Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7033626, 13 pages
https://doi.org/10.1155/2022/7033626

Review Article

Blockchain-Based Crowdsourcing Makes Training Dataset of Machine Learning No Longer Be in Short Supply

Haitao Xu,[1] **Wei Wei**,[2] **Yong Qi**,[1] **and Saiyu Qi**[1]

[1]*Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*
[2]*School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China*

Correspondence should be addressed to Yong Qi; qiy@mail.xjtu.edu.cn

Received 13 April 2022; Accepted 6 June 2022; Published 26 July 2022

Academic Editor: Alireza Souri

Recently, machine learning has become popular in various fields like healthcare, smart transportation, network, and big data. However, the labelled training dataset, which is one of the most core of machine learning, cannot meet the requirements of quantity, quality, and diversity due to the limitation of data sources. Crowdsourcing systems based on mobile computing seem to address the bottlenecks faced by machine learning due to their unique advantages; i.e., crowdsourcing can make professional and nonprofessional participate in the collection and annotation process, which can greatly improve the quantity of the training dataset. Additionally, distributed blockchain technology can be embedded into crowdsourcing systems to make it transparent, secure, traceable, and decentralized. Moreover, truth discovery algorithm can improve the accuracy of annotation. Reasonable incentive mechanism will attract many workers to provide plenty of dataset. In this paper, we review studies applying mobile crowdsourcing to training dataset collection and annotation. In addition, after reviewing researches on blockchain or incentive mechanism, we propose a new possible combination of machine learning and crowdsourcing systems.

1. Introduction

In recent years, machine learning (ML) techniques have made a lot of achievements in many fields [1–4]. However, a large number of labelled training dataset is one of the most cores of the technology. The quality, quantity, and diversity of training dataset have significant effects on the generalization capability and accuracy of trained models. Unfortunately, those data mainly originate from web crawlers or are provided by hired workers that leads to two main concerns. First, data originated from web, which may be not reliable, cause some issues on copyright [5] and privacy. Second, the service fee for hiring workers to collect special dataset or annotate dataset may be high. These bottlenecks hinder the further development of machine learning.

Thanks to the development of wireless technology and the popularity of mobile computing, crowdsourcing [6] appeared and the technology has gained great interest and adoption like Upwork [7], Amazon Mechanical Turk [8], and UBER [9]. Meanwhile, crowdsourcing systems are popular in environmental monitoring [10], smart transportation [11, 12], healthcare [13], and online marketplace [14]. The main idea of the technology is to find solutions of problems publicly, and it can easily increase energetic efficiency [3, 15, 16]. In addition, crowdsourcing in mobile computing domain is named mobile crowdsourcing (MCS). In particular, MCS systems dedicate to collecting data with sensors embedded on mobile devices. This technique provides a new idea for solving the problems above.

Now, we introduce a traditional MCS system. A naive MCS system mainly consists of three parties: task requesters, workers, and MCS system. Figure 1 shows the workflow, and the detail is introduced as follows: (1) the task requesters post tasks with payments by MCS system. (2) Workers choose the tasks which they are interested in by MCS system. (3) Workers work for the tasks and submit the results to MCS system. (4) The task requesters receive the results, and MCS system sends the payments to the workers.

In order to overcome the bottlenecks we introduced before, we can replace the above tasks with training dataset

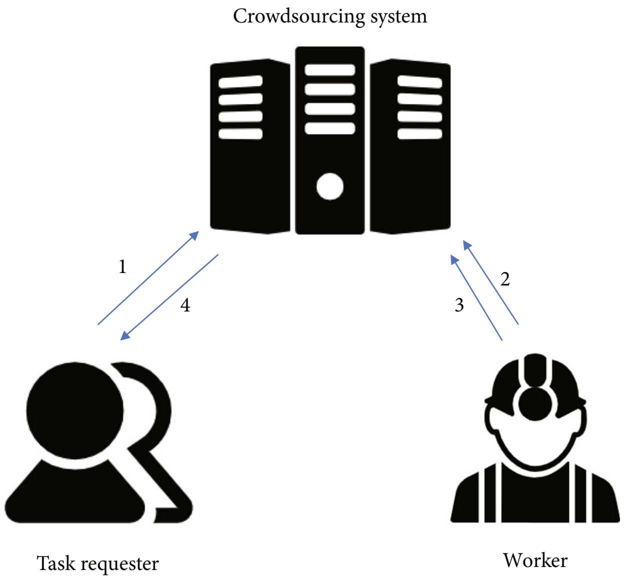

Figure 1: Workflow of traditional crowdsourcing system.

collection tasks or training dataset annotation tasks. If the function of MCS system is to collect training dataset, which we named TDCMCS system (training dataset collection mobile crowdsourcing system), workers need to collect data and annotate them before submitting the results. Otherwise, in the TDAMCS system (training dataset annotation mobile crowdsourcing system), task requesters need to submit the dataset which need workers to annotate. Fortunately, these two types of system already exist in recent years [17, 18], and we will introduce them in Section 3.

In fields of machine learning (ML), the training dataset consists of image dataset [19, 20], video dataset [17], text dataset [21], and other dataset [22]. The corresponding MCS systems have existed, and we will introduce them in Section 3.

However, traditional MCS systems are almost implemented on cloud servers; the detailed researches are shown in Section 3. Unfortunately, centralized platforms face many problems. First, centralized cloud servers inherently suffer from single point of failure and cannot ensure transparent operation, which will result in misbehavior. The order of passengers cannot be stopped at the end of services in April 17, 2015, as a service outage appeared because of hardware failure in Uber Technologies, China [23]. Second, the data (if the system is a crowdsensing system) produced by workers should belong to workers themselves, but when the data is stored on cloud servers, problems like unclear copyright [24] and information leakage [25, 26] will appear, which obviously does not meet the requirements of General Data Protection Regulation (GDPR) [5]. In December 2015, the Office of the Australian Information Commissioner (OAIC) reported that freelancer broke the Privacy Act by revealing true identity of users which included dummy accounts, IP addresses, and active account [27]. Third, centralized cloud servers are easy to be remotely hijacked and suffer mischief attacks and DDoS attacks, which leads to the unavailable services. In May 2014, Elance and oDesk is out of services for many workers because of DDoS attacks [28]. Fourth, the service fee of the centralized crowdsourcing system is usually high; for example, the sliding service fee for five percent to twenty percent is necessary for most of the crowdsourcing systems [7].

To overcome the challenges above, decentralized crowdsourcing systems [29], usually combined with blockchain [30] technology and cryptographic algorithms [24, 31], become a great idea. Blockchain-based CS system often regards smart contracts as a trusted third-party broker, which can automatically execute codes of smart contract, and the whole process is public and traceable, so that various functions can be realized in distributed ways. In fact, the distributed crowdsourcing system based on blockchain has applications in many fields like court processing of adjudication, posting and assigning tasks, data collection, copyright protection, data transactions, mobile health, and energy trading. Even so, there are still some problems existing when combining blockchain-based CS system with machine learning. The core problem is that the quality and quantity of training labelled dataset cannot be guaranteed. To ensure the quality of training data, building reputation systems is a great proposal. In the TDCMCS system, after receiving the training labelled dataset, the task requester can give a feedback for the dataset. The feedback influences the reputation of worker, which is positively related to the payments. Hence, worker will prefer to offer high quality of training labelled dataset. In the TDAMCS system, the task requester should give a feedback on the quality of annotation. Meanwhile, truth discovery [32], one of the carefully designed algorithms [33–35] based on mathematical theory [36], can enhance the quality of annotation. Truth discovery (TD) algorithm is a method to discover truth among unreliable users' information. In order to collect satisfactory quantity of training dataset, incentive method is necessary. The price of training dataset in the marketplace or the payments of hiring workers determines how many workers can be attracted to provide the dataset or annotate the dataset.

Recent researches make it possible to break through the bottlenecks of training dataset mentioned above. Note that, this is not the first survey on crowdsourcing and machine learning. Survey of Alenezi and Faisal [37] used either machine learning, crowdsourcing, or both of them in the proposed solutions which were aimed at surveying the current learning and e-learning schemes. Ørting et al. [38] reviewed studies published until July 2018 which combine crowdsourcing with analysis of the medical images. Jin et al. [39] surveyed the past quality control research of collect knowledge about various types of data items. Konečný et al. [40] described the types of annotation computer vision that researchers have collected using crowdsourcing in their survey.

Different from them, our contributions are shown as follows: (1) we review studies which combine crowdsourcing systems with the analysis of training dataset collection and annotation. (2) We also review some researches on blockchain, incentive mechanism, PPTD, etc., and extract some available techniques and methods which can break through the bottleneck of quality, quantity, and diversity of training dataset. (3) We summarize future development directions, and we propose an implementation of a decentralized (i.e., blockchain) MCS system that has the functions of collecting training datasets and annotating datasets.

Table 1 shows the summary of important abbreviations in our paper. In the rest of paper, we introduce the preliminaries in Section 2, and then, we introduce the related applications in Section 3. Section 5 provides future scope of machine learning with crowdsourcing technology. Next, the paper is concluded in Section 6.

2. Preliminaries

This section introduces all the core technologies and methods mentioned in this paper.

2.1. Machine Learning. Machine learning is designed to simulate human intelligence by learning among the surroundings which is a developing branch of computational algorithms. The core problem of ML is how to construct computers that automatically improve by learning experience [41]. ML is considered to be the core in the new age of big data. In addition, techniques based on ML have been utilized successfully in various fields ranging among computer vision, pattern recognition, finance, biomedical, computational biology, entertainment, medical applications, and spacecraft engineering [42].

In ML, a common task is to study and construct the algorithms which can be learned from data and then make predictions [43]. The algorithms make data-driven predictions or decisions by building mathematical models from input data [44]. These input data are commonly divided in multiple datasets. Particularly, three types of datasets (i.e., training datasets, validation datasets, and test datasets) are usually utilized in different phases of the generation of the model.

Table 1: Summary of abbreviations.

Machine learning	ML
Federated learning	FL
Crowdsourcing	CS
Mobile crowdsourcing	MCS
Training dataset collection mobile crowdsourcing	TDCMCS
Training dataset annotation mobile crowdsourcing	TDAMCS
General data protection regulation	GDPR
Office of the Australian Information Commissioner	OAIC
Truth discovery	TD
Privacy-preserving truth discovery	PPTD
Short message service	SMS

Training dataset is very important in fields of machining learning, but its development is limited because constructing a well-annotated and large-scale dataset is very difficult as the expensive fee of data acquisition and annotation [45].

2.2. Blockchain. A blockchain [46] can be defined as a database which is shared among the users. Blockchain allows its users to trade valuable assets in a pseudonymous and public setup without relying on any central authority or intermediary [47]. A blockchain system consists of three core elements: the block, the chain, and the activity.

Concretely, blocks are storage carriers based on consensus agreements by all validators or stakeholders. The storage contents also capture the interactions of the various parties, such as Bitcoin transactions. Activities can be represented in a service manner in the blockchain system. For instance, digital transactions can be the service contents in Bitcoin. In addition, the chain is the connection of all blocks and is one-way growing. The one-way chain growth is a core property of tamper-resistant.

2.3. Smart Contracts. Szabo [48] initially introduced smart contract in the 1990s. Smart contract originated from the idea to create a technological legal framework which can help businesses, reducing disputes and costs. Smart contracts allow any user to define and execute smart contracts on the blockchain which can maintain the balance and data storage of them. In addition, the access is completely controlled by its code. However, all contract data and balances are public on the blockchain. The program code captures the logic terms of the contract between multiple parties and then defines triggers and response actions. The functions are executed in a smart contract when triggered by events or times. For example, when transactions are added to the blockchain. The creation of autonomous agents is allowed on smart contracts, and their behaviors are completely dependent on the transactions sent to them and their code. Hence, contracts are transparent and decentralized. The financial transaction rules can be enforced without any trusted third-party through the smart contracts.

2.4. Crowdsourcing. Crowdsourcing [49] is a novel model where organizations or individuals obtain services or goods from a relatively open, large, and often rapidly growing group of participants. The services mainly include voting, ideas, finances, and microtasks.

Until 2021, crowdsourcing generally involves using the network (however, online is not always necessary) to attract and assign works between participants and then achieve cumulative results. The word "crowdsourcing" was allegedly created in 2005 [50–52]. Compared with outsourcing, crowdsourcing usually involves a more public and less-specific group [53–55].

Improving costs, speed, flexibility, quality, diversity, or scalability is the advantage of crowdsourcing [56, 57]. A general crowdsourcing method usually includes virtual labor markets, competitions, and public online collaboration. Some crowdsourcing give organizations ways which go beyond the ideas provided just by their employees like LEGO ideas [58, 59]. Amazon Mechanical Turk is one of the crowdsourcing where monotonous "microtasks" are performed by large, paid crowds at the same time. Some unprofitable organizations like Wikipedia have utilized crowdsourcing technology to develop common goods [60].

2.5. Crowdsensing. Crowdsensing [61], which can be called mobile crowdsensing sometimes, is a technology where a large number of crowd share and extract information collectively via their mobile devices (such as smartphones, wearables, and tablet computers). In conclusion, it means crowdsourcing of sensing data from mobile devices.

Raghu et al. coined "mobile crowdsensing" (MCS) in 2011 [62]. MCS belongs to three main types: social (like tracking personal exercise data in a community), infrastructure (like locating potholes), and environmental (like monitoring pollution). Current CS applications operate based on the core assumption that all participants submit the sensing data voluntarily which results in extensive user participation [63]. Also, it can indicate the way users who carry mobile devices form microcrowds based on specific CS activities [64].

2.6. Truth Discovery. In order to achieve the aim of solving conflicts from multiple data sources with noises, truth discovery (TD) is widely utilized in various areas [65]. Although differences exist in the algorithm to compute weights of users and ground truths, the procedure of existing TD algorithms can be concluded as follows: a TD algorithm commonly initializes a random number of ground truths (According to [66, 67], the ground truth value of each object can be also initialized as the average of the sum of the sensory values of the object by different users.) and then conducts truth update and weight update iteratively until convergence. Current TD algorithms almost follow two principles: (1) the data of the user who holds a higher weight has more effect in the execution. (2) The user who provides a more accurate data item will have a higher weight (i.e., it is closer to the aggregated truth).

Due to the concerns of privacy, privacy-preserving truth discovery (PPTD) algorithm appears in recent years. Weight of data provider, data provided by provider, and truth need to be preserved in cipher domain. Weight can be usually used to infer whether the provider is rich or poor. For instance, if a provider has a high weight, it can always provide more accurate data, and accurate data is usually collected by sensors embedded in high-end devices, so the provider is usually a wealthy person who can buy high-end mobile devices. Besides, data and truth belong to individuals and therefore need to be encrypted, and only a few people can decrypt them. By combination with homomorphic cryptographic algorithm or perturbation algorithm, PPTD algorithm implements the TD algorithm calculation in the cipher domain.

3. Related Applications

As shown in Table 2, we introduce papers in different fields with the aim of extracting suitable methods and techniques to make the combination of MCS and machine learning possible. Note that, we focus on the researches of training dataset collection or annotation. Hence, federated learning (FL) is not in our consideration. At the same time, we discuss the weaknesses and advantages of each type of system in Tables 3 and 4.

In fact, crowdsourcing systems can be divided into traditional crowdsourcing [68], mobile crowdsourcing (MCS) [69], vehicle crowdsourcing (VCS), and so on according to the tools used to complete the task. On the other hand, crowdsourcing systems can be divided into centralized crowdsourcing and decentralized crowdsourcing according to the environment in which the system is deployed. Decentralized crowdsourcing can be further divided into blockchain-based crowdsourcing and non-blockchain-based crowdsourcing. In addition, crowdsourcing systems can also be divided according to the task types or application fields, which we will not elaborate here.

3.1. Centralized Crowdsourcing System. The general steps of machine learning are as follows: (1) getting the dataset and preprocessing, (2) selecting the algorithm in machine learning and determining the model, (3) training the model and cross-validation, (4) verification curve and learning curve, (5) training dataset multiple times, (6) testing data test model, and (7) predicting new outcomes. We focus on crowdsourcing systems which can help to get various training dataset.

Swearngin and Yang [70] proposed a centralized crowdsourcing system to collect large-scale data of interface tappability on a variety of mobile applications.

Kuldeep et al. [71] proposed a centralized crowdsourcing system called SMSAssassin, which was aimed at collecting spam mails. The system is effective to filter email spams.

Amini and Yang [72] proposed a centralized crowdsourcing framework named CrowdLearner, which was aimed at generating recognizers by using mobile sensors input automatically. At the same time, the framework assigns the annotation tasks to the crowd.

Trivedi et al. [73] proposed a centralized crowdsourcing system which can assign the sensing tasks to workers. Workers sense the text datasets of mobile phone battery temperature via their mobile devices, and the system can infer ambient temperature.

Table 2: Summary of literatures on various fields.

Item	Related work	Key points
Centralized MCS system with dataset collection or annotation	[70–74]	[70] Training dataset collection of interface tappability [71] Training dataset collection of spam SMSes [72] Training dataset annotation [73] Training dataset collection of temperature [74] Training dataset collection and trading of images
Centralized traditional CS system with dataset collection or annotation	[17, 18, 75–87]	[17] Training dataset collection of videos [75] Training dataset collection of images with labels [76] Training dataset collection of audios [18] Training dataset annotation with three criteria to ensure high quality [77] Training dataset annotation with high quality and low costs [78–87] Other systems for training dataset annotation
Decentralized CS systems without blockchain	[88–91]	[88–91] Task completion in a distributed way but need a centralized system or trusted third-party servers
Decentralized CS systems with blockchain	[92–101]	[92] Court processing of adjudication [93] A general blockchain-based decentralized framework [94] Collaboratively building and updating a dataset [95] Copyright-preserving data trading [96] Data trading based on reverse auction [97] Task-worker matching [98] Healthcare remedy evaluation system [99] Novel mobile health applications [100, 101] Peer-to-peer energy trading platform
Reputation system with blockchain	[102]	[102] Blockchain-based MCS system with reputation system
PPTD	[65, 67, 103–114]	[65, 67, 103–114] PPTD systems implemented on cloud servers
	[115–118]	[115–118] PPTD systems implemented on blockchain

Table 3: Advantages and weakness of the systems.

Item	Transparency	Suffer single point of failure	Suffer remote hijacking	Suffer mischief attacks	Suffer DDoS attacks
Centralized traditional CS system with dataset collection or annotation	No	Yes	Yes	Yes	Yes
Centralized MCS system with dataset collection or annotation	No	Yes	Yes	Yes	Yes
Decentralized CS systems without blockchain	No	Yes/no	Yes/no	Yes/no	Yes/no
Decentralized CS systems with blockchain	Yes	No	No	No	No
Decentralized CS systems with blockchain and PPTD	Yes	No	No	No	No

Table 4: Advantages and weakness of the systems.

Item	Service fee	Data privacy protection	Data quality	Data quantity	Data diversity
Centralized traditional CS system with dataset collection or annotation	High	No	Low	Small	Low
Centralized MCS system with dataset collection or annotation	High	No	Low	Medium	Low
Decentralized CS systems without blockchain	High	No	Low	Small	Low
Decentralized CS systems with blockchain	Low	Yes	Low	Large	High
Decentralized CS systems with blockchain and PPTD	Low	Yes	High	Large	High

Lan et al. [74] proposed a centralized crowdsourcing system named CrowdBuy. It is a privacy-preserving image dataset trading system where a buyer can purchase desired image dataset from available mobile users efficiently, while quality is guaranteed and data ownership and privacy of users are respected.

Next, we introduce some researches on traditional crowdsourcing systems which were aimed at collecting training dataset or annotating dataset. In fact, we can easily convert these systems to MCS systems due to the development of mobile computing.

Sigurdsson et al. [17] proposed a data collection approach named the Hollywood in Homes. The corresponding crowdsourcing system was aimed at collecting dataset of videos, and the annotation works are assigned by workers.

Law and Ahn [75] proposed a research dataset, and it contains one hundred thousand images with English labels, which was generated from a crowdsourcing game named TagATune.

Orlandic et al. [76] presented a centralized CS system which was aimed at collecting dataset of audios. About 3,000 recordings of coughs are labelled by 4 experienced doctors to help diagnosing medical abnormalities. It is one of the largest dataset of cough which is annotated by experts currently.

Hsueh et al. [18] evaluated annotation data from both unprofessional workers hired from the Internet and professional workers in research laboratories. Meanwhile, the impact of noisy annotations on sentiment classification model performance and the utility of annotation selecting on classification efficiency and accuracy are examined.

Chang et al. [77] proposed a centralized crowdsourcing approach named Revolt, the idea of which is to convert expert annotation to crowd-based annotation. Experiments evaluated by them compare traditional crowdsourcing annotation with Revolt and proved that the approach can produce high-quality label with lower financial cost.

Other scholars [78–87] proposed their own crowdsourcing systems for annotate datasets, which will not be described in detail here.

In conclusion, there have been a lot of researches on MCS systems or traditional CS systems for collecting various types of training datasets or annotation. However, these systems often do not take incentives into account. We next focus on the traditional crowdsourcing systems which embed incentives for collecting datasets or some methods of hiring workers.

Zheng et al. [119] proposed a mobile crowdsensing data trading system named ARETE, where the data seller can obtain profit by offering data services after acquiring raw data, while data buyer can buy data services through queries. Additionally, reward sharing and the price design of online data are studied in depth.

Gao et al. [120] proposed a scheme in MCS, which was aimed at solving the recruitment problems of unknown workers. In the scenario, the prior sensing qualities of workers are unknown. Meanwhile, a combined multiarmed bandit problem was modeled to find an optimal solution algorithm.

Hu et al. [121] presented an incentive scheme based on dynamic demand in a mobile crowdsensing systems which is location-dependent. The mechanism changes the payment of sensing tasks dynamically to balance their popularity and demand at each round.

Liu et al. [122] aimed at the problem of online recruitment under the time and budget constraints and proposed a dynamic worker hiring strategy under truthful pricing. In addition, the worker hiring strategy contains an online pricing mechanism based on reverse auction which achieves individual rationality and truthfulness.

It can be seen that if we want to use the MCS system to collect the training data (or hire workers to annotate training dataset) needed for machine learning, incentives are essential. The crowdsourcing trading system or worker recruiting system are feasible. We need to model the specific scenarios (e.g., limited budget, limited time to recruit workers, or unknown quality of workers) by combining specific game theory methods, which will result in reasonable benefits for all parties. As a result, enough workers will participate in the works.

The quality, quantity, and diversity of training dataset have significant effects on the generalization capability and accuracy of trained models. Applying ML in crowdsourcing can easily overcome the bottleneck of the quantity and diversity of training dataset and make it no longer be in short supply.

3.2. Decentralized Crowdsourcing System. Because of the problems of centralized systems we mentioned in Section 1, distributed crowdsourcing systems come into being. As far as we know, few researchers proposed decentralized crowdsourcing systems which were aimed at collecting training dataset or hiring workers to annotate training dataset. Most researches that combine decentralized systems with machine learning focus on training models, i.e., federated learning [40, 123–126]. We first introduce some traditional decentralized crowdsourcing systems without combining blockchain technology.

Ryabinin and Gusev [88] proposed a software library to enable traditional mobile crowdsourcing applications to increase privacy of users without affecting the overall quality of crowdsourcing dataset. Additionally, they proposed Fougere, a decentralized approach, which can send data samples to third-party servers from user devices.

Zhang et al. [89] proposed a protocol named D2, which can help to design a DTN- (delay-tolerant network-) based distributed crowdsourcing system. The goal of them is completing calculation tasks collaboratively and minimizing time consumption.

Yang et al. [90] proposed a task assignment approach scheme that exploits social relations in the crowdsourcing systems. Their approach focused on load balancing in distributed environment.

Cheung et al. [91] proposed a distributed and asynchronous task assignment in the MCS.

Researches above focused on task completion in a distributed way, but they actually utilized a centralized system (or trusted third-party servers) to support services, which runs in the opposite direction of building the crowdsourcing system in a decentralized way.

3.3. Blockchain-Based Crowdsourcing System. The emerging blockchain technology is unique among distributed systems due to its unique advantages. Next, we introduce the decentralized CS systems which is based on blockchain.

Ast and Sewrjugin [92] presented a blockchain-based crowdsourcing application named CrowdJury, which court used to process adjudication. However, they did not provide the details of the crowdsourcing protocols design.

Li et al. [93] presented CrowdBC, a decentralized CS framework based on blockchain, where the tasks can be resolved by a group of workers without relying on any third trusty institution. In addition, privacy of users can be guaranteed, while low transaction fee is required.

Harris and Waggoner [94] presented a decentralized framework for users to build datasets collaboratively and use smart contracts to host updated models continuously. In order to keep the model accurate on some test sets, they proposed both nonfinancial and financial incentive structures to provide good data.

Sheng et al. [95] proposed a blockchain-based crowdsourcing data trading framework called CPchain, which was aimed at preserving copyright. Without any truthful broker, it can construct credible and truthful data trading between mutually untrusted users while guaranteeing the copyright and quality of data.

An et al. [96] presented a blockchain-based CS data trading system. Their highlight is that the system was the path-breaking work which utilized both blockchain and reverse auction to guarantee the truthfulness and fairness of the whole data transaction process.

Meanwhile, due to the existence of unique nonfungible tokens in the blockchain, we advise to combine it with digital copyright to protect data copyright.

Zhang et al. [97] presented a blockchain-based platform named PFcrowd, which performs matching algorithm of encrypted task-worker between different crowdsourcing systems without any trustful third-party institution. The core of their approach is to covert the task recommendation algorithm into a credible smart contract.

J. Park et al. [98] proposed CORUS, a healthcare remedy evaluation system utilizing blockchain-based CS and cloud computing platform. The system is an efficient and effective alternative scheme to the expensive clinical trials and objective evaluation on the remedies.

By combining crowdsourcing with mobile computing, cloud computing, Internet of Things (IoT), and blockchain, Fernández-Caramés et al. [99] proposed a system which stores data of patient and develops novel mobile health applications. The application supports diagnosing, public health actions, studying, and monitoring. In addition, it can help people to control diseases and increase global awareness on popularity of diabetes.

Shen et al. [100] presented a blockchain-based architecture and an optimization model. The goal of the architecture is to manage the operation of crowdsourcing energy systems which enables P2P energy trading. Users on Internet can easily and seamlessly trade energy on the system.

Jamil et al. [101] proposed an energy transaction platform built on blockchain which can provide day-ahead controlling, generation scheduling, and real-time support.

Hence, we recommend emerging blockchain technology to combine with the MCS system to collect training dataset or hire workers to label the dataset.

3.4. Reputation System and Privacy-Preserving Truth Discovery. As we mentioned in Section 1, to ensure high quality of training data, building reputation system is a great idea. Feedback of task requesters is the core of reputation system [127, 128].

K. Zhao et al. [102] proposed BC-MCS, an efficient and credible blockchain-based mobile crowdsensing system, which combined edge computing with the emerging blockchain technology in the MC scenario. The privacy-preserving reputation management approach is the key idea of the proposed system which can resist malicious users.

On the other hand, aggregating multiple annotations, which can make the annotations more accurate, is a great method. Truth discovery is an algorithm which stands out from the crowd.

An et al. [104] presented a light blockchain-based model which can assess quality of data. A couple of quality assessment processes on data are presented in their model. One of them is implemented in data quality assessment, and the other is implemented in the selection of users. Meanwhile, the service fee is low as smart contracts and consensus mechanism are carefully redesigned to be suitable for CS.

Miao et al. [65] proposed a PPTD framework for CS systems on cloud, which can protect the privacy of both sensory data and reliability scores generated by the TD algorithm of users. Their work was achieved by using homomorphic cryptographic scheme and threshold Paillier cryptosystem. By optimizing the algorithm, Zheng et al. [105] greatly reduces the communication and computing overhead. Compared with Miao, Zheng's designs achieved at least 30x and 10x savings on communication and computation of users, respectively.

Works of Cai et al. [106] are on stream data. They presented a CS system which enables PPTD algorithm and blockchain-based full-fledged knowledge monetization. Two noncollusion cloud servers were used to execute PPTD algorithm, and blockchain was used to construct a fair data trade market that does not need trusted third parties.

Tang et al. and Xu et al. [107, 108] presented the noninteractive PPTD systems which did not need users to be online all the time and guaranteed the privacy strongly.

Zheng et al., Xue et al., and Miao et al. [67, 109, 110] proposed novel PPTD systems, which can protect both sensory data and reliability scores of users. Miao's work is on both batch data and stream data.

Gao et al. [111] proposed an efficient and novel location-based PPTD mechanism, which can aggregate data accurately with both data privacy and location privacy of users protecting. In addition, superincreasing sequence techniques are used to guarantee efficiency and feasibility.

Xu et al. [112] presented the first verifiable and PPTD protocol in CS systems named V-PATD. Their openly verifiable approach lets any entity verify whether the aggregated truths returned from the cloud server are correct. Their verification approach is scalable and efficient as the cloud server carries most of the computation burdens.

Wang et al. [103] proposed a CS system which can execute a private and fair truth discovery (PFTD). The system is based on two noncollusion cloud servers and Paillier cryptosytem. Their goal is to not only protect the privacy of each workers but also guarantee the fairness in crowdsourcing.

PPTD algorithm mentioned above is all by using homomorphic cryptographic scheme, as Zhang et al. and Sun et al. [113, 114] utilized the method of adding perturbation to the data to protect privacy. This method greatly reduces the computational cost of PPTD.

The above work is based on the centralized cloud server to realize the PPTD algorithm. Due to the problems of centralization (mentioned in the Section 1), some distributed solutions have emerged one after another.

Liang et al. [115] presented a blockchain-based novel CS model which can control quality of crowdsensing data. In the model, TD algorithm and fuzzy theories are utilized.

By combining zero-knowledge proof, trusted hardware, and differential privacy, Duan et al. [116] presented a blockchain-based CS ecosystem which support data aggregation services which completely guarantee data aggregation correctness, differential privacy, confidentiality, and robustness.

Tian et al. [117] presented a blockchain-based framework which is secure and reliable while supporting PPTD algorithm. Without any central servers, the framework assigns the data processing tasks to decentralized workers, whose behaviors are forced and verified by using the novel smart contract technology. At the same time, because the blockchain cannot protect the privacy of on-chain data, they combined privacy-preserving solution with blockchain for protecting privacy.

Wang [118] proposed PrivSTD, a PPTD mechanism based on edge computing, which can gain high accuracy of truth aggregated by streaming crowdsourcing data with the privacy protection of workers.

There are a large amount of researches on blockchain-based PPTD algorithm. It is an innovative idea to combine with annotating training dataset.

4. Challenges for Combining Machine Learning with Crowdsourcing Technology

In this paper, we focus on crowdsourcing help to supply training dataset and annotate dataset, which lead to many challenges.

4.1. Quality of Training Dataset. Although the combination of ML and crowdsourcing can easily solve the problem of the diversity and quantity of training dataset, the quality of the datasets is difficult to guarantee because the workers in crowdsourcing may be professional or nonprofessional.

4.2. Incentives. If the reward for labeling tasks and collecting datasets is too low, no workers will participate in the work, and if it is too high, the employer's benefits will decrease. How to set the rational reward is a challenge.

In addition, crowdsourcing can help train the models of machine learning, i.e., federated learning (FL), which brings some new challenges.

4.3. Privacy and Security. As an efficient privacy-preserving method, FL can train a lossless ML model through local training and parameter transfer of the participants without directly obtaining the data source. However, there are also many security risks in FL. The main security threats include poisoning attacks, adversarial attacks, and privacy leaks.

4.4. Communication Efficiency. ML algorithms, especially complex deep learning algorithms, need to train a large number of parameters during the training process. For instance, CNN may need to train millions of parameters, and each update process needs to update millions of parameters. In addition, the state of network communication may also lead to high communication costs. For example, unstable network conditions and inconsistent speeds during parameter uploading and downloading will lead to excessive model training costs for the entire algorithm.

4.5. Heterogeneity. In FL system, another challenge is the heterogeneity among the client devices, including differences in storage, CPU computing power, and network transmission. The heterogeneity makes the computing time of the devices different and even cause individual devices to drop directly.

5. Future Scope of Machine Learning with Crowdsourcing Technology

The researchers believe that it has immense potential in many fields when machine learning meets crowdsourcing. We consider that there are two types of blockchain-based systems that hold great promise: TDCMCS system (training dataset collection crowdsourcing system) and TDAMCS system (training dataset annotation crowdsourcing system).

In the TDCMCS system, task requesters post dataset collection tasks on blockchain. Workers receive the tasks and collect training dataset by their sensors embedded on mobile devices. After annotating the training dataset, workers submit the results on blockchain. Finally, task requesters receive the result and give the feedback on the training labelled dataset. A good feedback will enhance the reputation of the worker. By combining reputation system, the quality of datasets can be improved, because the reputation directly affects rewards and whether the workers can be assigned tasks. To attract workers to the tasks, we advise set payment when posting tasks or build an auction system. If the incentive mechanisms are reasonable, the number of dataset will grow explosively. Game theory is an analytical tool which can ensure the rationality of the reputation system and the fairness of the trading system. Truth discovery is a great data aggregation algorithm that can make annotation of training dataset more accurate. In addition, blockchain-based crowdsourcing systems can solve problems such as single points of failure.

In the TDCMCS system, the workflow is a little different from in TDCMCS system. First, task requesters post dataset annotation tasks on blockchain with their dataset public online (e.g., IPFS [129] or NFS [130]). Second, workers get the dataset and annotate it. Third, the annotations will be uploaded on blockchain, and the smart contract will execute PPTD algorithm to make the annotations more accurate.

Finally, the feedback and quality of annotations jointly determine the change in reputation.

Such systems could shine in a large number of domains like environmental monitoring, smart transportation, and healthcare [131]. There will be an explosion of all kinds of high-quality datasets in many domains, leading to rapid advances in machine learning.

6. Conclusion

In this paper, we survey the researches on crowdsourcing systems, machine learning, truth discovery, etc., and propose the possible combination of machine learning and MCS systems. The goal of this survey is to provide a possible guide for combination of crowdsourcing and machine learning to explore potential research directions in related fields and may lead to exciting results.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Ministry of Education of the People's Republic of China under Grant No. 2020KJ010801.

References

[1] W. Wang, X. Zhao, Z. Gong, Z. Chen, N. Zhang, and W. Wei, "An attention-based deep learning framework for trip destination prediction of sharing bike," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4601–4610, 2021.

[2] D. Poap, M. Woniak, W. Wei, and R. Damaeviius, "Multithreaded learning control mechanism for neural networks," *Future Generation Computer Systems*, vol. 87, pp. 16–34, 2018.

[3] A. Zielonka, A. Sikora, M. Woźniak, W. Wei, Q. Ke, and Z. Bai, "Intelligent internet of things system for smart home optimal convection," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4308–4317, 2021.

[4] J. H. Wu, W. Wei, L. Zhang et al., "Risk assessment of hypertension in steel workers based on LVQ and fisher-SVM deep excavation," *IEEE Access*, vol. 7, pp. 23109–23119, 2019.

[5] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," in *A Practical Guide*, vol. 10-no. 3152676pp. 10–5555, Springer International Publishing, Cham, 1st edition, 2017.

[6] H. Jeff, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[7] UpworkFebruary 2022, https://www.upwork.com/.

[8] Amazon mechanical turkFebruary 2022, https://www.upwork.com/.

[9] UberFebruary 2022, https://www.uber.com/.

[10] M. Stevens and E. D'Hondt, "Crowdsourcing of pollution data using smartphones," in *Workshop on ubiquitous crowdsourcing*, pp. 1–4, ACM, 2010.

[11] W. Wang, F. Xia, H. Nie, Z. Chen, and W. Wei, "Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3567–3576, 2021.

[12] C. H. Liu, Z. Dai, H. Yang, and J. Tang, "Multitask-oriented vehicular crowdsensing: a deep learning approach," in *IEEE INFOCOM 2020 -IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2020.

[13] P. McCartney, "Crowdsourcing in healthcare," *MCN: The American Journal of Maternal/Child Nursing*, vol. 38, no. 6, p. 392, 2013.

[14] S. Gaikwad, D. Morina, R. Nistala et al., "Daemo: a self-governed crowdsourcing marketplace," in *Adjunct proceedings of the 28th annual ACM symposium on user interface software & technology*, pp. 101-102, 2015.

[15] Y. Liu, X. Sun, W. Wei, and W. Jing, "Enhancing energy-efficient and QoS dynamic virtual machine consolidation method in cloud environment," *IEEE Access*, vol. 6, pp. 31224–31235, 2018.

[16] H. Dou, Y. Qi, W. Wei, and H. Song, "Carbon-aware electricity cost minimization for sustainable data centers," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 211–223, 2017.

[17] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*, pp. 510–526, Cham, 2016.

[18] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: a study of annotation selection criteria," in *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pp. 27–35, 2009.

[19] G. Chen, C. Li, W. Wei, W. Jing, and R. Damaeviius, "Fully convolutional neural network with augmented atrous spatial pyramid pool and fully connected fusion path for high resolution remote sensing image segmentation," *Applied Sciences*, vol. 9, no. 9, p. 1816, 2019.

[20] Q. Ke, J. Zhang, W. Wei, R. Damaševičius, and M. Woźniak, "Adaptive independent subspace analysis of brain magnetic resonance imaging data," *IEEE Access*, vol. 7, pp. 12252–12261, 2019.

[21] L. Gao, S. Biderman, S. Black et al., "The pile: an 800gb dataset of diverse text for language modeling," 2020, https://arxiv.org/abs/2101.00027.

[22] M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 766–773, Boston, MA, USA, 2011.

[23] Uber china statement on service outageFebruary 2022, http://shanghaiist.com/2015/04/18/uber\chinese\operations\recently\hacked.php/.

[24] W. Wei, M. Woniak, R. Damasevicius, X. Fan, and Y. Li, "Algorithm research of known-plaintext attack on double random phase mask based on WSNs," *Journal of Internet Technology*, vol. 20, no. 1, pp. 39–48, 2019.

[25] W. Wei, S. Liu, W. Li, and D. Du, "Fractal intelligent privacy protection in online social network using attribute-based encryption schemes," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 736–747, 2018.

[26] S. W. Chen, D. L. Chiang, C. H. Liu et al., "Confidentiality protection of digital health records in cloud computing," *Journal of Medical Systems*, vol. 40, no. 5, pp. 1–12, 2016.

[27] FreelancerFebruary 2022, http://www.smh.com.au/business/freelancer-contests-20000-privacy-breach-finefrom-oaic-20160112-gm4aw2.html.

[28] Elance and odesk hit by ddosFebruary 2022, https://gigaom.com/2014/03/18/elance-hit-by-major-ddosattack-downing-service-for-many-freelancers/.

[29] D. Guo, J. Xie, X. Zhou, X. Zhu, W. Wei, and X. Luo, "Exploiting efficient and scalable shuffle transfers in future data center networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 4, pp. 997–1009, 2015.

[30] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," *Decentralized Business Review*, no. article 21260, 2008.

[31] S. Qi, Y. Lu, W. Wei, and X. Chen, "Efficient data access control with finegrained data protection in cloud-assisted IIoT," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2886–2899, 2021.

[32] X. Yin, J. Han, and S. Yu Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[33] W. Wei, H. Song, H. Wang, and X. Fan, "Research and simulation of queue management algorithms in ad hoc networks under DDoS attack," *IEEE Access*, vol. 5, pp. 27810–27817, 2017.

[34] J. Chen, D. Chenglie, Y. Zhang, P. Han, and W. Wei, "An adaptive clustering-based algorithm for automatic path planning of heterogeneous UAVs," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2021.

[35] S. Xia, D. Peng, D. Meng et al., "A fast adaptive k-means with no bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 1–99, 2020.

[36] H. M. Srivastava, Y. Zhang, L. Wang, P. Shen, and J. Zhang, "A local fractional integral inequality on fractal space analogous to Anderson's inequality," *Abstract and Applied Analysis*, vol. 2014, 7 pages, 2014.

[37] H. S. Alenezi and M. H. Faisal, "Utilizing crowdsourcing and machine learning in education: literature review," *Education and Information Technologies*, vol. 25, no. 4, pp. 2971–2986, 2020.

[38] S. Ørting, A. Doyle, A. van Hilten et al., "A survey of crowdsourcing in medical image analysis," 2019, https://arxiv.org/abs/1902.09159.

[39] Y. Jin, M. Carman, Y. Zhu, and Y. Xiang, "A technical survey on statistical modelling and design methods for crowdsourcing quality control," *Artificial Intelligence*, vol. 287, article 103351, 2020.

[40] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: strategies for improving communication efficiency," 2016, https://arxiv.org/abs/1610.05492.

[41] M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[42] I. El Naqa and M. J. Murphy, *What Is Machine Learning?*, Springer International Publishing, Cham, 2015.

[43] F. Provost and R. Kohavi, "Glossary of terms," *Journal of Machine Learning*, vol. 30, no. 2/3, pp. 271–274, 1998.

[44] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4, Springer, 2006.

[45] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, V. Kurková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds., pp. 270–279, Springer International Publishing, Cham, 2018.

[46] S. Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System*, Technical report, Manubot, 2019.

[47] M. Risius and K. Spohrer, "A blockchain research framework," *Business & Information Systems Engineering*, vol. 59, no. 6, pp. 385–409, 2017.

[48] N. Szabo, "Smart contracts: building blocks for digital markets," *EXTROPY: The Journal of Transhumanist Thought*, vol. 18, no. 2, 1996.

[49] CrowdsourcingFebruary 2022, https://en.wikipedia.org/wiki/Crowdsourcing/.

[50] E. Estellés-Arolas and F. G.-L.-d. Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information Science*, vol. 38, no. 2, pp. 189–200, 2012.

[51] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Anatomy of a crowdsourcing platformusing the example of microworkers.com," in *2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing*, pp. 322–329, Seoul, Korea (South), 2011.

[52] E. Schenk and C. Guittard, "Crowdsourcing: what can be outsourced to the crowd, and why," in *Workshop on open source innovation*, vol. 72, p. 3, Strasbourg, France, 2009.

[53] D. C. Brabham, *Crowdsourcing*, Mit Press, 2013.

[54] D. C. Brabham, "Crowdsourcing as a model for problem solving," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[55] J. Prpic and P. Shukla, "Crowd science: measurements, models, and methods," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 4365–4374, Koloa, HI, USA, 2016.

[56] R. Buettner, "A systematic literature review of crowdsourcing research from a human resource management perspective," in *2015 48th Hawaii International Conference on System Sciences*, pp. 4609–4618, Kauai, HI, USA, 2015.

[57] J. Prpić, A. Taeihagh, and J. Melton, "The fundamentals of policy crowdsourcing," *Policy & Internet*, vol. 7, no. 3, pp. 340–361, 2015.

[58] W. Liu, J. Moultrie, and S. Ye, "The customer-dominated innovation process: involving customers as designers and decision-makers in developing new product," *The Design Journal*, vol. 22, no. 3, pp. 299–324, 2019.

[59] D. Schlagwein and N. Bjorn-Andersen, "Organizational learning with crowdsourcing: the revelatory case of lego," *Journal of the Association for Information Systems*, vol. 15, no. 11, pp. 754–778, 2014.

[60] A. Taeihagh, "Crowdsourcing, sharing economies and development," *Journal of Developing Societies*, vol. 33, no. 2, pp. 191–222, 2017.

[61] CrowdsensingFebruary 2022, https://en.wikipedia.org/wiki/Crowdsensing/.

[62] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[63] Y. Wang and J. Ma, *Mobile Social Networking and Computing: A Multidisciplinary Integrated Perspective*, CRC Press, 2014.

[64] A.-S. K. Pathan, *Crowd Assisted Networking and Computing*, CRC Press, 2018.

[65] C. Miao, W. Jiang, L. Su et al., "Cloud-enabled privacy-preserving truth discovery in crowd sensing systems," in

Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, pp. 183–196, 2015.

[66] Q. Li, Y. Li, J. Gao et al., "A confidence-aware approach for truth discovery on long-tail data," *Proceedings of the VLDB Endowment*, vol. 8, no. 4, pp. 425–436, 2014.

[67] Y. Zheng, H. Duan, and C. Wang, "Learning the truth privately and confidently: encrypted confidence-aware truth discovery in mobile crowdsensing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2475–2489, 2018.

[68] X. Liu, P. Zou, W. Zhang et al., "CPSFS: a credible personalized spam filtering scheme by crowdsourcing," *Wireless Communications and Mobile Computing*, vol. 2017, 9 pages, 2017.

[69] Y. Qiao, J. Zhang, Q. He et al., "Truthful profit maximization mechanisms for mobile crowdsourcing," *Wireless Communications and Mobile Computing*, vol. 2022, 13 pages, 2022.

[70] A. Swearngin and L. Yang, "Modeling mobile interface tappability using crowdsourcing and deep learning," in *Artificial Intelligence for Human Computer Interaction: A Modern Approach*, pp. 73–96, Springer, 2021.

[71] K. Yadav, P. Kumaraguru, A. Goyal, A. Gupta, and V. Naik, "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering," *12th Workshop on Mobile Computing Systems and Applications*, pp. 1–6, 2011.

[72] S. Amini and L. Yang, "Crowdlearner: rapidly creating mobile recognizers using crowdsourcing," *26th annual ACM symposium on User interface software and technology*, pp. 163–172, 2013.

[73] A. Trivedi, P. Bovornkeeratiroj, J. Breda, P. Shenoy, J. Taneja, and D. Irwin, "Phone-based ambient temperature sensing using opportunistic crowdsensing and machine learning," *Sustainable Computing: Informatics and Systems*, vol. 29, article 100479, 2021.

[74] Z. Lan, Y. Li, X. Xiang, X. Y. Li, and L. Qiang, "Crowdbuy: privacy-friendly image dataset purchasing via crowdsourcing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, USA, 2018.

[75] E. Law and L. Von Ahn, "Input-agreement: a new mechanism for collecting data using human computation games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1197–1206, 2009.

[76] L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.

[77] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: collaborative crowdsourcing for labeling machine learning datasets," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2334–2346, 2017.

[78] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in twitter data with crowdsourcing," in *Proceedings of the NAACL Workshop on Creating Speech and Text Language Data With Amazon's Mechanical Turk*, 2010.

[79] H. Fromreide, D. Hovy, and A. Søgaard, "Crowdsourcing and annotating NER for Twitter# drift," *LREC*, pp. 2544–2547, 2014.

[80] S. Park, P. Shoemark, and L.-P. Morency, "Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization," in *Proceedings of the 19th international conference on Intelligent User Interfaces*, pp. 37–46, 2014.

[81] M. Lease, "On quality control and machine learning in crowdsourcing," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[82] T. W. Cenggoro, F. Tanzil, A. H. Aslamiah, E. K. Karuppiah, and B. Pardamean, "Crowdsourcing annotation system of object counting dataset for deep learning algorithm," *IOP Conference Series: Earth and Environmental Science*, vol. 195, no. 1, article 012063, 2018.

[83] B. M. Good, M. Nanis, C. Wu, and A. I. Su, "Microtask crowdsourcing for disease mention annotation in PubMed abstracts," in *Pacific Symposium on Biocomputing Co-Chairs*, pp. 282–293, World Scientific, 2014.

[84] J. Yang, J. Fan, Z. Wei, G. Li, T. Liu, and X. du, "Cost-effective data annotation using game-based crowdsourcing," *Proceedings of the VLDB Endowment*, vol. 12, no. 1, pp. 57–70, 2018.

[85] S. Hantke, Z. Zhang, and B. W. Schuller, "Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world," *Interspeech*, pp. 3951–3955, 2017.

[86] P. Roit, A. Klein, D. Stepanov et al., "Controlled crowdsourcing for highquality qa-srl annotation," 2019, https://arxiv.org/abs/1911.03243.

[87] S. Hao, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *The 4th Human Computation Workshop, HCOMP@AAAI 2012*, vol. WS-12-08 of *AAAI Technical Report*, Toronto, Ontario, Canada, 2012.

[88] M. Ryabinin and A. Gusev, "Towards crowdsourced training of large neural networks using decentralized mixture-of-experts," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3659–3672, 2020.

[89] S. Zhang, J. Wu, and S. Lu, "Minimum makespan workload dissemination in DTNs: making full utilization of computational surplus around," *The fourteenth ACM international symposium on Mobile ad hoc networking and computing*, pp. 293–296, 2013.

[90] P. Yang, Q. Li, Y. Yan et al., ""Friend is treasure": exploring and exploiting mobile social contacts for efficient task offloading," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5485–5496, 2016.

[91] M. H. Cheung, R. Southwell, F. Hou, and J. Huang, "Distributed time-sensitive task selection in mobile crowdsensing," *The 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 157–166, 2015.

[92] F. Ast and A. Sewrjugin, *The crowdjury, a crowdsourced justice system for the collaboration era*, Technical report, 2015, August 2015, https://www. http://weusecoins.com/assets/pdf/library/The.

[93] M. Li, J. Weng, A. Yang et al., "Crowdbc: a blockchain-based decentralized framework for crowdsourcing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 6, pp. 1251–1266, 2019.

[94] J. D. Harris and B. Waggoner, "Decentralized and collaborative AI on blockchain," in *2019 IEEE international conference on blockchain (Blockchain)*, pp. 368–375, Atlanta, GA, USA, 2019.

[95] D. Sheng, M. Xiao, A. Liu, X. Zou, and S. Zhang, "Cpchain: a copyright-preserving crowdsourcing data trading framework based on blockchain," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, Honolulu, HI, USA, 2020.

[96] B. An, M. Xiao, A. Liu, G. Gao, and H. Zhao, "Truthful crowdsensed data trading based on reverse auction and blockchain," in *Database Systems for Advanced Applications*, G. Li, J. Yang, J. Gama, J. Natwichai, and Y. Tong, Eds., pp. 292–309, Springer International Publishing, Cham, 2019.

[97] C. Zhang, Y. Guo, H. Du, and X. Jia, "Pfcrowd: privacy-preserving and federated crowdsourcing framework by using blockchain," in *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*, Hang Zhou, China, 2020.

[98] J. Park, S. Park, K. Kim, and D. Lee, "Corus: blockchain-based trustworthy evaluation system for efficacy of healthcare remedies," in *2018 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 181–184, Nicosia, Cyprus, 2018.

[99] T. M. Fernández-Caramés, I. Froiz-Míguez, O. Blanco-Novoa, and P. Fraga-Lamas, "Enabling the internet of mobile crowdsourcing health things: a mobile fog computing, blockchain and IoT based continuous glucose monitoring system for diabetes mellitus research and care," *Sensors*, vol. 19, no. 15, p. 3319, 2019.

[100] W. Shen, A. F. Taha, J. Wang, K. Kvaternik, and A. Hahn, "Energy crowdsourcing and peer-to-peer energy trading in blockchain-enabled smart grids," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 8, pp. 1612–1623, 2019.

[101] F. Jamil, N. Iqbal, S. Ahmad, and D. Kim, "Peer-to-peer energy trading mechanism based on blockchain and machine learning for sustainable electrical power supply in smart grid," *IEEE Access*, vol. 9, pp. 39193–39217, 2021.

[102] K. Zhao, S. Tang, B. Zhao, and Y. Wu, "Dynamic and privacy-preserving reputation management for blockchain-based mobile crowdsensing," *IEEE Access*, vol. 7, pp. 74694–74710, 2019.

[103] Z. Wang, X. Cheng, S. Sen, and L. Wang, "Achieving private and fair truth discovery in crowdsourcing systems," *Security and Communication Networks*, vol. 2022, 15 pages, 2022.

[104] J. An, J. Cheng, X. Gui et al., "A lightweight blockchain-based model for data quality assessment in crowdsensing," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 84–97, 2020.

[105] Y. Zheng, H. Duan, X. Yuan, and C. Wang, "Privacy-aware and efficient mobile crowdsensing with truth discovery," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 1, pp. 121–133, 2020.

[106] C. Cai, Y. Zheng, and C. Wang, "Leveraging crowdsensed data streams to discover and sell knowledge: a secure and efficient realization," in *38th IEEE International Conference on Distributed Computing Systems, ICDCS 2018*, pp. 589–599, Vienna, Austria, 2018.

[107] X. Tang, C. Wang, X. Yuan, and Q. Wang, "Non-interactive privacy-preserving truth discovery in crowd sensing applications," in *2018 IEEE Conference on Computer Communications, INFOCOM 2018*, pp. 1988–1996, Honolulu, HI, USA, 2018.

[108] G. Xu, H. Li, and R. Lu, "Practical and privacy-aware truth discovery in mobile crowd sensing systems," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2312–2314, 2018.

[109] K. Xue, B. Zhu, Q. Yang, N. Gai, D. S. Wei, and N. Yu, "Inpptd: a lightweight incentive-based privacy preserving truth discovery for crowdsensing systems," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4305–4316, 2021.

[110] C. Miao, W. Jiang, L. Su et al., "Privacy-preserving truth discovery in crowd sensing systems," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 1, pp. 1–32, 2019.

[111] J. Gao, F. Shaojing, Y. Luo, and T. Xie, "Location privacy-preserving truth discovery in mobile crowd sensing," in *29th International Conference on Computer Communications and Networks, ICCCN 2020*, pp. 1–9, Honolulu, HI, USA, 2020.

[112] G. Xu, H. Li, S. Xu et al., "Catch you if you deceive me: verifiable and privacy-aware truth discovery in crowdsensing systems," in *ASIA CCS '20: The 15th ACM Asia Conference on Computer and Communications Security*, pp. 178–192, Taipei, Taiwan, 2020.

[113] C. Zhang, C. Xu, L. Zhu, Y. Li, and H. Wu, "An efficient and privacy-preserving truth discovery scheme in crowdsensing applications," *Computers Security*, vol. 97, article 101848, 2020.

[114] P. Sun, Z. Wang, Y. Feng, L. Wu, and Z. Wang, "Towards personalized privacy-preserving incentive for truth discovery in crowdsourced binarychoice question answering," in *IEEE International Conference on Computer Communications (INFOCOM 2020)*, Toronto, ON, Canada, 2020.

[115] D. Liang, J. An, J. Cheng, Y. He, and R. Gui, "The quality control in crowdsensing based on twice consensuses of blockchain," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, UbiComp/ISWC 2018 Adjunct*, pp. 630–635, Singapore, 2018.

[116] H. Duan, Y. Zheng, Y. Du, A. Zhou, C. Wang, and M. H. Au, "Aggregating crowd wisdom via blockchain: a private, correct, and robust realization," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, Kyoto, Japan, 2019.

[117] Y. Tian, J. Yuan, and H. Song, "Secure and reliable decentralized truth discovery using blockchain," in *7th IEEE Conference on Communications and Network Security, CNS 2019*, pp. 1–8, Washington, DC, USA, 2019.

[118] C. Wang, *Privacy-Preserving Truth Discovery in Mobile Crowdsensing: Challenges, Solutions, and Opportunities*, 2018.

[119] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "Arete: on designing joint online pricing and reward sharing mechanisms for mobile data markets," *IEEE Transactions on Mobile Computing*, vol. 19, no. 4, pp. 769–787, 2020.

[120] G. Gao, J. Wu, M. Xiao, and G. Chen, "Combinatorial multi-armed bandit based unknown worker recruitment in heterogeneous crowdsensing," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2020.

[121] J. Hu, Z. Wang, J. Wei et al., "Towards demand-driven dynamic incentive for mobile crowdsensing systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4907–4918, 2020.

[122] W. Liu, Y. Yang, E. Wang, and J. Wu, "Dynamic user recruitment with truthful pricing for mobile crowdsensing," in *IEEE INFOCOM 2020 – IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2020.

[123] P. Kairouz, H. B. McMahan, B. Avent et al., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[124] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[125] K. Bonawitz, H. Eichner, W. Grieskamp et al., “Towards federated learning at scale: system design,” *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.

[126] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, “Federated learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[127] J. Yang, H. Wang, Z. Lv et al., “Multimedia recommendation and transmission system based on cloud platform,” *Future Generation Computer Systems*, vol. 70, pp. 94–103, 2017.

[128] Z. Ji, H. Pi, W. Wei, B. Xiong, M. Wozniak, and R. Damasevicius, “Recommendation based on review texts and social communities: a hybrid model,” *IEEE Access*, vol. 7, pp. 40416–40427, 2019.

[129] J. Benet, “Ipfs-content addressed, versioned, p2p file system,” 2014, https://arxiv.org/1407.3561.

[130] S. Shepler, B. Callaghan, D. Robinson et al., *Rfc3530: Network File System (Nfs) Version 4 Protocol*, 2003.

[131] J. Li, G. Deng, W. Wei, H. Wang, and Z. Ming, “Design of a real-time ECG filter for portable mobile medical systems,” *IEEE Access*, vol. 5, pp. 696–704, 2017.