Bridging Predictive Analytics and Mobile Crowdsensing for Future Risk Maps of Communities Against COVID-19

Sedevizo Kielienyu
skiel067@uottawa.ca
University of Ottawa
Ottawa, ON

Burak Kantarci
burak.kantarci@uottawa.ca
University of Ottawa
Ottawa, ON

Damla Turgut
turgut@cs.ucf.edu
University of Central Florida
Orlando, FL

Shahzad Khan
shahzad@gnowit.com
Gnowit Inc.
Kanata, ON

ABSTRACT

Crowd monitoring and management is an important application of Mobile Crowdsensing (MCS). The emergence of COVID-19 pandemic has made the modeling and simulation of community infection spread a vital activity in the battle against the disease. This paper provides insights for the utility of MCS to inform the decision support systems combating the pandemic. We present an MCS-driven community risk modeling solution against COVID-19 pandemic with the support of smart mobile device users (i.e., MCS participants), who opt-in to crowdsensing campaigns and grant access to their mobile device's built-in sensors (including GPS). Each community is defined by the spatio-temporal instances of MCS participants that are clustered based on the projected future movements of these participants. The MCS platform keeps track of the mobility patterns of the participants and utilizes unsupervised machine learning (ML) algorithms, more specifically k-means, Hidden Markov Model (HMM), and Expectation Maximization (EM) to predict a risk score of COVID-19 community spread for each community ahead of time. Through numerical results from simulating a metropolitan area (e.g., Paris), it is shown that communities' COVID-19 risk scores at the end of a set of MCS campaign can be predicted 20% ahead of time (i.e., upon completion of 80% of the MCS time commitments) with a dependability score up to 0.96 and an average of 0.93. Further tests with a larger population of participants show that community risk scores can be predicted 20% ahead of time with a dependability score up to 0.99 and an average of 0.98.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; • **Networks → Location based services**; • **Computing methodologies → Cluster analysis**.

KEYWORDS

Mobile crowdsensing, Internet of Things (IoT), COVID-19, crowd monitoring, unsupervised learning, machine learning

ACM Reference Format:
Sedevizo Kielienyu, Burak Kantarci, Damla Turgut, and Shahzad Khan. 2020. Bridging Predictive Analytics and Mobile Crowdsensing for Future Risk Maps of Communities Against COVID-19. In *18th ACM International Symposium on Mobility Management and Wireless Access (MobiWac'20), November 16–20, 2020, Alicante, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3416012.3424624

1 INTRODUCTION

Mobile crowdsensing is a natural utility due to the wide availability of non-dedicated sensors by encouraging smart mobile device users to share their sensor data from their devices [1]. The sensors are non-dedicated in that they are not used exclusively by a single application, are not locked up during operation, and are available to be utilized for other purposes. Crowd management and public health [7] are two application areas that can benefit from mobile crowdsensing.

With the surge of COVID-19, Internet of Things (IoT) sensing has been considered a viable solution to model outbreaks by obtaining social contact and/or epidemic networks. Chamola et al. [3] have presented a comprehensive review of the emerging technologies, including IoT and Artificial Intelligence, to battle the COVID-19 pandemic. The availability of IoT sensing will enable a better understanding of the community spread phenomenon, given that up to 30% of the positive cases may remain asymptomatic [8]. With this in mind, Simsek and Kantarci [14] proposed utilizing MCS-assisted and machine learning-based easily relocatable mobile assessment centers to deliver frictionless low-cost testing capability in districts with emerging clusters of confirmed cases based on daily reports. In that particular study, the authors aim to mimic the behavior of a Self-Organizing Feature Map that is proposed to model adversarial settings in a mobile crowdsensing environment [25].

Risk monitoring and assessment for communities, particularly vulnerable populations, is of paramount importance amid the COVID-19 pandemic. As reported by Zhang et al. [24], aggregated public data reveal insights such as infrastructural or demographic details that lead to useful analysis and decision-making features. Predicting the near future risks of community infection spread to help decision-makers in government and healthcare agencies is an open

MobiWac'20, November 16–20, 2020, Alicante, Spain

ACM ISBN 978-1-4503-8119-2/20/11…$15.00
https://doi.org/10.1145/3416012.3424624

area of research. It requires a spatiotemporal extraction and analysis of acquired data from clusters of individuals. Therefore, MCS campaigns are strong candidates to obtain such data for the prediction of future spatiotemporal risks.

In this paper, we review existing IoT sensing solutions for community health monitoring to pave the way for MCS to help with community infectious disease outbreaks. We also propose a novel scheme that employs MCS to obtain predictive COVID-19 risk scores at the community level, to help public health departments to allocate scarce resources more effectively. To this end, MCS participants are monitored for up to forty minutes following their initial report to an MCS campaign, and based upon the mobility pattern of each participant and other participants, a machine learning-based estimation is proposed under three different unsupervised methods: k-means, Expectation Maximization (EM) and Hidden Markov Model (HMM). In an urban region (i.e., Paris), MCS campaigns for 10,000 and 30,000 users have been tested by monitoring the mobility pattern of each participant for forty minutes. This value has been set according to the maximum observation time recorded across all users, which was 41 minutes. Since users may remain in the same geo-location even after the last observational time, we set the geo-location of any user by its last recorded location whose observation time is less than the maximum observable time. The estimated mobility patterns of all participants reveal clusters of risk groups based on the estimated distance between them. Our simulations indicate that the EM-based estimation of mobility patterns can obtain COVID-19 risk scores of multiple communities 20% ahead of the time (of them leaving the MCS platform) and make highly dependable suggestions for the risk levels within the communities. A recent survey on human mobility models is presented in [16].

The rest of the paper is organized as follows. Section 2 presents an overview of IoT and mobile sensing for community health monitoring solutions. Section 3 presents the proposed scheme for predictive analytics-backed MCS for look ahead community risk maps. Numerical results are presented and discussed in detail in Section 4. The paper is concluded in Section 5 with future directions.

2 IOT AND MOBILE SENSING FOR COMMUNITY HEALTH MONITORING AGAINST OUTBREAKS

Several studies have tackled the role of IoT in community health monitoring against outbreaks. For instance, the study in [4] investigates the viability of using wearable sensors to monitor the populations at risk and patients with mild symptoms of COVID-19. While discussing the integration of Artificial Intelligence with IoT for healthcare services, the study in [6] positions IoT and mobile sensing as the gluing component between AI-based decisions and edge/fog computing facilities/infrastructures, particularly for location-aware solutions to monitor and fight epidemic diseases.

With the recent surge of the COVID-19 pandemic, IoT-based smart decision surveillance systems have been considered as potential enablers to analyze and monitor the outbreaks through the existing surveillance and/or communication infrastructure such as personal mobile devices [12]. The study in [18] presents the viability of interaction between four digital technologies, namely IoT, big data analytics, AI and blockchains to transform the legacy public response strategies against COVID-19 pandemic.

Leveraging the connectivity of IoT devices, Singh et al [15] identified various IoT applications to mitigate the impact of COVID-19 pandemic. The study suggests using IoT devices in record management for hospitalized patients while extending IoT-based self-monitoring services to all other patients. The survey study conducted by Nasajpour et al. [9] presents the role of the IoT technology in response to COVID-19 in three phases: early diagnosis, quarantine time, and post-recovery.

There have been several efforts to model personal contacts such as using Bluetooth Low Energy signals on smartphones [10]. Similarly, social contact between mobile devices is also another dimension investigated in an analytical study considering the infection rates of COVID-19 under a Social Internet of Things (SIoT) network [21]. Apart from the majority of the literature, the study in [11] advocated the use of a dedicated IoT sensor to acquire contact tracing data against COVID-19, where it is presumed that the users of the dedicated IoT sensors have agreed to opt-out from any privacy concern due to contact tracing at the expense of timely outbreak reactions by the governments and/or healthcare organizations. This assumption also aligns with the objective of addressing the trade-off between the value of information and the cost of privacy in IoT [19]. By utilizing the social sensing trend of MCS, Cecilia et al. [2] presents the potential use of various posts related to COVID-19 in social networks in the development of early warning systems. The analyses involves the interpretation of feeds and posts on social media concerning epidemiological control measures.

The study conducted by Swayamsidda and Mohanty [17], introduces the concept of Cognitive Internet of Medical Things to acquire patient data and enable rapid diagnosis. They suggest the utilization of a cognitive radio network to acquire data from electroencephalogram, electrocardiogram, blood pressure, pulse oximeter, and electromyography sensors, and contact tracing data of confirmed COVID-19 cases.

Based on the state of the art and relevant works centered around sensing, sensor networks, and COVID-19, there are no existing studies that leverage MCS campaigns that integrate with various applications and services via smart mobile devices. That said, this study aims to supplement rather than supplant the utility of existing efforts to fight COVID-19 pandemic with IoT-enabled solutions.

3 PREDICTIVE ANALYTICS-BACKED MCS FOR LOOK AHEAD COVID-19 RISK MAPS OF COMMUNITIES

The proposed framework builds on the cloud-inspired sensing as a service concept in MCS to leverage embedded sensors in smart mobile devices.

Crowdsensing campaigns can obtain mobility patterns of MCS participants who have opted to provide their GPS locations to the MCS platform. GPS accuracy of mobile smart devices varies between 6 to 9 meters; hence it is impossible to measure the exact distance between two individuals solely relying on the GPS signals. Although BLE-based contact tracing solutions have been proposed, mobility patterns of MCS campaign participants are needed to forecast their future locations and estimate the distance between

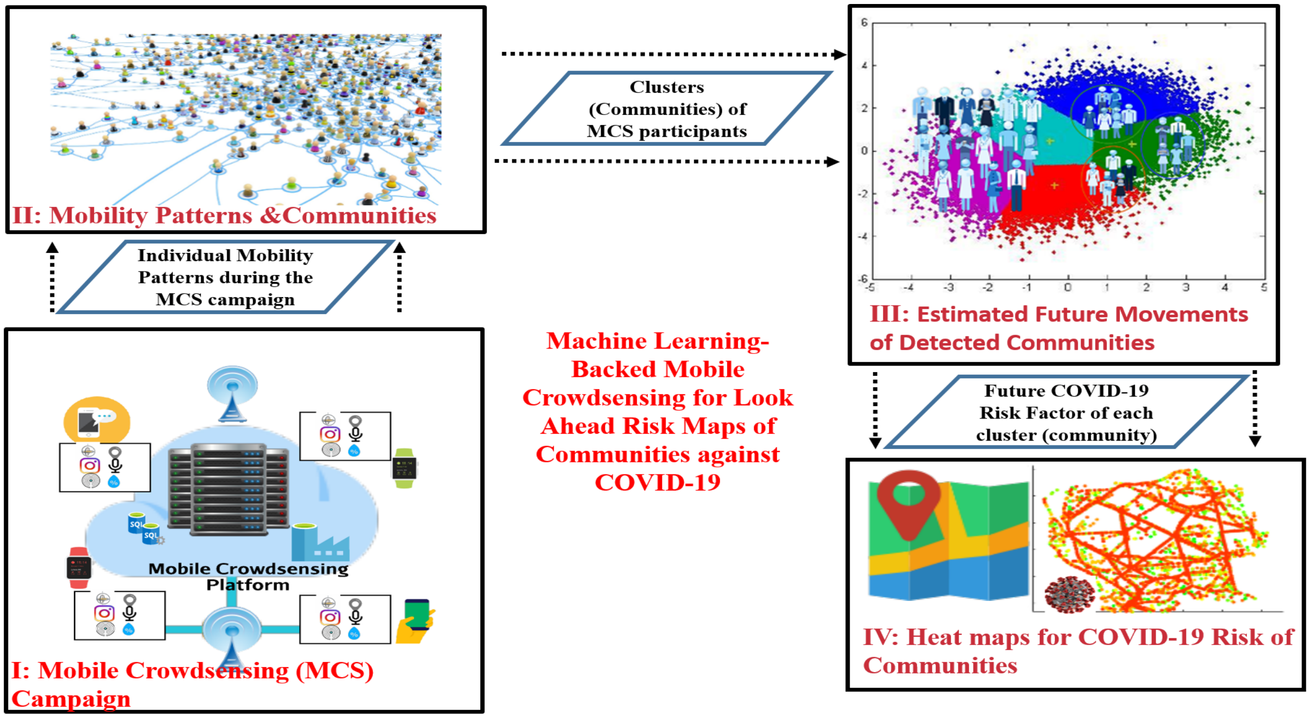
II: Mobility Patterns &Communities
Clusters (Communities) of MCS participants
III: Estimated Future Movements of Detected Communities
6
4
2
0
-2
-4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
Individual Mobility Patterns during the MCS campaign
Machine Learning-Backed Mobile Crowdsensing for Look Ahead Risk Maps of Communities against COVID-19
Future COVID-19 Risk Factor of each cluster (community)
Mobile Crowdsensing Platform
SQL
SQL
I: Mobile Crowdsensing (MCS) Campaign
IV: Heat maps for COVID-19 Risk of Communities

Figure 1: Minimalist illustration of the proposed COVID-19 community risk mapping via MCS

participants to calculate the risk scores. Fig. 1 presents a minimalist illustration of the proposed COVID-19 community risk mapping via MCS. Below, we explain the function of each building block of the proposed solution.

3.1 MCS Campaigns and Data Acquisition

The first building block shown in Fig. 1 denotes a standard data acquisition procedure in an MCS campaign. Thus, the prediction of future COVID-19 risk scores of communities begins with the launch of an MCS campaign. We can see that a typical MCS campaign requests access to multiple sensors in users' mobile devices. The proposed framework is designed to cooperate with MCS platforms to extract GPS data from reported multi-sensor data of MCS participants. Here, what makes it further challenging is that each participant needs to be observed for at least forty minutes to be able to obtain a reasonable social contact network of communities. This time window is used in the formulation of a distinct movement pattern for each participant.

3.2 From Individual Mobility Patterns to Communities

Mobility patterns of MCS participants are defined by time and their geo-coordinates. Since the mobility data is high dimensional, this step involves the application of dimensionality reduction technique, namely t-distributed Stochastic Neighbor Embedding (t-SNE) [20], to the mobility patterns. The dimensionality reduction also enables intuitive visualization of mobility patterns on a 2D-plot. The reduced dimensions are fed into an unsupervised machine learning algorithm to detect clusters within the acquired data where each cluster stands for a community in the monitored region. This study mainly employs and evaluates k-means, EM, and HMM methods to detect communities and predict their future behavior regarding movements, which is explained under the next building block.

3.3 Estimation of Future Behaviour of Communities

This step elaborates on Part III of Fig. 1. The probability of having a community spread by an airborne viral infection is higher when people are in close contact with each other. As known worldwide, the recommended social distance between individuals is two meters to prevent airborne transmission. Spatio-temporal behavior of MCS participants that reveal changes in their geo-coordinates in time reveals clusters (i.e., communities) such that the estimated risk of community spread in each cluster can be computed using the predicted future coordinates of the users.

This study aims to improve the foresight and dependability of the estimations of future community spread risks. Thus, this particular building block's objective is to integrate a look-ahead mechanism to determine the future community risk at $t\%$ ahead of time, which is referred to as the *foresight* in this study. The movement pattern of a user is not random, i.e., a user is likely to walk in the same route for a certain number of steps for a while, following upon their first initial step. This phenomenon is observed in everyday life from the commuting pattern of individuals. Any two instances of the mobility pattern of a user will be relatively close to each other with some distance between them. Based on this fact, a clustering method can be employed on these patterns from their relative distance to each other or through some probability function with a marginal error. The clustering is essentially performed to predict the instances of future movements of each user. It is worth noting that this study focuses on evaluating the efficacy of the clustering algorithms to determine the clusters. Following upon the detection of clusters (i.e., MCS communities), for each participant in a community, a 'contact list' is generated based on the estimated distance to other participants. The generation of the 'contact list' for an instance of a participant/user i in cluster c is defined by the list of participants within a circle of pre-determined radius centered around the participants as he/she moves throughout the MCS campaign. Eqs. 1 - 2 formulate the contact list of an instance (i.e., geo-location) of user i in cluster c.

$$CL_c^i = \{j : j \in c \wedge D_{ij} \leq r \wedge i \neq j\} \tag{1}$$

$$D_{ij} = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_i)\cos(\phi_j)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{2}$$

As seen in the first equation, the contact list is a set of contacts of user i that are within the pre-determined radius r (in meters) around an instance of user i. The distance is obtained from a distance matrix, D where each entry D_{ij} denotes the haversine distance [13] between a geo-coordinate instance of user i and a geo-coordinate instance of user j from its set of instances (i.e., $\{j_{ins} | j_{ins} \in M_{j,c}, j_{ins} \subset M_{j,c}\}$). $M_{j,c}$ contains all the different instances of user j in cluster c. For example, $M_{j,c}$ contains 10 instances of user $j = 5$ (i.e., $j_{ins} = 10$). In Eq. 2, $\Delta\phi$ denotes the latitudinal difference (in radians) between an instance of user i and an instance of user j. Similarly, $\Delta\lambda$ denotes the longitudinal difference in radians. R is the radius of the earth which is multiplied by 1000 to represent the distance in meters.

For every instance (i.e., geo-location) of user i in cluster c, a contact list is generated. Thus, a user's contact list dynamically changes as time elapses. It is worth noting that some mobility patterns of a user may fall into multiple clusters. The contact list of a user for a specific cluster is calculated by taking into account geo-spatial overlaps with multiple instances of other users. Thus, user j may appear in the contact list of user i in multiple instances, and every occurrence of user j is considered as a new increase in the risk score of user i. The rationale for this is that the transmission probability increases when an infected person stays around an individual for more extended period than someone who passes by and walks away. Given the fact that a significant portion of the infected cases remain asymptomatic for COVID-19, instead of investigating whether a contact is infectious, this study considers all contacts potentially risky, and each contact at every instance translates into an additional increase in the risk score of an individual. For a comparative study, we also compute a different number of clusters in all the identified clustering algorithms. Any participant in the monitored region can be susceptible to a viral infection. Measuring the symptoms of every individual in a city is still a challenging task, and it may not even be possible due to scarcity of resources. Isolating each case of an infected person would require an immediate action of contact tracing, which is not the scope of this study. Instead, this study associates a risk score to an individual (MCS participant) based on their contact list, such that a community risk can be derived from a set of participants and communities are ranked with respect to their risk scores.

3.4 Risk Factor generation and Estimation of Future Community Risk

This step presents the method for calculating the risk factor of a cluster/community, facilitating the transition to Part IV in Fig. 1. Note that the term 'community' in this risk study does not represent a geographical area in the monitored region. Rather, it represents various spatio-temporal instances of MCS participants under the MCS campaign, which are clustered based on the estimation of their future movements. Thus, in the monitored region, multiple instances of users contributing to a community are spread across the terrain along with other instances of different communities. The risk score/factor of each user i in every community is obtained by keeping track of their contact list, which is mathematically formulated by Eqs. 3-4 where rf_c^i denotes the risk factor of user i in a cluster c, $I_{i,c}$ stands for the number of instances (geo-location) of user i in cluster c, M is the number of communities /clusters, and U is the number of unique users across the M communities.

$$rf_c^i = (\sum_{i=1}^{I_{i,c}} CL_c^i)/globalmax \tag{3}$$

$$globalmax = \max_{\substack{1<c\leq M;\\ 1<i\leq U}} \sum_{i=1}^{I_{i,c}} CL_c^i \tag{4}$$

Every mobility instance of a user reveals a new footprint of his/her contact network. A contact network footprint of user i contains the list of participants that are present within a pre-determined circle of the user at a given mobility instance of that user. The size of a contact network footprint is represented by the total number of users contained in it. The sum of the sizes of all contact network footprints of user i is normalized by the maximum value of the sums of the contact network footprints of all users in the entire monitored region. This value is referred to as the *globalmax* value defined in Eq. 4. The reason behind taking the maximum summation value among all users is that there is a possibility of clusters' varying in their area. Thus, by setting a maximum global value, it is possible to determine the relative risk scores/factors amongst different clusters/communities. The average risk factor calculated for all users in cluster c denotes the risk factor of the community represented by that cluster as formulated in Eq. 5 where N stands for the number of unique users in cluster c.

$$RF_c = (\sum_{i=1}^{N} rf_c^i)/N \tag{5}$$

4 SIMULATIONS AND NUMERICAL RESULTS

4.1 Simulation Settings

Mobility patterns of MCS participants are recorded using Crowdsensim [5]. Paris, a metropolitan city, is chosen to acquire participants' mobility patterns and detect communities through MCS campaigns to generate and acquire sensory data in a realistic simulating environment. Crowdsensim deploys a crowdsensing campaign that allows participants (individuals who have opted to grant access to the built-in sensors in their mobile devices) to share their sensory data, including their mobility patterns. The mobility patterns recorded by the MCS campaigns help forecast the future community risks under the COVID-19 pandemic. We investigate the impact of the participant pool in MCS campaigns through two scenarios with populations of 10,000 and 30,000. Considering the population of Paris, these correspond to 0.47% and 1.4% participation ratio, respectively. The minimum and maximum travel times of a participant are set to 20 and 120 minutes. The MCS platform can recruit a participant anytime throughout his/her travel. Furthermore, to keep the scenarios realistic, no time enforcement to remain in the campaign has been put in place for the participant, meaning that a participant may choose to leave before the maximum travel time. Wi-Fi antennas of participants' mobile devices are used to connect the participants to the MCS platform.

Each participant's mobility pattern is defined by the tuple <UserID, Latitude, Longitude, Day, Hour, Min, Sec>. A heatmap correlation matrix is used to test any correlation in the features. After analyzing the correlation matrix, the final tuple is reduced to <UserID, Latitude, Longitude, Hour, Min, Sec>. The tuple uniquely defines the behavior of each user. It is worth to note that although the mobility behavior is tracked in minute granularity, since participants in an MCS campaign push sensed data to the MCS platform in the order of seconds, mobility data are extracted out of the reported sensory data of the participants, and are represented in the order of minutes.

The difference between two consecutive instances/mobility data of a user is one minute. Thus, in every minute, the latitude-longitude values of a user are recorded. While the final tuple contains the *Sec* attribute to denote time granularity at the order of seconds, the change in the geo-coordinate of a user is observed at the order of minutes which is recorded by the *Min* attribute. Hence, the *Sec* attribute is discarded in determining the geo-location of the MCS participants in this study. Each participant is observed for up to forty minutes following the first recorded instance for that user. This observation leads to sufficient length in the travel distance for each user. In this case, for 10,000 users, the mobility pattern data set would contain 410K instances, while for 30,000 users, the mobility pattern data set would contain 1.23M instances.

4.2 Evaluation Metrics

Two evaluation metrics have been considered to estimate the future community risk of COVID-19 as defined below:

Dependability: To evaluate the estimation of future community risk, we introduce an error metric to denote the average error between the actual risk factor and the predicted risk factor for M communities against COVID-19 over the monitored region. The sum of the differences between the actual and predicted risk factors in all communities is averaged by the total number of communities obtained by the clustering algorithm. One's complement of the risk prediction error is defined as the dependability of the prediction.

Foresight: The predicted COVID-19 risk factor of a community is computed $t\%$ ahead of time. A prediction being $t\%$ ahead of time indicates that the participants have been contracted by the MCS platform for (1-t)% of their time commitments for the MCS campaigns. In this work, we also aim to recommend the most feasible value for the foresight (i.e., *t% ahead*) to make this future prediction by testing the error between the actual and the predicted risk factor. The total instances of the users are partitioned so to predict the future mobility behavior of the users. In our experiments, 20% to 50% ahead predictions are tested to determine the most feasible value for the foresight.

The average dependability metric can be mathematically formulated in Eq. 6 where RF_c denotes the actual risk factor while $RF_c^{'}$ denotes the predicted risk factor with a foresight at $t\%$ ahead for M communities.

$$Dependability(Avg) = (\sum_{c=1}^{M} |RF_c - RF_c^{'}|)/M \tag{6}$$

Under the M communities of a particular clustering algorithm, it is possible to identify the community having the least error between its actual risk factor and predicted risk factor by observing their absolute differences. Thus, a community having the minimum difference in the actual-predicted risk factor will result in the maximum dependability of the prediction as formulated in Eq. 7.

$$Dependability(Max) = \min_{c \in M} |RF_c - RF_c^{'}| \tag{7}$$

4.3 Dimensionality Reduction and Clustering

T-distribution-based Stochastic Neighbor Embedding (t-SNE) is applied in all instances to map the feature set onto two dimensions before selecting clusters. The output of t-SNE is a 2D vector representing the mobility patterns. The outputs are the data points that will be clustered based on an estimation of future movements. In each of the three clustering methods, namely Expectation Maximization (EM), Hidden Markov Model (HMM), and k-means, the number of communities M is set to 5, 6, 7, and 8, and for each clustering algorithm, the value of M is determined based on the maximum dependability of risk prediction. From the sets of participant instances (concerning the aimed foresight), t-SNE computes M clusters over the 2D vector. These clusters are referred to as communities. The communities are also determined over the users' entire committed times for the MCS campaigns.

4.4 Numerical Results

According to the Official U.S. government information about the GPS and related topics, GPS accuracy of the new generation of smartphones is 4.9m on average. However, this value can vary depending on surrounding physical settings. Moreover, the GPS accuracy of the previous generations of smartphones vary between 6-13 meters. We set the pre-determined radius (r) for each participant's risk circle to 10 meters, taking into consideration the factors discussed above.

Table 1 presents the average dependability performance of the three methods for COVID-19 risk estimation under Scenario 1 and Scenario 2. These results represent the best combination for foresight and number of communities for each clustering algorithm. For simplicity purposes, we present the performance of these algorithms under the best input parameter selections. In Scenario 1, EM results in the maximum average dependability of COVID-19

	Scenario 1:	**Participants: 10,000**	**Participation: 0.47%**
Clustering Method	Number of communities	Dependability (avg)	Foresight
EM	**8**	**0.9316**	**20% ahead**
HMM	8	0.9301	20% ahead
k-means	8	0.9025	40% ahead
	Scenario 2:	**Participants: 30,000**	**Participation: 1.4%**
EM	**8**	**0.9823**	**20% ahead**
HMM	5	0.9820	20% ahead
k-means	8	0.9485	20% ahead

Table 1: Average dependability of COVID-19 risk estimation for 10,000 and 30,000 MCS participants

Community	50% ahead	40% ahead	30% ahead	20% ahead
Community 8	0.08	0.093	0.076	0.04
Community 2	0.139	0.112	0.196	0.066
Community 3	0.14	0.161	0.057	0.044
Community 1	0.089	0.125	0.071	0.055
Community 5	0.111	0.133	0.062	0.11
Community 4	0.128	0.132	0.051	0.101
Community 7	0.085	0.113	0.06	0.059
Community 6	0.124	0.15	0.089	0.071

Table 2: COVID-19 risk factor estimation error with respect to the actual risk factor of each community under EM Clustering at various foresight for 10,000 participants. Communities are presented and color-coded starting from the one with the highest risk (red) to the one with the lowest risk (green) score. Actual and predicted risks vary between zero and one.

Community	50% ahead	40% ahead	30% ahead	20% ahead
Community 8	0.021	0.068	0.021	0.017
Community 2	0.024	0.088	0.0005	0.012
Community 4	0.038	0.12	0.026	0.071
Community 7	0.136	0.099	0.035	0.001
Community 1	0.049	0.124	0.13	0.009
Community 6	0.033	0.103	0.046	0.002
Community 3	0.048	0.135	0.038	0.012
Community 5	0.044	0.09	0.035	0.017

Table 3: COVID-19 risk factor estimation error with respect to the actual risk factor of each community under EM Clustering at various foresight for 30,000 participants. Communities are presented and color-coded starting from the one with the highest risk (red) to the one with the lowest risk (green) score. Actual and predicted risks vary between zero and one.

risk estimation, i.e., 0.9316, while the number of communities is 8, and foresight is at 20% ahead of time. In Scenario 2, EM and HMM outperform k-means in terms of dependability of their risk predictions while EM performs slightly better than HMM with an average dependability of risk estimation at 0.9823 with eight communities and a 20% ahead foresight.

Below we provide a comprehensive discussion and analysis of the performance results.

Scenario 1 - 10,000 users: Table 2 presents the COVID-19 risk estimation error at various foresight with respect to the actual risks of eight communities formed under EM. The definition of a community in the context of this risk study is defined as the different instances of MCS participants who have been clustered intelligently based on an estimation of their future movements. Each row in the table highlights the risk estimation error of a particular community at different foresight (i.e., 20% to 50% ahead of time), and the communities are sorted (in decreasing order) with respect to their actual risk scores at the end of the MCS campaigns, and based on their risks at the end of the MCS campaigns, each row is also color coded accordingly varying from red to green.

Each COVID-19 risk factor estimation error represents the absolute difference between the predicted risk (i.e., t% ahead of the end of the participants' MCS commitments) and the actual risk (i.e., at the end of the participants' MCS commitments) of a community. The lower the error, the higher the dependability of risk predictions. The results in the table particularly help to investigate the most feasible foresight for the community risk factors by answering the following question: How much time ahead of the last MCS commitments can the community risk scores be estimated with the lowest possible error? The average of the absolute difference across all communities under each foresight is used to find the foresight that would lead to the lowest error. It is shown in the table that the risk factor estimation error is minimum at a foresight that is 20% ahead of the participants' last commitments for the MCS campaign.

Under 20% foresight, the minimum error of risk factor estimation (0.04) is obtained for Community 8. Thus, the maximum dependability for the risk estimation is obtained for Community 8, and its value is 0.96. The average of the risk estimation errors across all communities is 0.07, which translates into an average dependability of 0.93, as observed under Scenario 1 of Table 1. These low error values (and their corresponding high dependability values) enable estimation of the nature of risks associated with an outbreak within communities. It is worth to note that risk estimations are made solely based on the analysis of the participants' movements during MCS campaigns with fewer instances of individuals than the actual populations. Based on these observations, it can be concluded that the proposed framework can help public health services in improving decisions against a pandemic such as COVID-19 as they can be informed about the projection of future risks within communities through an IoT-enabled solution.

Scenario 2 - 30,000 users: Table 3 presents the risk estimation error of EM with 30,000 participants at various foresight with respect to the actual community risks. The table follows the same structure as Table 2, i.e., each row highlights the risk estimation error of the predicted risks to the actual risks at different foresight. Similar to Scenario 1, in Scenario 2 (i.e., Table 3), as well, 20% ahead foresight results in the lowest error in risk factor estimation. As seen in the table under 20% ahead foresight prior to the completion of participants' MCS contributions, the proposed framework achieves the minimum risk estimation error (0.001) for Community 7, and the error translates into 0.99 dependability. The average of the risk estimation errors across all communities is 0.0176, which translates into an average dependability of 0.98.

When the average dependability under Scenario 1 is compared to that under Scenario 2, it can be concluded that the higher the participation, the better the estimations. This work suggests that ML-Backed MCS platforms with sufficient participation can efficiently estimate the future risk scores of communities against COVID-19 pandemic.

Look Ahead Risk Maps of Community Risk: Figs 2-3 illustrate the risk heatmap between the actual community risk and predicted risk for Scenario 1 (10,000 MCS participants) and Scenario 2 (30,000 MCS participants) under EM clustering to form the eight communities. In both plots, the legend displays the risk factor of 8 clusters, each associated with a color code varying from green (lowest risk) to red (highest risk) through varying a hue factor value. Each geo-coordinate of a participant is associated with a cluster. Each participant is plotted on a representative (averaged) coordinate color coded by the actual (left) or predicted (right) COVID-19 risk score of his/her community. COVID-19 hot spots of risk areas can be determined by zooming into the risk maps as illustrated in the figures. When the enlarged plots of actual and predicted community risks are compared, it is observed that predictions with 20% ahead foresight align with the actual COVID-19 risks of the communities. Observing the contours of both the enlarged plots, it is noticeable that the risk factor decreases. This means fewer people co-exist in those locations.

Another observation is that, as the number of participants increases, the risk heatmap of predictions approaches the actual risk heatmap of communities in terms of the hot spots although there is not a dramatic change in the actual average community risk scores as the average dependability is above 90% in both scenarios from reference to Table 1.

5 CONCLUSIONS

Mobile crowdsensing has appeared as a viable non-dedicated sensing paradigm for the Internet of Things. With the surge of COVID-19 pandemic, crowd monitoring and decision-making systems can significantly benefit from MCS campaigns to understand the community behavior and predict the upcoming risks against the epidemic. In this paper, we presented a four-block COVID-19 community risk mapping framework on a machine learning-backed MCS platform. The proposed framework consists of mobile crowdsensing campaigns, mobility patterns, communities extracted out of the mobility patterns, estimated future movements of the detected communities, and the projected heatmaps for COVID-19 risks of the communities. A feasibility study with multiple unsupervised machine learning algorithms has been conducted, and it is shown that the Expectation Maximization (EM) algorithm integrated into the proposed framework can predict the community risk scores 20% ahead of time with average dependability of 93% and 98% with 10,000 and 30,000 MCS participants, respectively.

As this paper presents a proof of concept and a feasibility study, future extensions and challenges have also been identified. MCS can be implemented in either an opportunistic or a participatory manner. The latter requires overt user involvement in accepting the monitoring, reporting sensor and other information. Thus, participatory MCS campaigns can provide information about future locations of participants based on the tasks they have opted in. However, this still needs to be complemented by an accurate trajectory estimation. Therefore our ongoing agenda to extend this work includes effective statistical and ML-based solutions to improve the community detection and risk mapping for COVID-19. Last but not least, this work has assumed that all participants were incentivized to join MCS campaigns; however incorporating effective incentives schemes from the literature such as [22, 23] into this work is also included in our agenda.

ACKNOWLEDGMENTS

This work was supported in part by the the Natural Sciences and Engineering Research Council of Canada (NSERC) under the Alliance Program under Grant Number ALLRP 552696-2020, and Compute Canada (https://www.computecanada.ca). The authors would also like to thank all healthcare professionals and essential workers who have been at the front line of the battle against COVID-19 pandemic.

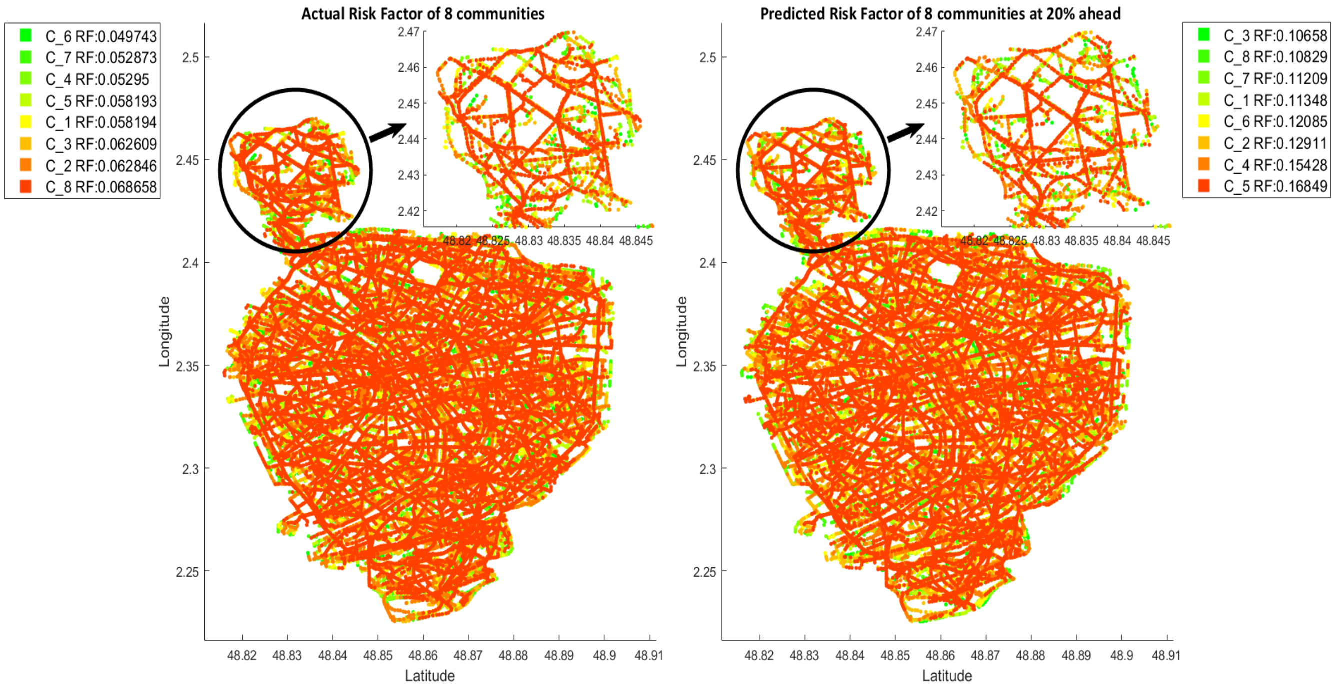

Figure 2: Look ahead risk map of communities against COVID-19 for Scenario 1: 10,000 users under EM clustering.

REFERENCES

[1] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry. 2019. A Survey on Mobile Crowdsensing Systems: Challenges, Solutions, and Opportunities. *IEEE Communications Surveys Tutorials* 21, 3 (thirdquarter 2019), 2419–2465.

[2] J. M. Cecilia, J.-C. Cano, E. Hernández-Orallo, C. T. Calafate, and P. Manzoni. 2020. Mobile crowdsensing approaches to address the COVID-19 pandemic in Spain. *IET Smart Cities* 2, 2 (2020), 58–63.

[3] V. Chamola, V. Hassija, V. Gupta, and M. Guizani. 2020. A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. *IEEE Access* 8 (2020), 90225–90265.

[4] X. Ding, D. Clifton, N. JI, N. H. Lovell, P. Bonato, W. Chen, X. Yu, Z. Xue, T. Xiang, X. Long, K. Xu, X. Jiang, Q. Wang, B. Yin, G. Feng, and Y. Zhang. 2020. Wearable Sensing and Telehealth Technology with Potential Applications in the Coronavirus Pandemic. *IEEE Reviews in Biomedical Engineering* (2020), 1–1.

[5] C. Fiandrino, A. Capponi, G. Cacciatore, D. Kliazovich, U. Sorger, P. Bouvry, B. Kantarci, F. Granelli, and S. Giordano. 2017. CrowdSenSim: a Simulation Platform for Mobile Crowdsensing in Realistic Urban Environments. *IEEE Access* 5 (2017), 3490–3503.

[6] L. Greco, G. Percannella, P. Ritrovato, F. Tortorella, and M. Vento. 2020. Trends in IoT based solutions for health care: Moving AI to the edge. *Pattern Recognition Letters* 135 (2020), 346 – 353.

[7] L. A. Kalogiros, K. Lagouvardos, S. Nikoletseas, N. Papadopoulos, and P. Tzamalis. 2018. Allergymap: A Hybrid mHealth Mobile Crowdsensing System for Allergic Diseases Epidemiology : a multidisciplinary case study. In *IEEE PerCom Workshops*. 597–602.

[8] H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang, and C.-K. Tang. 2020. Coronavirus disease 2019 (COVID-19): current status and future perspective. *International Journal of Antimicrobial Agents* (2020), 105951.

[9] M. Nasajpour, S. Pouriyeh, R. M. Parizi, M. Dorodchi, M. Valero, and H. R. Arabnia. 2020. Internet of Things for Current COVID-19 and Future Pandemics: An Exploratory Study. *ArXiv* abs/2007.11147 (2020).

[10] P. C. Ng, P. Spachos, and K. Plataniotis. arXiv, 2020. COVID-19 and Your Smartphone: BLE-based Smart Contact Tracing. arXiv:2005.13754 [cs.LG]

[11] A. Polenta, P. Rignanese, P. Sernani, N. Falcionelli, D.N. Mekuria, S. Tomassini, and A.F. Dragoni. 2020. An Internet of Things Approach to Contact Tracing—The BubbleBox System. *Information* 11 (2020), 347.1–347.12. Issue 7.

[12] Md S. Rahman, N. C. Peeri, N. Shrestha, R. Zaki, U. Haque, and S. H. Ab Hamid. 2020. Defending against the Novel Coronavirus (COVID-19) Outbreak: How Can the Internet of Things (IoT) help to save the World? *Health Policy and Technology* 9 (June 2020), 136–138. Issue 2.

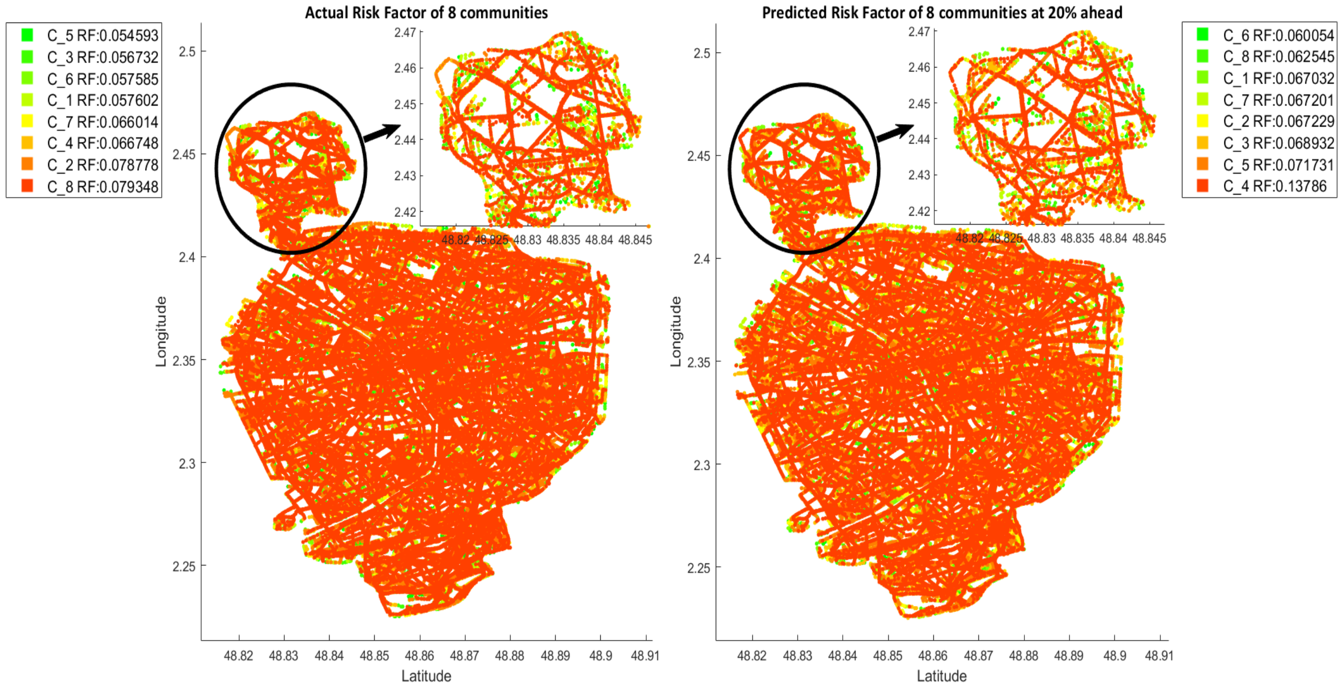

Figure 3: Look ahead risk map of communities against COVID-19 for Scenario 2: 30,000 users under EM clustering.

[13] C. C. Robusto. 1957. The cosine-haversine formula. *The American Mathematical Monthly* 64, 1 (1957), 38–40.

[14] M. Simsek and B. Kantarci. 2020. Artificial Intelligence-Empowered Mobilization of Assessments in COVID-19-like Pandemics: A Case Study for Early Flattening of the Curve. *Int. Journal of Environmental Research and Public Health* 17 (2020), 3437.1–3437.17. Issue 10.

[15] R. P. Singh, M. Javaid, A. Haleem, and R. Suman. 2020. Internet of things (IoT) applications to fight against COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (July-August 2020), 521–524. Issue 4.

[16] G. Solmaz and D. Turgut. 2019. A survey of human mobility models. *IEEE Access* 7, 1 (December 2019), 125711–125731. DOI: 10.1109/ACCESS.2019.2939203.

[17] S. Swayamsiddha and M. Chandana. 2020. Application of cognitive Internet of Medical Things for COVID-19 pandemic. *Diabetes Metabolic Syndrome: Clinical Research Reviews* 14, 5 (2020), 911 – 915.

[18] D. S. W. Ting, L. Carin, V. Dzau, and T. Y. Wong. 2020. Digital technology and COVID-19. *Nature medicine* 26, 4 (2020), 459–461.

[19] D. Turgut and L. Bölöni. 2017. Value of Information and Cost of Privacy in the Internet of Things. *IEEE Communications Magazine* 55, 9 (2017), 62–66.

[20] L. van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[21] B. Wang, Y. Sun, T. Q. Duong, L. D. Nguyen, and L. Hanzo. 2020. Risk-Aware Identification of Highly Suspected COVID-19 Cases in Social IoT: A Joint Graph Theory and Reinforcement Learning Approach. *IEEE Access* 8 (2020), 115655–115661.

[22] D. Yang, G. Xue, X. Fang, and J. Tang. 2016. Incentive Mechanisms for Crowdsensing: Crowdsourcing With Smartphones. *IEEE/ACM Transactions on Networking* 24, 3 (2016), 1732–1744.

[23] X. Zhang, Z. Yang, W. Sun, Y. Liu, S. Tang, K. Xing, and X. Mao. 2016. Incentives for Mobile Crowd Sensing: A Survey. *IEEE Communications Surveys Tutorials* 18, 1 (2016), 54–67.

[24] Y. Zhang, Y. Li, B. Yang, X. Zheng, and M. Chen. 2020. Risk Assessment Of COVID-19 Based On Multisource Data From A Geographical View. *IEEE Access* (2020), 1–1.

[25] Y. Zhang, M. Simsek, and B. Kantarci. 2020. Empowering Self-Organized Feature Maps for AI-Enabled Modelling of Fake Task Submissions to Mobile Crowdsensing Platforms. *IEEE Internet of Things Journal* (2020), 1–1.