penalties are feedbacked to the platform operator as a reference for model pricing.

The workflow of the incentive mechanism integrated within the FL framework is as follows.

- 1) The platform operator defines a hazard identification application scenario and creates standard validation dataset, then publishes an FL task to the open ecosystem.
- 2) City-gas company participants conduct FL model training with the help of centralized coordination and secure model aggregation service provided by the platform

operator.

- 3) Participants' contributions are evaluated through WT-Shapley and the computational resources used are recorded.
- 4) The FL global model is downloaded by the model user and deployed for real-time inference in real-world application.
- 5) The business value generated by the AI model is evaluated and transferred to the platform operator.
- 6) The platform operator distributes rewards to ecosystem participants according to their contributions.

Since the validation dataset created by domain experts is composed of a variety of potential hazards and updated from time-to-time newly detected potential hazards, the model validation accuracy, e.g., F1 Score used in this paper, is usually consistent to its application performance for most model users. Consequently, participants' SVs calculated based on the validation accuracy of sub-models corresponding to participants coalitions, indicating their contribution to the FL model's business value. In real-world case, a very small portion of household inspection photos have potential hazards. These scarce photos are much more beneficial to improve FL model accuracy than the normal ones. Therefore, participants who contribute more photos with potential hazards tend to get higher SVs, and our designed end-to-end incentive mechanism would provide higher monetary rewards to them accordingly. In this way, a positive cycle can be derived, as shown in Figure 3, encouraging participants to provide more photos with potential hazards for the purpose of promoting the performance of the FL model, accelerating AI model adoption with higher business value obtained, achieving win-win situation among participants and model users within a vibrant ecosystem.

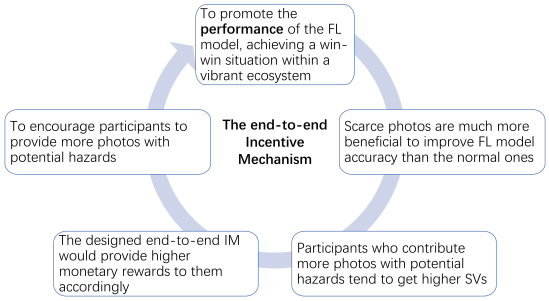

Figure 3: Incentive mechanism derives a positive cycle for a vibrant ecosystem

6 Experiments and Results

In order to evaluate the efficiency and effectiveness of our proposed incentive mechanism, comprehensive experiments are conducted using a public dataset and real-world household inspection photos. Firstly, experiments are carried out using MNIST dataset (LeCun, Cortes, and Burges 2010), to verify that our proposed WT-Shapley has advantage over the state-of-the-art GTG-Shapley, to achieve better trade-off between computation cost and SV calculation accuracy. Secondly, the aforementioned advantage of WT-Shapley is further verified in household inspection application using photos collected by multiple city-gas companies. Thirdly, we prove that a validation dataset that incorporates domain expertise is essential to the effectiveness of our designed incentive mechanism for the formation of a positive cycle in the ecosystem. And finally, we report how an intelligent hazard identification model helps household inspection application, as well as how to quantify its business value and distributing monetary rewards to participants.

6.1 Experiments on MNIST Dataset

As introduced in Section 2, the canonical SV calculation is impractical for FL, and MR algorithm is the most accurate gradient-based approach. Our experiments take MR as the benchmark for measuring the accuracy of SVs calculated by WT-Shapley and GTG-Shapley. Since the computation cost is mainly introduced by constructing the sub-models and evaluating its accuracy over the validation dataset, the computation cost is represented by the number of different sub-models involved in the SV approximate calculation algorithms.

In order to compare the performance of WT-Shapley and GTG-Shapley, experiments are carried out using MNIST dataset divided and assigned to 10 FL participants in same (i.i.d., MNIST-IID) and imbalanced (non-i.i.d., MNIST-NON-IID) distribution scenarios and their performance are reported in Figure 4 and 5 respectively. The left-upper and left-lower sub-figures respectively plot the SV calculation error rate and the number of sub-models with respect to the decreasing threshold, while the right sub-figure reveals the trade-off between the number of sub-models (i.e., computation cost) and the accuracy of participant contribution evaluation.

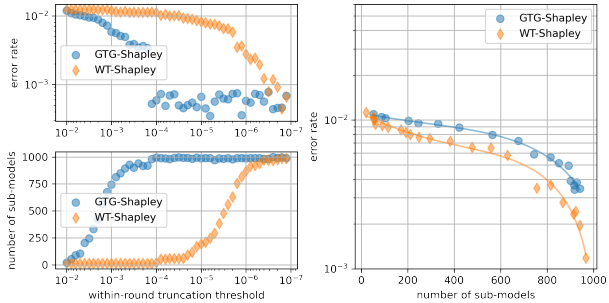

Figure 4: Performance comparison between WT-Shapley and GTG-Shapley using MNIST-IID dataset

As discussed in Section 4, the performance of WT-Shapley and GTG-Shapley algorithms are sensitive to the within-round truncation threshold. In our experiments, the threshold is set at equal intervals in the range of 1.0 and 1e-7 in log_{10} scale. As shown in Figure 4 and 5, when the threshold decreases, the number of sub-models increases and the error rate decreases. It is because that, the smaller threshold, the smaller of the estimated marginal utility of those

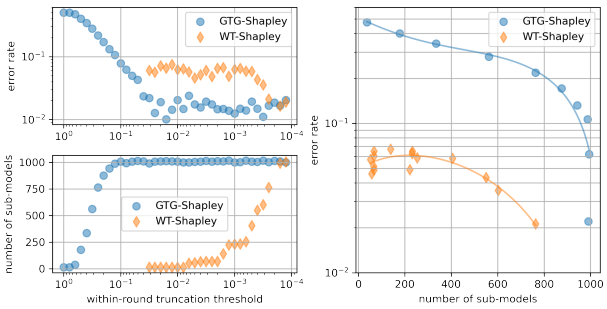

Figure 5: Performance comparison between WT-Shapley and GTG-Shapley using MNIST-NON-IID dataset

truncated coalitions, according to Equ 7. As a result, the approximate Shapley value is more close to the result of MR algorithm, and according the more participant coalition samples involved.

When the threshold is reduced to a certain value, saturation occurs, the number of sub-models or the error rate no longer changes. As the minimum error rate is controlled by convergence criterion of Monte-Carlo sampling (refer to line 6 of algorithm 1).

Aforementioned observations are valid for both WT-Shapley and GTG-Shapley. However, due to the weight introduced in WT-Shapley (line 15 of Algorithm 1), the corresponding thresholds are shifted when the number of sub-models and SV error rate start to change significantly.

As analyzed in Section 4, after introducing the weight of participant coalition, the truncation threshold in our WT-Shapley could correctly implicate the marginal utility in SV calculation. Hence the risk of truncating coalitions with marginal utility higher than expectation could be reduced compared with GTG-Shapley. It is verified by the results reported in the right sub-figures of Figure 4, and 5, which plot the trade-off between computation cost (i.e., the number of sub-models) and SV accuracy (i.e., error rate). The difference between the area enclosed by the two polynomial fitting curves and the coordinate axis can be considered as the advantage of our proposed WT-Shapley, i.e., the lower curve which is closer to the origin of the coordinates with lower computation cost and lower error rate. Intuitively, with the comparable numbers of the sub-models, WT-Shapley could obtained higher SV calculation accuracy; to achieve comparable SV calculation accuracy, WT-Shapley takes less participant coalitions into account.

Table 2, takes a close look at the trade-off by revealing exact value of the number of sub-models and SV calculation in moderate situation. Consistent with the observation on Figure 4 and 5, WT-Shapley achieves better trade-off and its advantages is most significant in MNIST-NON-IID scenario. WT-Shapley achieves an even lower error rate with about 1/5 - 1/2 number of sub-models of GTG-Shapley.

It is an interest observation that the advantages of WT-Shapley over GTG-Shapley is most significant in MNIST-NON-IID scenarios. This might because the weight introduced in WT-Shapley plays a better selection effect, and the sub-models' marginal utilities with greater impact on SV are retained.

	error rate @ comparable number of sub-models		**number of sub-models @ comparable error rate**	
Experiment ID	GTG-Shapley	WT-Shapley	GTG-Shapley	WT-Shapley
MNIST-IID	0.950% @246	**0.797% @211**	677 @0.723%	**383 @0.719%**
MNIST-NON-IID	40.0% @178	**6.71% @139**	1008 @5.00%	**222 @4.90%**

Table 2: Performance of WT-Shapley and GTG-Shapley under comparable conditions of computation cost and SV calculation accuracy

City-gas company	**Normal**	**Hazardous**	**Irrelevant**	**Total**
C1	850	50	100	1000
C2	700	200	100	1000
C3	400	500	100	1000

Table 3: City-gas companies' training dataset in household safety inspection application

In real-world applications, non-i.i.d. is more prevalent due to various sizes of city-gas companies, and imbalance in dataset size or proportion.

6.2 Household Inspection Application Evaluation

The WT-Shapley algorithm and the WT-Shapley -based incentive mechanism are further evaluated by the application of household natural gas safety inspection. More specifically, YOLOv4-tiny object detected model (Wang, Bochkovskiy, and Liao 2021) is trained to detect the water heater and chimney flue on inspection photos. The photo is identified as *irrelevant* if no water heater is detected, as *normal* if both water heater and chimney flue are detected, as *hazardous* if water heater is detected but no chimney flue is detected.

In order to join the FL task initiated by the FL platform operator, each city-gas company annotates its inspection photos manually using a box to indicate the position of water heater and chimney flue, if any, on them. In our experiments, three companies jointed, each of which has *1000* human-annotated photos, including hazardous, normal, and irrelevant categories as shown in Table 3. Note that, there are differences in the distribution of these categories among the three companies, in order to compare the impacts on SV-based contribution to the FL model.

WT-Shapley Algorithm Advantages For both of WT-Shapley and GTG-Shapley algorithms, the between-round truncation criterion is met after 10 rounds of FL model aggregation, which can reduce computation cost dramatically compared with canonical and MR approximate SV algorithms. Different from GTG-Shapley, our proposed WT-Shapley taking the probability of participant coalitions into account. Similar to experiments on public dataset, Figure 6 shows that WT-Shapley could reduce the computation cost

Validation dataset	Normal	Hazardous	Irrelevant	Total
V1 (Standard)	60	480	60	600
V2 (Alternative)	330	210	60	600

Table 4: The standard and alternative validation dataset in household safety inspection application

and improve SV accuracy in the household safety inspection application. For instance, to achieve comparable error rate at around 2.56%, WT-Shapley and GTG-Shapley requires 53 and 49 sub-models respectively, indicating around 7.5% computation cost saving. The advantage of WT-Shapley is expected to be more significant with the increasing number of FL participants.

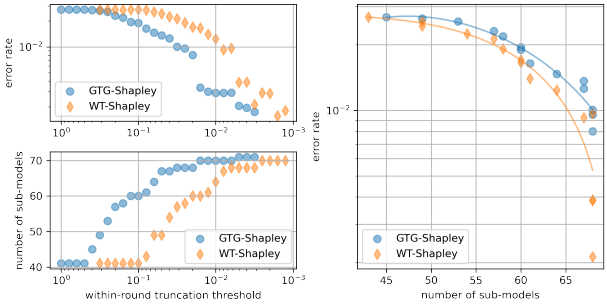

Figure 6: Performance comparison between WT-Shapley and GTG-Shapley using household safety inspection dataset

Effectiveness of Incentive Mechanism As introduced in Section 5, the standard validation dataset (V1), as shown in Table 4, is co-created by domain experts and data scientists based on their knowledge and experience. The proportion of hazardous photos is relatively high for the purposes of 1) verifying that the model could identify most potential hazards under various situations with high business value in the real-world application; 2) keep the dataset within a reasonable size to save the cost of SV calculation and application adoption. In order to illustrate the importance of the standard validation dataset to the effectiveness of our incentive mechanism, we purposely introduce an alternative validation dataset (V2) with more balanced number of normal and hazardous photos, as shown in Table 4.

Refer to Table 3, the percentage of hazardous photos in city-gas company C1, C2 and C3 increases in turn. Consistently, we can observe their increasing Shapley Values, in Figure 7. The experimental results verified our analysis in Section 5, that companies providing more hazardous photos make higher contribution to the FL model, and thus receiving higher monetary rewards. Since our end-to-end incentive mechanism could encourage participants to contribute data more important to FL model, it is effective to form the positive cycle for vibrant ecosystem.

In the case of the alternative validation dataset V2 (refer to Table 4), participants' contribution does not increase with

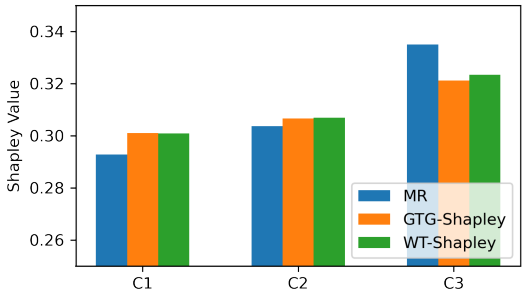

Figure 7: Participant Shapley Values using V1 (**standard** validation dataset) in household safety inspection application

the percentage of hazardous photos as shown in Figure 8. As a result, the incentive mechanism could not encourage city-gas companies to contribute hazardous photos which are important to intelligent hazard identification FL models.

Therefore, it proves that the standard validation dataset is essential, and it is necessary for data scientists and domain experts to work closely to construct a standard validation dataset that can truly reflect business values.

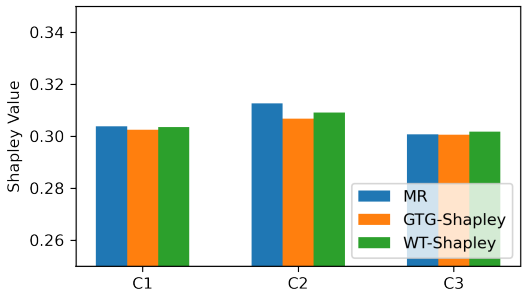

Figure 8: Participant Shapley Values using V2 (**alternative** validation dataset) in household safety inspection application

It is worthy to point out that, as shown in both Figure 7 and 8, WT-Shapley results is more close to MR benchmark than GTG-Shapley, with the comparable computation cost in terms of number of sub-models. For Figure 7, at the same number of sub-models: 49, error rates of GTG-Shapley and WT-Shapley are 2.64% and 2.44% respectively, for Figure 8, at the same number of sub-models: 45, error rates of GTG-Shapley and WT-Shapley are 0.785% and 0.520% respectively.

Application Value of FL Model As one of the biggest natural gas operator in China, ENN Group's business has expanded to more than 200 cities, providing gas services for 24m families and 190k enterprises.

The reported household safety inspection application was launched internally in ENN on Jun 30, 2021, with an average

of 100k calls per day, with accuracy around 90%.

From the following two aspects, the AI model proves its business value. Firstly, the AI model can check hundreds of thousands of photos in full rather than sampling manually, thus 140 hours of labor can be saved for 100k photos inspection. And secondly, the AI model pre-inspects daily incremental photos, reducing the scope of manual inspections dramatically.

Meanwhile, more city-gas companies request the use of AI models to inspect accumulated photos. With application feedback, the standard validation dataset is iteratively enriched and improved, thus the participants' contribution evaluated by WT-Shapley always fits the real business value.

7 Conclusions and Future Work

In this paper, we reported our works and experience of cross-silo FL application on intelligent safety inspection in energy sector. We proposed 1) WT-Shapley, an efficient and improved Shapley Value approximation algorithm based on state-of-the-art approach for participant contribution evaluation; 2) an end-to-end incentive mechanism based on WT-Shapley by leveraging knowledge of both data scientists and domain experts. Since the reported application launched on Jun, 2021, the business values generated by AI model is evaluated and the proposed end-to-end incentive mechanism is verified to be effective, the contributions evaluated by WT-Shapley can truly representing business values.

However, both WT-Shapley and GTG-Shapley are based on Monte-Carlo sampling, which have the common problem of ignoring the best sub-combination, a more efficient method still needs to be explored. In terms of application, we will promote the online application to safety inspectors and residential users, and continue to explore co-create intelligence in other scenarios in energy sector.

References

Ali, A.; Ilahi, I.; Qayyum, A.; Mohammed, I.; Al-Fuqaha, A.; and Qadir, J. 2021. Incentive-Driven Federated Learning and Associated Security Challenges: A Systematic Review.

Castro, J.; G´omez, D.; and Tejada, J. 2008. Polynomial Calculation of the Shapley Value based on Sampling. *Computers & Operations Research*, 36(2009): 1726 – 1730.

ChinaDaily. 2021. Gas explosion leaves 12 dead, 138 hurt.

GDPR. 2018. General Data Protection Regulation.

Ghorbani, A.; and Zou, J. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. arXiv:1904.02868.

Huang, J.; Talbi, R.; Zhao, Z.; Boucchenak, S.; Chen, L. Y.; Stefanie; and Roos. 2020. An Exploratory Analysis on Users' Contributions in Federated Learning. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 20–29.

IEEE. 2021. IEEE Guide for Architectural Framework and Application of Federated Machine Learning. *IEEE Std 3652.1-2020*, 1–69.

iResearch. 2020. iResearch Serial Market Research on China's AI Industry (III).

Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gürel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. J. 2019. Towards Efficient Data Valuation Based on the Shapley Value. In Chaudhuri, K.; and Sugiyama, M., eds., *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, 1167–1176. PMLR.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST Handwritten Digit Database.

Liu, Z.; Chen, Y.; Yu, H.; Liu, Y.; and Cui, L. 2021. GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning. arXiv:2109.02053.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A.; and Zhu, J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 1273–1282. PMLR.

Molnar, C. 2019. *Interpretable Machine Learning*.

Okhrati, R.; and Lipani, A. 2020. A Multilinear Sampling Algorithm to Estimate Shapley Values. arXiv:2010.12082.

PIPL. 2021. Personal Information Protection Law of the People's Republic of China.

Shapley, L. S. 1953. A Value for N-Person Games. *Contributions to the Theory of Games*, 2(28): 307 – 317.

Song, T.; Tong, Y.; and Wei, S. 2019. Profit Allocation for Federated Learning. In *2019 IEEE International Conference on Big Data (Big Data)*, 2577–2586.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2021. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13029–13038.

Wang, G.; Dang, C.; and Zhou, Z. 2019. Measure Contribution of Participants in Federated Learning. 2597–2604.

Wei, S.; Tong, Y.; Zhou, Z.; and Song, T. 2020. *Efficient and Fair Data Valuation for Horizontal Federated Learning*, 139–152. Cham: Springer International Publishing. ISBN 978-3-030-63076-8.

Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; and Yu, H. 2019. *Federated Learning*. Morgan & Claypool Publishers.

Yu, H.; Liu, Z.; Liu, Y.; Chen, T.; Cong, M.; Weng, X.; Niyato, D.; and Yang, Q. 2020. A Fairness-aware Incentive Scheme for Federated Learning. In *Proceedings of the AAAI/ACM Conference on Ai, Ethics, and Society*, 393–399.

Zeng, R.; Zeng, C.; Wang, X.; Li, B.; and Chu, X. 2021. A Comprehensive Survey of Incentive Mechanism for Federated Learning. arXiv:22106.15406.

Zhan, Y.; Zhang, J.; Hong, Z.; Wu, L.; Li, P.; and Guo, S. 2021. A Survey of Incentive Mechanism Design for Federated Learning. Forthcoming.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*, 41: pages647–665.