Collaborative Self Organizing Map with DeepNNs for Fake Task Prevention in Mobile Crowdsensing

Murat Simsek, Senior Member, IEEE, Burak Kantarci, Senior Member, IEEE, Azzedine Boukerche, Fellow, IEEE

Abstract—Mobile Crowdsensing (MCS) is a sensing paradigm that has transformed the way that various service providers collect, process, and analyze data. MCS offers novel processes where data is sensed and shared through mobile devices of the users to support various applications and services for cutting-edge technologies. However, various threats, such as data poisoning, clogging task attacks and fake sensing tasks adversely affect the performance of MCS systems, especially their sensing, and computational capacities. Since fake sensing task submissions aim at the successful completion of the legitimate tasks and mobile device resources, they also drain MCS platform resources. In this work, Self Organizing Feature Map (SOFM), an artificial neural network that is trained in an unsupervised manner, is utilized to pre-cluster the legitimate data in the dataset, thus fake tasks can be detected more effectively through less imbalanced data where legitimate(fake tasks ratio is lower in the new dataset. After pre-clustered legitimate tasks are separated from the original dataset, the remaining dataset is used to train a Deep Neural Network (DeepNN) to reach the ultimate performance goal. Pre-clustered legitimate tasks are appended to the positive prediction outputs of DeepNN to boost the performance of the proposed technique, which we refer to as pre-clustered DeepNN (PrecDeepNN). The results prove that the initial average accuracy to discriminate the legitimate and fake tasks obtained from DeepNN with the selected set of features can be improved up to an average accuracy of 0.9812 obtained from the proposed machine learning technique.

Index Terms—Mobile Crowdsensing, Machine Learning, Self Organizing Feature Map, Deep Neural Networks, Fake Task Prevention, Sensing as a Service,

I. INTRODUCTION

Mobile crowdsensing (MCS) has been integrating a useful data collection capability into the smart environments by using embedded sensing capabilities of smart devices in either opportunistic or participatory manner [1]. Smart devices equipped with various sensors are the building blocks of large scale sensing networks [2] and mobility and sociability directly affect the context of MCS data [3]. MCS use cases are various such as healthcare [4], [5], public safety [6], transportation, traffic, and environmental monitoring [7].

MCS platforms exploit the ubiquity of smart mobile devices and assume that their sensing, processing, and communication resources are readily available anytime and anywhere. Interrupting the resource allocation adversary affects the task compilation and resource ratios by keeping user and platform utilities as high as possible during MCS campaign. Malicious tasks can be designed in an intelligent way to consume the resources of the devices as well as MCS servers in the platform [8]. There have been prior solutions that aims to prevent the malicious users from deceiving the MCS platform using game theoretic strategies [9]. The literature lacks thorough studies on fake tasks, hence a detailed investigation is required. However, there are few useful studies, for instance, the authors in [10] detect the fake tasks that target the attack surface of the MCS servers or devices, through machine learning algorithms. In the event of a successful fake task submission attempt, fake tasks can drain the resources of both participating devices and MCS servers. Recent studies [10]–[12] have pointed out that ensemble techniques, if succeed in detecting and eliminating the fake tasks from the MCS platform, can achieve remarkable battery savings against such attacks, and lead to significant savings in battery lifetimes of the smart devices.

Previous works empowered a knowledge-based prior knowledge input method that employed a deep neural network (DeepNN) with four hidden layers in combination with a Self Organizing Feature Map (SOFM) [8], [13] in order to enhance legitimacy detection performance in [14]. Moreover, the DeepNN performance had been shown to improve when combined with ReliefF-based sequential feature selection [15]. The motivation in this study is to cope with the class imbalance between legitimate and illegitimate tasks through early detection and elimination of legitimate tasks by empowering an unsupervised method in the training dataset. Thus, the test performance of the legitimacy detection model can be further improved. In the proposed technique, SOFM is used of pre-clustering legitimacy tasks for both training and test datasets. As remaining mixed (legitimate and fake) tasks in the training and test datasets can be used for training and test of DeepNN model, it is referred to as pre-clustered DeepNN (PrecDeepNN). After the test performance of PrecDeepNN is completed, pre-clustered legitimate-only (PrecL) tasks in the test dataset are combined with test performance of PrecDeepNN to obtain the PrecDeepNN/PrecL model over all results. Although SOFM is utilized in MCS applications such as intelligent attack anticipation [8], [13], [16], this is the first time SOFM driven pre-clustering of legitimate tasks is proposed to boost the overall performance of the DeepNN.

The paper is organized in six sections. Section II presents the motivation and the recent related studies. Section III introduces the characteristics of submitted tasks and their generation process. Section IV presents the proposed methodology in detail. Section V summarizes the simulation results and verifies the performance of the proposed methodology. Finally, concluding remarks and future work are presented in Section VI.
II. RELATED WORK AND MOTIVATION

Despite their efficiency, MCS systems are vulnerable due to relying on the uncertainty of the availability of the distributed sensing devices and their heterogeneity [17] and potential sensing inaccuracy [18]. The study in [19] studies the performance of machine learning classifiers against data poisoning attacks. Additionally, the authors of [20] address intelligent malicious attacks with ensemble learning methods via context recognition through machine learning techniques. There are also recent studies that investigate identification and filtering of fake MCS tasks applying machine learning solutions [10], [21]. To ensure the security of MCS platforms, deep learning methods have also been studied [15], [22].

The contribution of this paper is transformative compared to the previous studies that were incremental improvements of illegitimate task detection: In this study, by applying SOFM-based clustering for early detection of legitimate tasks, the number of legitimate tasks is reduced in the training dataset. By doing so, the class imbalance of the training dataset is mitigated so to improve the training performance with the ultimate goal of better performance on the test set. Once the SOFM-based clustering is applied to the test dataset by aiming at detecting legitimate only tasks (i.e., true positives) before pursuing prediction through the trained model. After applying all machine learning strategies, the overall performance of the proposed methodology outperforms the baseline performance.

III. LEGITIMATE AND FAKE CROWDSSENSING TASKS GENERATION IN MCS PLATFORM

Fake tasks are injected by adversaries to maliciously affect the performance of MCS campaigns. Thus, not only participants experience as dissatisfied but also MCS resources, such as smartphone sensors and computational capacity of MCS platform, are drained by adversarial attacks. To create defense mechanism against fake tasks, Crowdsensim simulator [23] is used to simulate a realistic environment. Using this simulator, fake tasks should be designed to demonstrate the real impact on the MCS system. To achieve this purpose, fake tasks have been modelled by taking into account certain features as listed in Table I in [21].

TABLE I: Simulation settings for task generation

| Feature Selection |
|-------------------|
| Hour, Minute, Day, Resources, Coverage, On Peak Hour, Grid Number |

| Task Features | Battery consumption of smartphone |
|---------------|----------------------------------|
| 80% in {10%; 10% in {1% -6%} |

All possible features for submitted tasks in MCS can be seen in Fig. 1. The location of each task is defined as GPS coordinates such as ‘latitude’ and ‘longitude’. Time-based features such as ‘day’, ‘hour’ and ‘minute’, determine the task-specific time requirements. The ‘day’ feature indicates the date among 6 days of the test week. Since this feature is varies from 1 to 6, and not specify distinctive property of legitimate and fake tasks, it is not used as one of the task’s features. ‘Duration’ specifies the activation time of each task in minutes. "Resources” defines the required battery level which is the percentage of consumption value of a smartphone by the tasks. 'Legitimacy’ is a binary label to identify legitimate tasks from fake tasks by using 1 for legitimate and 0 for fake tasks. ‘Grid Number’ defines grid based location of tasks that are covered by pre-defined square area. Moreover, to show the impact of busy communication time between 7am and 11am, “on-peak hours” is added as binary value referring whether the tasks is appeared in this time or not. To sum up all feature set, fake tasks are produced through statistical distribution in Table I.

IV. SELF ORGANIZING FEATURE MAP (SOFM) Driven PRE-CLUSTERED DEEP NEURAL NETWORK (PREDNN) APPROACH

Machine learning has been mostly used as a multi-disciplinary technique to design intelligent strategy embedding the engineering design problems. These designs can be seen as intelligent attack or defence strategies in MCS systems. In this paper, machine learning methodology has been applied to detect legitimacy of MCS tasks which try to affect the performance of legitimate tasks during MCS campaigns. To obtain baseline classification performance as explained in the previous study [14], ReliefF feature selection method is firstly applied to 10 meaningful features of the MCS dataset as depicted in Fig. 1 then a Deep Neural Network (DeepNN) with 4 hidden layers outputs the average accuracy result of ten runs which are required to eliminate the random initial weights of DeepNN training process. SOFM was used for prior knowledge to reduce the complexity of DeepNN. In this study, SOFM emerges clustering performance to partition the legitimate-only tasks and mixed tasks with legitimate and fake. General workflow of the proposed methodology can be seen in Fig. 2 regarding the
algorithmic details. Each DeepNN structure is trained 10 times with random initial conditions whereas the second part of the flow chart provides details about how to apply an unsupervised method to detect legitimate tasks to mitigate imbalance of legitimate and fake tasks. Further details are given in the following subsections.

A. Self Organizing Feature Map (SOFM) for pre-clustered Data

SOFM is an artificial neural network-based model that is trained in an unsupervised manner; thus, multiple clusters of similar data can be covered by a pre-defined number of neurons. Neurons in a SOFM are organized as 2D and they are updated through weights which are equal to the number of inputs. When a sample is applied to SOFM, inputs of each sample and weights of each neuron are compared to each other through a distance-based metric. The winner neuron which is the closest neuron to the inputs is called as Best Matching Unit (BMU). The weights of BMU are updated to the closer to inputs of sample data. Therefore, SOFM organizes itself according to the whole samples in the dataset. Algorithmic details about SOFM can be seen in [8], [14].

As seen in Fig. 2 SOFM generates two sets of clusters: legitimate-only and mixed. The latter contains legitimate and fake tasks together for the training dataset. As seen in Fig. 3 similar clustering results are obtained by training and testing data. Mixed dataset that contains legitimate and fake tasks together is utilized in the pre-clustered Deep Neural Network (PrecDeepNN) to yield a more specific prediction output on the test set. Although fake tasks are always outnumbered by the legitimate tasks in the MCS dataset (which is a realistic assumption), less imbalanced data can be obtained from mixed data through extracting pre-clustered legitimate-only data out of the training data.

SOFM can generate legitimate-only data through clusters \{1, 3, 6, 7, 9, 10, 11, 12, 14, 15\} so legitimate classification is realized by pre-clustered legitimate tasks process as shown in Fig. 2. In addition to clustering legitimate-only data, SOFM can generate mixed data through clusters \{2, 3, 4, 8, 13, 16\} to train the pre-clustered DeepNN as depicted in Fig. 2. The mixed data consists of legitimate and fake tasks; however, legitimate data rates are mitigated in this new training dataset so fake tasks can be classified better than the original training dataset. In the test phase, SOFM generates not only legitimate-only tasks which are used for classification performance of the test data but also

![Flowchart](image-url)
Output of the PrecDeepNN model after the training process and given the pre-clustered inputs by SOFM is defined in (3) where $Y_{PrecDeepNN}^T$ indicates trained PrecDeepNN model.

$$Y_{PrecDeepNN} = PrecDeepNN(x_{fc,SOFM}^{T})$$  \hspace{1cm} (3)

Legitimate and fake task prediction results of proposed PrecDeepNNPrecL model are depicted in (4) and (5), respectively. $P_L\{PrecDeepNN_T\}$ indicates legitimate task prediction of the PrecDeepNN model and $L_{Prec}$ indicates the pre-clustered legitimate test data by SOFM in (4). $P_F\{PrecDeepNN_T\}$ denotes fake task prediction of the PrecDeepNN model in (5).

$$Leg\_tasks = P_L\{PrecDeepNN_T (x_{fc,SOFM}^{T})\} + L_{Prec}$$ \hspace{1cm} (4)

$$Fake\_tasks = P_F\{PrecDeepNN_T (x_{fc,SOFM}^{T})\}$$ \hspace{1cm} (5)

The general structure of the proposed method and input-output relationships can be seen in Fig. 4.

V. PERFORMANCE EVALUATION

In this study, Timmins region in Canada is chosen to collect tasks generation data from MCS platform. Crowdsensim that is a mobile crowd sensing simulator generates 10 features for submitted tasks. While the centers of fake task submission zones with 200m radius are selected randomly, fake/legitimate task proportion is set to 1/9 for generating 4,000 sensing tasks. The total number of data is 14306 in the six-days campaign and the first 80% of the total data is used for training dataset (11,145) and remaining 20% of the total data is selected for test dataset. Since test dataset is not selected inside the training dataset, more challenging extrapolation performance can be obtained from test dataset to demonstrate the efficiency of the proposed methodology.

Simulation results have been obtained by MATLAB code generating environment and the neural network Toolbox. latitude, longitude, coverage and grid number features has been selected through ReliefF sequential feature selection algorithm is used in this study by following the same setting in our previous work [14]. In addition, DeepNN is modelled with four hidden layers where each hidden layer has 15 neurons that was [14]. Since training and test sets are totally different than each other, all numerical results have been collected as extrapolation performance.

SOFM clustering results show that 5,217 pre-clustered legitimate-only data can be separated from total training data. Remaining 4,550 legitimate and 1,378 fake tasks data constitute 5,928 mixed data to train the PrecDeepNN model. In the test phase, 1,428 PrecL data points which are pre-clustered as legitimate-only before test phase are appended to the prediction output of the PrecDeepNN, thus proposed PrecDeepNNPrecL model testing performance can be obtained from the whole process. Before testing performance of PrecDeepNN, 1,331 legitimate and 401 fake tasks are evaluated by the model to
generate the prediction results. All numerical values on pre-clustering performance of SOFM are presented in Table II.

The test performance of PrecDeepNN and PrecDeepNNPrecL are compared in Fig. 5 to demonstrate the contribution of PrecL that is appended to PrecDeepNN to obtain the proposed PrecDeepNNPrecL results. Overall testing improvement between the initial accuracy of DeepNN and final accuracy of PrecDeepNNPrecL is demonstrated in Fig. 6. This figure verifies the efficiency of the proposed methodology since 0.9735 average accuracy is increased to 0.9812 after applying the PrecDeepNNPrecL model.

VI. CONCLUSION

Task submission and data collection in MCS have led to vulnerabilities against adversaries. Sensing resources of participant devices and MCS platforms can be drained by fake task submissions. In this paper, the supervised DeepNN model is combined with an unsupervised SOFM method to improve the task legitimacy detection performance of the MCS system. Pre-clustered legitimate-only tasks are initially clustered by

| Dataset          | Tasks       | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | pre-clustered |
|------------------|-------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|---------------|
|                 | Legitimate-only | 418| 320| 466| 653| 702| 1056| 203| 237| 603| 559| 5217|
| Training Data    | Legitimate   | 640| 989| 863| 841| 923| 294| 124| 1378|
|                  | Fake         | 100| 435| 182| 245| 292| 292| 124| 4550|
|                  |              |    |    |    |    |    |    |    | 1362|
| Test Data        | Legitimate-only | 133| 87 | 142| 156| 184| 289| 90 | 106| 139| 103| 1429|
|                  | Legitimate   | 199| 239| 245| 294| 95 | 46 | 133| 401 |
|                  | Fake         | 44 | 105| 47 | 64 | 95 | 46 | 133| 401 |

Fig. 4: Pre-clustering based DeepNN process for training and test. Cluster numbers are just for illustrative purposes.

Fig. 5: In the proposed methodology, pre-clustered DeepNN results as average accuracy of 0.9656 is improved to 0.9812 as average accuracy of pre-clustered Deep-NN with pre-clustered legitimate tasks through adding pre-clustered legitimate-only tasks which are obtained by SOFM clusters 1,3,6,7,9,10,11,12,14,15 in Table II.
SOFM, and the remaining (i.e., mixed) data are used to train the DeepNN model. Test results of obtained from the PrecDeepNN model can predict fake and legitimate tasks with an average accuracy of 0.9656. Pre-clustered legitimate-only test data that are initially clustered by SOFM are appended to the prediction results of the PrecDeepNN so PrecDeepNNPrecL has achieved an overall average accuracy of 0.9812. Hence, the initial average accuracy of DeepNN (0.9735) is improved up to an average of 0.9812 by the proposed PrecDeepNNPrecL.

The improvement of the proposed methodology is sufficient to demonstrate the efficiency of PrecDeepNNPrecL, however our ongoing agenda includes studying the detection performance impact of different structures as well as various distributions of legitimate and fake tasks in the MCS platform. In addition to these, our ongoing agenda includes implementation of the proposed defense mechanism on a realistic test bed under a real application such as vehicular crowdsensing. Furthermore, collaboration between fake task submitters have not been considered in our threat model, which remains an open issue that is currently included in our ongoing research agenda. Last but not least, incorporation of a game theoretic model between the MCS platform and the task submitters can help eliminate the submission of fake tasks even before the detection module. Therefore, building on the existing literature, we are planning to bridge a game theoretic incentive-based approach and the proposed fake task detection scheme in the extension of this study.

ACKNOWLEDGMENT

This work was supported in part by the the Natural Sciences and Engineering Research Council of Canada (NSERC) under the DISCOVERY Program, Canada Research Chairs Program and NSERC CREATE TRANSIT Project.

REFERENCES

[1] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry, “A survey on mobile crowdsensing systems: Challenges, solutions and opportunities,” IEEE Communications Surveys Tutorials, vol. 21, no. 3, pp. 2419–2465, 2019.

[2] Y. Zhang and C. Chen, “Secure heterogeneous data deduplication via fog-assisted mobile crowdsensing in 5g-enabled iiot,” IEEE Transactions on Industrial Informatics, pp. 1–1, 2021.

[3] M. Girolami, D. Belli, S. Chessa, and L. Foschini, “How mobility and sociality reshape the context: A decade of experience in mobile crowdsensing,” Sensors, vol. 21, no. 19, 2021.

[4] Y. Guo, B. Guo, Y. Liu, Z. Wang, X. Ouyang, and Z. Yu, “Crowdsafe: Detecting extreme driving behaviors based on mobile crowdsensing,” in IEEE SmartWorld, pp. 1–8, Aug 2017.

[5] Y. Liu, L. Kong, and G. Chen, “Data-oriented mobile crowdsensing: A comprehensive survey,” IEEE Communications Surveys Tutorials, vol. 21, pp. 2849–2885, thirdquarter 2019.

[6] R. F. El Khatib, N. Zorba, and H. S. Hassanein, “Rapid sensing-based emergency detection: A sequential approach,” Computer Communications, vol. 159, pp. 222–230, 2020.

[7] P. Huang, X. Zhang, L. Guo, and M. Li, “Incentivizing crowdsensing-based noise monitoring with differentially-private locations,” IEEE Transactions on Mobile Computing, pp. 1–1, 2019.

[8] Y. Zhang, M. Simsek, and B. Kantarci, “Empowering self-organized feature maps for ai-enabled modeling of fake task submissions to mobile crowdsensing platforms,” IEEE Internet of Things Journal, vol. 8, no. 3, pp. 1334–1346, 2021.

[9] L. Xiao, Y. Li, G. Han, H. Dai, and H. V. Poor, “A secure mobile crowdsensing game with deep reinforcement learning,” IEEE Transactions on Information Forensics and Security, vol. 13, pp. 35–47, Jan. 2018.

[10] M. Simsek, B. Kantarci, and A. Boukerche, “Utility-aware legitimacy detection of mobile crowdsensing tasks via knowledge-based self-organizing feature map,” IEEE Transactions on Mobile Computing, pp. 1–1, 2021.

[11] M. Simsek, B. Kantarci, and Y. Zhang, “Detecting fake mobile crowdsensing tasks: Ensemble methods under limited data,” IEEE Vehicular Technology Magazine, vol. 15, no. 3, pp. 86–94, 2020.

[12] Y. Zhang, M. Simsek, and B. Kantarci, “Ensemble learning against adversarial ai-driven fake task submission in mobile crowdsensing,” in IEEE Intl. Conf. on Communications (ICC), pp. 1–7, 2020.

[13] X. Chen, M. Simsek, and B. Kantarci, “Locally reconﬁgurable self organizing feature map for high impact malicious tasks submission in mobile crowdsensing,” Internet of Things, vol. 12, p. 100297, 2020.

[14] M. Simsek, B. Kantarci, and A. Boukerche, “Self-organizing feature map-integrated knowledge-based deep network against fake crowdsensing tasks,” in IEEE Global Communications Conf., pp. 1–6, 2020.

[15] A. Sood, M. Simsek, Y. Zhang, and B. Kantarci, “Deep learning-based detection of fake task injection in mobile crowdsensing,” in IEEE Global Conf. on Signal and Information Processing (GlobalSIP), IEEE, 2019.

[16] K. Quintal, E. Kara, M. Simsek, B. Kantarci, and H. Viktor, “Sensory data-driven modeling of adversaries in mobile crowdsensing platforms,” in IEEE Global Communications Conf. (GLOBECOM), pp. 1–6, 2019.

[17] F. Khan, A. U. Rehman, J. Zheng, M. A. Jan, and M. Alam, “Mobile crowdsensing: A survey on privacy-preservation, task management, assignment models, and incentives mechanisms,” Future Generation Computer Systems, vol. 100, pp. 456 – 472, 2020.

[18] X. Dong, Z. You, T. H. Luan, Q. Yao, Y. Shen, and J. Ma, “Towards data poisoning attacks in crowd sensing systems,” in ACM Intl. Symp. on Mobile Ad Hoc Networking and Computing, pp. 111–120, 2018.

[19] S. Gisdakis, T. Giannetos, and P. Papadimitratos, “Security, privacy, and incentive provision for mobile crowd sensing systems,” IEEE Internet of Things Journal, vol. 3, pp. 839–853, Oct. 2016.

[20] Z. Chen, M. Simsek, and B. Kantarci, “Federated learning-based risk-aware decision to mitigate fake task impacts on crowdsensing platforms,” in ICC 2021 - IEEE International Conference on Communications, pp. 1–6, 2021.

[21] L. Xiao, D. Jiang, D. Xu, W. Su, N. An, and D. Wang, “Secure mobile crowdsensing based on deep learning,” China Communications, vol. 15, pp. 1–11, Oct. 2018.

[22] C. Fiandrino, A. Capponi, G. Cacciarelli, D. Kliazovich, U. Sorger, P. Bouvry, B. Kantarci, F. Granelli, and S. Giordano, “CrowdSimSim: A simulation platform for mobile crowdsensing in realistic urban environments,” IEEE Access, vol. 5, pp. 3490–3503, 2017.