Bireylerin Kovid-19 Riskinin Uzay-zamansal Olarak Belirlenmesi

Araştırma Makalesi/Research Article

Hayri Volkan AGUN

Department of Computer Engineering, Bursa Technical University, Bursa, Turkey
hayri.agun@btu.edu.tr
(Geliş/Received:23.06.2022; Kabul/Accepted:07.01.2023)
DOI: 10.17671/gazibtd.1135014

Özet— Mevcut çalışmalar örneğin şüpheli-bulaş-eksiltme modeli ve makine öğrenmesi modelleri her bir kişi ve alan için bulaş riskinin hesaplanmasına uygun değildir. Bu çalışmada mevcut yaklaşımların eksik yönlerinin giderilmesi için toplanan verilerin uzaysal ve zamansal tahminleme modeli olarak bir araya getirildiği bir dönüt işleme tasarımı önerilmektedir. Önerilen tasarım üç ana işleme aşaması içermektedir. Bunlar verinini üretilmesi, geri dönüş analizi ve gerçek zamanlı uzaysal ve zamansal değerlendirme süreçleridir. Verilerin üretilmesi aşamasında her bir bireyin Kovid-19 durumunun Markov olasılık işlemi kullanılarak üretildiği süreç yer alır. Bu aşamada hastalığın çoğalma parametreleri, semptonlu hastaların ve semptonsuz hastaların görülme sıklığı, toplam nüfus, hastalığı geçirmekte olan nüfus, ve hareket halinde olan nüfus sayıları kullanılarak her bir hasta için Kovid durumu ve hareket halinde olma durumu rastsal olarak güncellenir. Hareket verisi ise rastsal olarak belirlenen özel alanlar için oluşturulur. Bu veride kişilerin belirli bir alan içerisindeki etkileşimleri rastsal olarak hesaplanır. Geri dünüş analizi aşamasında toplanan istatistikler ve yerel olay verileri birleştirilerek doğrusal bir model yardımıyla her bir bireyin Kovid-19 riski tahmin edilir. Bu bağlamda yerel istatistilerin elde edilmesinde olasılıksal bir yakınsama yaklaşımı kullanılabilir. Değerlendirme aşamasında, geri dönüş analizinden elde edilen tüm etkileşimler kişilerin periodik olarak güncel Kovid-19 riskinin hesaplanmasında kullanılır. Daha sonra her bir kişinin üretilen verideki Kovid-19 bilgisi kullanılarak tamin başarısı o zaman aralağı için hesaplanır. Populasyon sayısı, yer/zaman ve hareketlilik oranınında bağımsız olarak her bir birey etkileşimi için hesaplanan Kappa önerilen tasarımın etkisinin önemli olduğunu göstermiştir.

Anahtar Kelimeler— pandemi yayılma tahmini, uzay-zamansal analiz, akış işleme, risk hesaplama

A Spatio-Temporal Approach For Determining Individual's Covid-19 Risks

Abstract— Current state of art approaches such as the susceptible-infected-removed model and machine learning models are not optimized for modeling the risks of individuals and modeling the effects of local restrictions. To improve the drawback of these approaches, the feedback processing framework is proposed where previously accumulated global statistics and the model estimates generated from the spatial-temporal data are combined to improve the performance of the local prediction. The proposed framework is evaluated in three processing stages: generation of the simulation dataset, feedback analysis, and evaluation for the spatial-temporal and real-time pandemic analysis. In the data generation stage, the corresponding state of the illness for each person is modeled by a Markov stochastic process. In this stage, the parameters such as the reproduction rate, symptomatic rate, asymptomatic rate, population count, infected count, and the average mobility rate are used to update the individual's Covid-19 status and the individual's movements. The movement data of each person is generated randomly for several places of interest. In the feedback analysis stage, both the aggregated statistics and the local event data are combined in a linear model to infer a score for the Covid-19 probability of the person. In this respect, a stochastic model can be used to approximate the local statistics. In the evaluation stage, the result of the feedback analysis for all the interactions is used to classify the state of the individuals periodically. Later the accuracy of the evaluation for each person is obtained by comparing the individual's prediction with the real data generated in the same time interval. The Kappa scores independent from different populations, locations, and mobility rates obtained for every interaction indicate a significant difference from the random statistics.

Keywords— pandemic spread prediction, spatial-temporal analysis, stream processing, risk computation

1. INTRODUCTION

Pandemic analysis models have been frequently used to assess the risks of the Covid-19 spread. Common methods use the parametric models to estimate the number of patients for each stage of the illness [1-3]. Prediction of the state of the pandemic is done through aggregating the statistics and using those statistics in models. These models are used to track the state of the Covid-19 pandemic for a short time [4]. In general, aggregated statistics work very well for predicting the cases even if the data is noisy. However, these parameters may become inaccurate when the dynamics of seasonal, and locational changes of the pandemic are considered. The common mathematical model of the spread analysis is known as the susceptible infected recovered (SIR) model where the number of cases is predicted based on the infection rates. This spread model has been modified to overcome different challenges in Covid-19 prediction [5].

The mathematical models use the rates of the change of the variables of the pandemic [6]. These variables include the number of infected people, the reproduction number, the number of symptomatic patients, and the number of deads. Each of these variables is linked to each other by the constants obtained from the real cases. The linking of the variables creates a dynamic mathematical model known as susceptible, exposed, and asymptomatic (SIR) model [6]. During the Covid-19 several variations of the SIR model have been proposed. For example, in the SEAIR model the variables S, E, A, I, and R are used to denote the fraction of individuals which are respectively susceptible, exposed, asymptomatic, infectious, and recovered [7]. Similar extensions of the SIR model have been successfully applied on the Covid-19 data [8].

The previously mentioned SIR models and their derivations use only the parameters of global spread rates of the pandemic. On the other hand, the mobility pattern tracking approaches use both global parameters and the parameters of the spatial rates obtained through spatially linked cases and spatial clustering methods. These approaches have been successfully applied in the analysis of HIV transmission in Kenya [9], and Covid-19 in Oman [10], and in United States of America [11, 12]. All of these approaches have the capability of modeling different areas through mobility patterns and connections between these locations. Since a lockdown in the local area, creates a difference in the connections and the mobility patterns, thus use of the location information may improve the prediction score of the SIR approach.

To overcome the limitations of the SIR model, machine learning (ML) methods have been applied to pandemic datasets. During the Covid-19 pandemic, ML methods become one of the most studied prediction approaches. ARIMA time series forecasting [13], linear regression models such as Support Vector Regression [13], Gaussian mixture models [2], and random forest classification [14, 15] has been applied in prediction. Similarly, recurrent deep learning models such as GRU and LSTM [12, 16] have been successfully model the parameters of the Covid-19 pandemic. The main advantage of deep learning approaches is the ability to model long-range interactions and the ability to use a wide range of feature sets including the parameters of the SIR model. These parameters are the number of interactions, temporal patterns, census features, and reproduction numbers [17] In [17], a neural network model has been demonstrated as a successful combination of time-series data, cross-country specific features, and local features such as the number of hospitals, healthcare workers, and percentages in a neural representation.

Machine learning methods can model recent trends in time series data of the Covid-19 pandemic. In this respect, deep learning methods such as LSTM have been successful. A deep learning study proposed the mean percentage error measure in the prediction of the number of patients according to each social determinant of health (SDH) such as age group, education, etc [18]. In this approach, a convolution neural network classifier is trained for each region by using the SDH parameters. The prediction accuracy of Covid-19 is not the single contribution of the deep learning models. In a deep LSTM approach [17], the representative vector is modeled from the interaction of the features used in the prediction of the Covid-19 cases. The findings of this study suggest that the census features such as age/sex, race, ethnicity, household/family type, school enrollment, poverty status, income, etc. are correlated with all the other features such as mobility, transportation rate, mortality rate, and the Covid-19 reproduction numbers. Embedding the interactions of such features in the pandemic analysis is an effective generalization ability that can be used in other pandemic cases [19].

Spatial-temporal dynamics in spread models are very important for tracking the virus spread. Especially hot-spot analysis approaches are used to track the pandemic cases. In [20], the number of Covid-19 cases is computed in a distance-based correlation index for identifying the hotspots. In [21], the reproduction numbers are used in geospatial clustering to model the interactions of the Covid-19 cases based on the locations. In [22], a center of gravity model is proposed to localize the hubs and flow effect in pandemic parameters. Similarly, kernel-based spatial-temporal clustering methods have been applied for analysis of epidemiological diseases such as childhood leukemia and asthma [23]. Along with kernel density estimates, a Poisson- and Bernoulli-based prospective space-time scan is proposed to find the dense and highly probable spread clusters [24]. Moreover, in [10, 20, 25], a SIR model and the mobility network analysis is combined to estimate the recent reproduction numbers.

2. RISK ANALYSIS

In the Covid-19 pandemic, several mathematical models are proposed to predict the rate of the spread. The susceptible exposed-infectious resistant-susceptible (SEIRs) and susceptible exposed infectious recovery models (SEIR) are used for estimating the infection or

reinfection rates in a population [7]. These models are the extended versions of the susceptible-infected-removed (SIR) model which is used for modeling the parameters of the virus spread [25]. The SIR model with a time delay function for reinfection is demonstrated by Equation (1) [26]. In Equation (1), S(t), I(t), R(t), and C(t) are the rate of susceptible, infected, recovered, and cross-immune people respectively at a given time t, and the population size N(t).

$$\begin{aligned} N(t) &= S(t) + I(t) + R(t) + C(t) \\ S(t) &= \sigma\big(1 - S(t)\big) - \epsilon S(t) I(t-\tau) + \beta C(t) \\ I(t) &= \xi S(t) I(t-\tau) + \sigma\xi C(t) I(t) - (\eta + \alpha) I(t) \\ R(t) &= (1-\sigma)\xi C(t) I(t) + \alpha I(t) - (\eta + \gamma) R(t) \\ C(t) &= \gamma R(t) - \xi C(t) I(t) - (\eta + \beta) C(t) \end{aligned} \quad (1)$$

The reproduction parameters of the pandemic described in the SIR model are found by disease-free equilibrium (DFE) [8]. DFE proposes that the change in the number of infected people is dependent on other parameters. The prediction searches the time or the iteration when the changes in these parameters get fixed and the system gets to an equilibrium state. The SIR model uses the statistics gathered from the pandemic to determine these parameters. In Equation (1), these parameters are denoted by the symbols of eta, beta, alpha, tau and sigma.

In Covid-19, the SIR model is applied in certain intervals especially during the initial outbreak. In Figure 1, the periods are shown for describing the application of training and testing phases. The model is constructed from aggregated statistics gathered in the stages of Training 1 and Training 2. Later this model is verified through the validation time-frame and applied to predict the state of the pandemic in evaluation. The features such as the average number of mobility, the number of active cases, the number of symptomatic cases are accumulated through the training period and used to obtain the model parameters so that the number of active patients, symptomatic patients, and serious cases for the evaluation period can be predicted. The regulations and local restrictions enforce the time range of the prediction to be kept minimum and reduce the prediction performance of the model. The stochastic SIR model, the Markovian model, and the machine learning model have been used in the prediction of the aggregated statistics of the Covid-19 pandemic.

3. DATA GENERATION

In general, the pandemic datasets contain the statistics such as the percentages of active patients, serious cases, and the deads on a weekly basis for each location [27]. The parameters such as the time and the number of visits of each census block group (CBG) to a place of interest (POI) are not given [12] in those datasets. There are also mobility tracking datasets such as SafeGraph [29] where the tracking information for the individuals including the visited locations are available. However, none of these datasets contain GPS tracking information of the individuals thus they don't pinpoint the location of the violations of the social distance rules. Similarly, the datasets collected by the companies such as Google and Facebook do not include the tracking information. It can be claimed that the performance limitations of the previous analytic models are occurred not only because of the complexity of the Covid-19 pandemic but also the details of the gathered information. In this study, in order to overcome these limitations, a simulation approach based on a Markov chain model is proposed. This model approximates the true statistics by tracking the states of each individual and also by approximating the actual statistics of the pandemic.

A Markov chain is a random process model constructed from the finite (discrete) state Markov stochastic process. Markov property is the fundamental assumption of the Markov chain where the probability of an event in a given state is conditionally dependent on the previous state. In a Markov chain, the conditional property is used to determine the transitions from one state to another. In order to generate a dataset, we used weekly aggregated data of Covid-19 pandemic such as the number of tested patients, the number of infected and the number of serious cases. These numbers are used to create the transition probability matrix of the Markov chain. An example Markov chain transition diagram for 1st week of April 2020 is given in Figure 1.

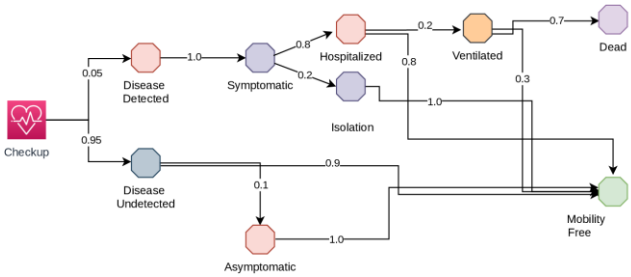

Figure 1. Markov transition diagram

In Figure 1, the number of checkups, the number of patients, and the number of healthy individuals are used to compute the transition probabilities for the disease-detected (positive) and disease-undetected (negative) cases. For example, if 100 people have been checked-up for the virus, and 80 people have not been infected then the transition probability to the state of disease undetected becomes 0.8. In order to generate a dataset, the transition matrix is used to approximate true posterior probability. In this case, there are three parameters are used to create a dataset, the number of weeks starting from a given date, the population size, total number of healthy people, total number of infected people, the percentage of the mobility in population, the probability of getting the disease from an infected person and the number of different permitted locations. These numbers and also the Markov chain matrix is used to decide whether a person can travel through the permitted location in a randomly determined time range. The person gets infected or not based on the transition matrix and the interactions of the person. In this respect, if the person is close enough to another person then

his/her getting an infection becomes more likely. Similar to the SIR model, a person might get infected again and shows either signs of illness (symptomatic) or not (asymptomatic). If the person is symptomatic then he/she is in isolation and by no chance, he/she can infect others. However, if the person is asymptomatic then he/she might infect others. The Markov network shown in Figure 1, is represented by a transition matrix given in the Equation (2). The element-wise multiplication of the transition matrix gives the estimate of the current transition probabilities at time h for a given person.

$$P^h = \begin{pmatrix} p_{11}^h & \cdots & p_{1n}^h \\ \vdots & \ddots & \vdots \\ p_{n1}^h & \cdots & p_{nn}^h \end{pmatrix} \qquad (2)$$

$$P^h = \underbrace{P \times P \times P \ldots P}_{h-times}$$

The probability of the transition from state i to state j is given in the ith row and jth column of the matrix. If the ith state is not connected to the jth state then the probability value is 0.0. The decision for each person is computed according to Equation (3).

$$\begin{aligned} V &= Normal(\mu, \alpha) \\ S &= [1.0,0.0,0.0,0.0] \\ S\_next &= S \times P^h \\ S_{next} &= [0.05, 0.8,0.2,0.3] \\ S_{next} &= Random(S_{next}) \\ S_{next} &= [0.0, 1.0, 0.0,0.0] \end{aligned} \qquad (3)$$

In Equation (3), the state probabilities are given in the S vector for a person. If there is not any state vector (S) for a person then the first state vector is determined by a normal distribution where the mean is the average number of check-ups and the standard deviation is the deviation from the mean for the current month. In this case, whether a person get a check-up or not is determined by randomly tossing a bias coin where heads are sampled from a normal distribution of checkups in the current month. After the S vector is constructed than the next S vector is computed via multiplication of the transition matrix. The multiplication creates another state vector. The state vector is used to determine the next vector by random selection. In this case, if the probability of the state is higher than any other state probability then it is more probable to be selected as the next state.

The states of each person are updated daily. So if a person got infected during the day, then he/she can infect others as well during the day. If he/she is tested positive then he/she cannot infect others. Each person is selected for a check-up by pure chance driven from the normal distribution. A step by step example is given in the Table 1. The theory for the above calculations is given in [28]. The algorithm of the data generation is given in Figure 2.

Table 1. Example calculations

Name	Operation	Example
V	A vector of values where checkups are stored with 1's and no checkup with 0's.	A population size of 4 with a mean of 3 and standard deviation of 0.25 V=[1, 1, 1, 0]
S	A state vector where a non zero value positioned at the index represent the probability of being at that state for the person	Person is in 0th state where it is represented by a checkup. S=[1.0, 0.0,0.0]
S_{next}	State vector is multiplied with the state probability matrix so that the next state is calculated.	State vector now indicate greater than 0 values for the possible states S = [0.05, 0.8, ..,0.3]
S_{next}	The next state index is determined by randomly selecting the state index by using the probabilities in the state vector	The state vector is now becomes a one hot vector such as [0.0, 1.0, 0.0,..0.0]. It is computed and stored for each person separetely.
P^h	The new transition matrix is calculated by expanding the current transition matrix with multiplication from left.	The next transition matrix is calculated from the previous one. Now the transition probabilities becomes modified.

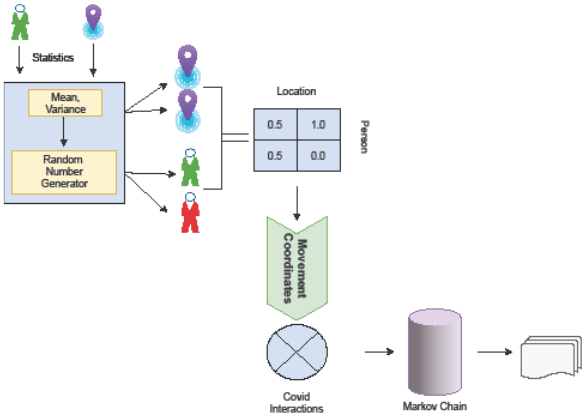

Figure 2. Data generation

In Figure 2, the statistics such as the population size, the percentage of mobility, the percentage of Covid-19 patients, and the number of different locations are used to generate the geo-coordinates (latitude and longitude) and the movement between these coordinates. All the data is generated through a Gaussian random number generator. At first, the number of locations, healthy people, symptomatic patients, and asymptomatic patients are determined. Later, the interaction between the location and the person is created by a random number generator and represented by a matrix. The contents of the matrix

represent the percentage of the time which is spent by a person in the location. Through using the matrix, the movement coordinates and the Covid-19 interactions are generated. If two-person violates the social distance rule then the maximum of the Covid-19 probability of the location and the Covid-19 probability of the interacted people is entered as the new Covid-19 probability of two-person. For example, if a person is tested Covid-19 positive then his/her interactions have the Covid-19 probability score of 1.0. If the person tested positive or the person gets Covid-19 then based on the stored and memorized stage of the person in the Markov chain, he/she gets the symptomatic or asymptomatic stage. A person exits the illness according to the time, and according to the next random probable state determined by the approximations in the transition states of the Markov chain.

In the data generation not only the states of the people are simulated but also the coordinates of their movement are modeled. In this case, a person either moves or waits according to the randomly determined degree. If a person is waiting then he/she can move backward or forwards. The backward movement happens only once in a one-time step. The time step is determined as 1 second, and the step size of the movement is 1 meter. These parameters are constant for every individual. There are two constraints: two people can not be in the same coordinate at the same time and the area boundary can not be cross passed so that all the people move freely in a closed area.

Based on the constraints described above the dataset is created by the Markov transition diagram is given in Figure 1. and the calculations are given in Equation (3) randomly. The state of each individual is stored by the state vector. At every time-step a movement data is created by moving an individual by a random direction in a closed area. The probability of the movement may create a collition. If a collision occurs the movement changes the direction by certain degree at random angles and continuous until no collision occurs. At every time-step the interactions are computed, if the movement yield an interaction then the state vector is updated based on the infectiousness of the interaction randomly. Approximately half of the interactions are assumed to be infectious if any of the two person is infected otherwise both person's states stays the same. In the final phase of data generation, each movement is recorded for each place of interest and person in the form of latitude and longitude. An example movement sequence is given in Table 2.

In Table 2. The small fragment of the generated dataset is shown. The dataset contains rows which represent a step. The rows are generated sequentially by updating the states of the individuals. The dataset contains latitude and longitude of the movement of a person inside a given place. The movement is determined randomly for each person. Each movement is done approximately in 1 meter diameter circle. The POI is the identifier of the place. The person is identified by CBG number. The Covid-19 status is defined in C19 column where the infections are marked by 1. In this dataset, there are two person. The second person is infected by Covid-19. The movement data is collected in a 11 seconds step range and the first person get close to the infected person at second 9. After this stage, the person becomes infected. When he/she gets infected the infection state is updated immediately and it is represented by the C19 column of the following step.

Table 2. An example movement data

Latitude	Longitude	POI	CBG	T	C19
50.0001	29.32001	122	001	1	0
50.0001	29.32001	122	001	2	0
50.0001	29.32002	122	001	3	0
50.0001	29.32003	122	001	4	0
50.0001	29.32004	122	001	5	0
50.0001	29.32005	122	001	6	0
50.0001	29.32006	122	001	7	0
50.0001	29.32007	122	001	8	0
50.0001	29.32008	122	001	9	1
50.0001	29.32009	122	001	1	1
50.0001	29.320010	122	002	2	1
50.0001	29.320011	122	002	3	1
50.0001	29.320012	122	001	4	1
50.0001	29.320013	122	001	5	1
50.0001	29.320013	122	001	6	1
50.0001	29.320012	122	001	7	1
50.0001	29.320011	122	001	8	1
50.0001	29.320009	122	001	9	1
50.0001	29.320008	122	001	10	1
50.0001	29.320009	122	001	11	1

The dataset contains the coordinates, place, time and Covid-19 status of each individual. During the evaluation only the information about the interactions are classified. In this respect, first the interactions are found, second the interaction states are determined based on the individuals. In this respect, only the new states of the individuals are predicted. The new state represents the state of an unknown, and previously undetected person. The evaluation of these new states gives a more reliable prediction accuracy since the interactions between two known Covid-19 patients do not contain any state changes and it can be easiy said that they are already patients.

4. FEEDBACK INFERENCE MODEL

The feedback model is built on the stream analyzing framework where the probability of an individual having a disease inferred through geospatial analysis of the previously generated statistics as in Figure 3. The feedback model estimates the Covid-19 probability of every individual through using the mobility patterns such as individuals' connections, the global and the local statistics of the previous time frame such as a day, or a week. The local and global statistics extracted from previous time frames are used to approximate the true probability score for the infections of the individuals.

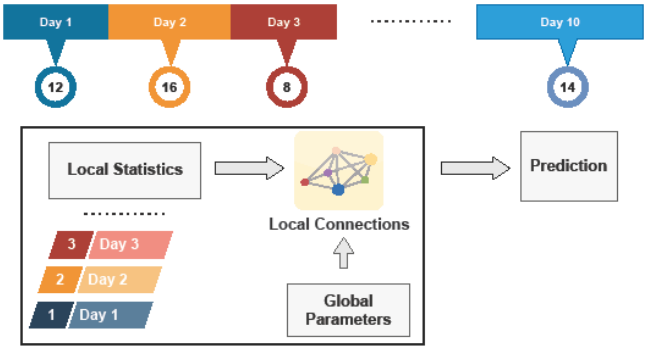

Figure 3. Feedback modelling

In Figure 3, the feedback modeling is depicted. The local and global statistics for the first 9 days is combined to predict Covid-19 risks for the 10th day. The local statistics are represented by the interaction risks of the individuals in a given specific location and in a given time frame. The global statistics represent the reproduction constants. For instance, if two person interacts the probability of getting an infection from each other is determined by these constants. This probability for each individual is calculated according to the equation given in Equation (4) and Equation (5).

$$score_i^h = \vec{\mu}_{local} \times \underbrace{\begin{bmatrix} 0.9 & \cdots & 0.1 \\ \vdots & \ddots & \vdots \\ 0.5 & \cdots & 0.4 \end{bmatrix}}_{Person-Location\ Matrix} + \vec{\mu}_{global} \qquad (4)$$

In Equation (4), the Covid-19 score of an individual becoming infected is calculated by multiplying the local parameters with the person-location matrix and adding to the global constants. The person-location matrix is the probability of getting infections in the specified location by a person. In this matrix, the rows represent the people and the columns represent the locations. Each element of this matrix represents the person's probability of infection in the specified location. For example, if the first person visits the second and thrid locations then the first row of this matrix will contain a zero value for the first location and the infection probability for the second and third locations. If there is not any data for the person-location matrix, then the probability of infection for the visited locations can be assumed as one. The Covid-19 score is not only dependent on the person-location matrix. It is also dependent on the interaction vector ($\vec{\mu}_{local}$) . In Equation (4), $\vec{\mu}_{local}$ represents the interaction risks of the ith person with the other three people in the h'th iteration. These risks can be estimated according to the time and distance proximity and the probability of being the Covid-19 patient. Through multiplication of the $\vec{\mu}_{local}$ vector with the person location matrix, only the interactions of the people on the same location will be accumulated. Also, there are base conditions related to the rate of the Covid-19. These are given as a global estimate vector represented by $\vec{\mu}_{global}$. The final result (probability) is the score of being a Covid-19 patient for a person in the specified location. There cannot be any conclusion drawn for the majority of the risk without knowing the local vector because the location-specific risk is assumed to be not only dependent on the locations but also the interaction with the people in those locations.

$$\vec{\mu}_{local}[j] = maximum(\frac{score_j + score_i}{2.0}, score_i) \qquad (5)$$

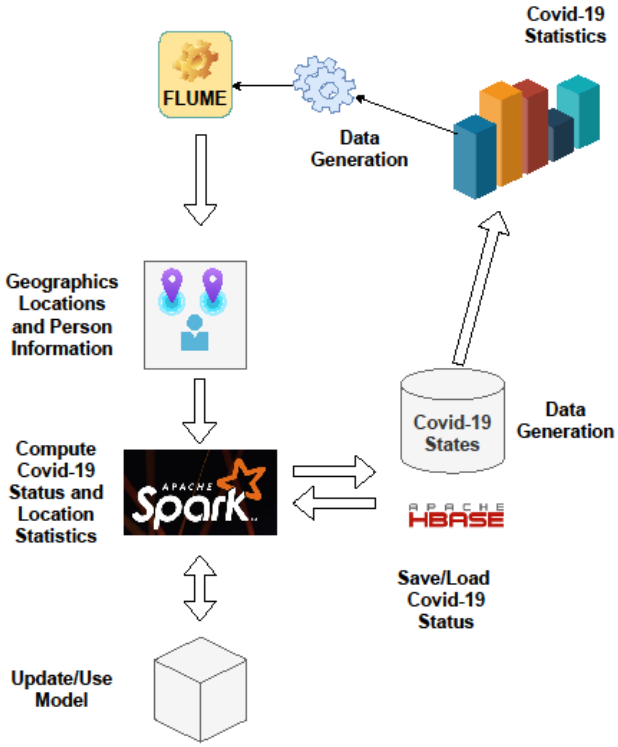

Figure 4. The Stream processing steps

Having an infection or not is determined by maximum averaging. The maximum averaging is given in Equation (5). In this equation, $score_i$ and $score_j$ represent the Covid-19 risk of the ith and jth person respecively. The equation assigns the maximum score of two person to the jth local index of the ith person. Thus, in the final decision the magnitute/norm of the $\vec{\mu}_{local}$ vector is used to calculate the total Covid-19 risk of the ith person. Each element of the vector represents the Covid-19 infection of the current person from the jth person. The infection risk is not only dependent on the Covid-19 risk of the other person but also the risks of the individual. Thus, taking the maximum of the average risk and the current person's risk is more appropriate than only considering the current interaction risk. Also, it should be noted that if a person has Covid-19 and his/her state is known. The risk of this patient is not updated by maximum averaging but his/her risk is used in both the person-location matrix in Equation (4) and maximum averaging in Equation (5). In this study, the global scores ($\vec{\mu}_{global}$) are not used. However, in general the global scores can be drawn from the total risk for every individual. The risk of an individual given by $score_i^h$ is the vector, which represents the total risks in each location. In order to find individual's final risk score, the norm of the vector is used.

In Figure 4, the stream processing framework is shown. All the theoretical model described in this paper is implemented by the tools in Apache Big Data Ecosystem. In order to process the streaming data, Apache Flume and Apache Spark stream processing frameworks are used. The data is generated for each person and location by using the Covid-19 states of the people and places. Information about the people and places is stored in Apache HBase. The generated data is pushed to the Apache Flume streaming engine and the data is fetched from Apache Spark for processing. During the processing, the Apache Hbase is used for storing the latest Covid-19 probability of individuals, the interactions, and the main statistics about the places. The Covid-19 probability and the place averages are computed in the Apache Spark streaming engine every 2 minutes. During the stream processing, the person and location identifiers are used to load/save the latest Covid-19 states and status from Apache HBase.

The stream processing framework (Apache Spark) is used to capture the interactions of geo-spatial events. These events contain the time and the coordinate of each mobile individual. To capture the interactions of individuals, the event data is indexed by geospatial hashing. Each coordinate is converted to 45 bit (1/0) geohash where any changes in the last bit correspond to approximately a 2.3-meter difference. The geocode of the event is hashed with the minute window of the time of the event so that the geospatial index can represent the time and the location of the event. Using geospatial indexing the events are clustered into bins. All the events in the same bin are sorted and approximate locations for as most as 6 seconds occurred one another are accepted as interactions. Later all the interactions are processed as explained above.

5. EVALUATIONS

The main concern in the evaluation of the pandemic model is the accuracy of the predictions. In many applications, the predictions include the number of patients, the number of hospitalized patients, and the number of deads for the next day or the next week. The number of deads is a major concern in pandemic analysis. So the prediction of the risks and applying the isolation procedures on time is very affective in reduction of the number of Covid-19 patients as well as the number of deads. For this reason, the evaluation interval for the streamed data is chosen as 2 minutes. Apache Flume and Apache Spark processing framework is used to capture the streamed data. The streamed data contains the person identifier, the coordinates of each person, the time, infected information, and the location id. The infected information denote whether a person is infected or not. This information is used for only evaluation.

In the data generation and evaluation stages several parameters are used. In Table 3, these parameters are given. The population size is denoted by CBG and the number of places is denoted by POI respectively. The population size is the total number of people who visited all the places. The maximum steps in the maximum amount of step taken by each individual. The number of steps are determined randomly in between the 300 steps and maximum steps. An individual takes a step in each second so that the step size of an individual is equal to the time spend in the place. In every step or second an individual moves by one meter. The social distance constant is the safe distance for every individual. A virus transmission may occur when any two individual get closer than the social distance. Thus, increasing the social distance increases the risk of infection by assuming that two person are not safe with-in a large social distance. Batch frame constant is the window-sampling time for all the events in the dataset. Using two minute window, we can group more events in the same window bucket. If the events of two individuals gets in the same window than these events may contain an interaction. The time constant is the maximum duration for the interaction. If any two events in the same location have occured in 6 seconds gap than these events are assumed to be an interaction. Increasing the time constant will eventually increase the number of interactions.

Table 3. Parameters for data generation and evaluation

Parameter Name	Parameter Type	Parameter Value
Population size	Variable	100, 400, 1600
Location size	Variable	1, 4, 16
Patients	Percentage	70
Asymptomatic Cases	Percentage	30
Sumptomatic Cases	Percentage	70
Maximum Seconds	Range	600, 1800,5400
Social Distance	Constant	2 meters
Covid-19 Threshold	Constant	0.88
Batch Frame Time	Constant	2 Minutes
Time Constant	Constant	6 seconds
Data Range	Constant	10 days
Isolation Range	Constant	7 days

The date range is the number of days that is used for generating the data. The date range is inversely proportional to the density of the mobility. If the date range is large then the possibility of an interaction is low, else vice-versa. The isolation range is the number of days in isolation. If a person is known to be a patient then he/she gets isolated for 7 days. The Covid-19 threshold is the value to accept whether a person is infected or not. The Covid-19 value is determined based on the number of positive interactions. If the calculation of the Equation (5) is above the Covid threshold then it is assumed to be an infection. The threshold is chosen as 0.88 because it is assumed that the number of interactions with more than three Covid-19 patients (same or different) infects the healty individual.

Each person also has a Covid-19 probability value which is continuously updated in each interaction and each person also has a true Covid-19 state which is generated and updated during the generation stage. The batch frame time is used to predict the Covid-19 probability of the person. Every 2 minutes, the evaluation for the event dataset occurs. If the person has a Covid-19 value greater than the

threshold then he/she is accepted as positive otherwise he/she is accepted as negative. The prediction for each person of having a Covid-19 positive or negative is compared with the true state of the person. The predictions are measured according to the positive and negative cases separately. The average scores such as true-positive, true-negative, false-positive, and false-negative are calculated.

The Covid-19 prediction of each person is computed by measuring the maximum Covid-19 probability of the interaction. An interaction contains two people and a place. The person and the place are represented by a unique identifier and a Covid-19 score. The estimate of the current Covid-19 risk for a person is the person's score and the average Covid-19 score of the people during visiting the place is the place's score. If the interaction doesn't contain any risks then the person gets the global constant score based on the calculation given in the feedback inference model. If the person has a high Covid-19 probability then his interactions get the same Covid-19 score too. If the probability of Covid-19 is greater than the threshold then he/she is accepted as Covid-19 positive and his/her interactions would have an above-average Covid-19 score. So, a person's having an infection is conditioned on his/her interactions, the location visits, and the average number of people infected in these places. In the calculations, the global constant vector is discarded, and the person-location risk matrix is computed by the average number of interaction scores in the location. For the initial value of the person-location matrix, the values are set to the percentage of visits to the location. In this case, if a person visits 3 locations the row values of this person will be 1/3. So both the global constant vector and the person-location matrix are chosen as same for every individual.

6. RESULTS AND DISCUSSION

In this study, the datasets are created randomly. Kappa statistics are used to measure the significance of the proposed method according to the random prediction [30]. To measure the Kappa statistics, the confusion matrix of the predictions is used. A Kappa score above 0.5 implies that the proposed approach is significantly better than random chance. In the evaluations, along with the Kappa score, the F-measure is used. These evaluation measures are given in Table 4 and Table 5 where true positive (tp), false positive (fp), true negative (tn), and false negative (fn) rates are used for f-score and Kappa score calculations. In Table 6, these scores are given for each population size, place of interest, and day range.

Table 4. Confusion matrix

	Actual Positive	Actual Negative	Total
Positive	tp	fp	m1
Negative	fn	tn	m0
Total	n1	n0	n

Table 5. Evaluation measures

	Formulations
n0	$fp + tn$
m1	$tp + fp$
precision	$tp/(tp + fp)$
recall	$tp/(tp + fn)$
p0	$(tp + tn)/n$
pe	$((tp + tn) * (tp + fn) + (fn + tn) * (fp + tn))/n^2$
Kappa	$(p0 - pe)/(1 - pe)$
f-score	$2 * precision * recall / (precision + recall)$

In Table 5. the calculation of evaluation measures are given. In Kappa score, the p0 indicate the accuracy of the proposed framework and pe is the random prediction accuracy based on the ratio of the most probable cases. The measures of F-Measure and Kappa for different datasets are given in Table 6. In Table 6. 18 different datasets are given. These datasets are randomly generated where individual's steps are determined randomly. Each row of the dataset corresponds to geographic location of a person in a given POI at each time step. The step size corresponds to minimum number of steps for an individual to randomly take. Increasing the step size increases the mobility density. For example, the dataset with a 200 step size has a lower mobility density than the dataset with 5400 steps. The mobility density is also proportional to CBG size. Oppositely, the mobility density is inversely proportional to POI size. Along with POI size, each dataset has a random number of interactions. Because the interaction size is dependent on the movement coordinates and the movement of individuals is randomly generated. The number of interactions is given in interaction size.

Table 6. Evaluation results

POI Size	CBG Size	Step Size	Interaction Size	F-Score	Kappa
1	100	600	480	84.197	72.66
1	100	1800	825	86.019	75.492
1	100	5400	1718	85.466	74.610
1	400	600	2020	80.575	73.251
1	400	1800	7450	82.255	77.023
1	400	5400	8215	84.615	71.962
4	100	600	3483	83.172	71.259
4	100	1800	41632	84.538	73.245
4	100	5400	45712	75.061	83.962
4	400	600	10243	77.544	64.664
4	400	1800	11798	84.718	73.621
4	400	5400	49347	84.895	73.787
16	100	600	34604	84.620	73.101
16	100	1800	35769	85.284	74.379
16	100	5400	39872	85.284	74.379
16	400	600	51304	84.673	73.452
16	400	1800	111349	84.538	73.245
16	400	5400	84895	84.538	73.245

The performance measures are obtained by comparing the predicted Covid-19 status with the generated case for every new interaction. According to theses results, increasing the number of interactions between people neither has a negative impact on the prediction performance nor on the Kappa score. Thus, it can be said that determining the individuals Covid-19 risk through other individuals risks, location risks and the interactions is done independently from the mobility density. On the contrary, if the risks are computed by assuming that each interaction with a Covid-19 patient infects the other person, we can not observe Kappa scores above 0.5 because every person would be infected and the prediction performance will be close to 1. From the Kappa scores, it can be said that the increased number of interactions increases the possibility of getting an infection and the model prediction approximates this change appropriately.

In Table 6. the scores indicate that the perfomance of the Covid-19 infection for each persons' interaction is better than determining the persons getting infected from a Covid-19 patient by pure chance. The performance scores are significant according to the Kappa scores. Kappa scores indicate that the prediction is reliable.

The final remarks for the dataset is the baseline performance. The baseline performance of pure chance can be calculated by selecting whether a person infected or not by 0.5 chance. Then for half of the interactions may emit an infection if one of the person is a Covid-19 patient. In this case, let's say mobility size and population size increased to a hypothetical limit where all the interactions are infectious. Then the accuracy will be approximately 50%. So, the question is how the proposed framework perform better than 50%. Because the model uses the total number of interactions of a person as well as the location risk. In this case, whether the person get infected is dynamically calculated based on these priors. If the density increases the possibility of getting infection increases. The proposed framework approximates this possibility appropriately.

7. CONCLUSION

In this study, a geo-spatial analyzing framework is proposed for simulating the pandemic conditions of every person. First, the movement and status data for each individual is generated, then the mobility of each person is aggregated in an interval and the probability of his/her infection is estimated using his/her interactions, and the visited places. The framework consists of three stages; the data generation stage, the processing stage, and the evaluation stage. Four variables are used in the data generation stage; these are the population size, the number of places, the maximum step size, and the time interval. Along with these variables, the data generation of each individual consists of the geospatial movement patterns, and the Covid-19 state of the individual. The Covid-19 states of the individuals are derived from the Markov chain where the probability estimates of the global Covid-19 rates are modelled. During the data generation, the chance of getting infected and the paths of the infection are determined randomly by using the transition probabilities of the Covid-19 in the Markov chain.

For varying set of parameters, the performance of the proposed framework is evaluated through the Kappa and F-Measure of Covid-19 status of every interaction. Only the interactions of the individuals having unknown Covid-19 status are considered in evaluations. Based on the Kappa statistics, the proposed framework is significantly better than random guesses and the F-measure scores indicate that the event/location based interaction statistics is an effective measure for Covid-19 prediction of individuals.

REFERENCES

[1] Y. Zeng, X. Guo, Q. Deng, S. Luo, H. Zhang, "Forecasting of COVID-19: spread with dynamic transmission rate", *Journal of Safety Science and Resilience*, 1(2), 91–96, 2020.

[2] A. Singhal, P. Singh, B. Lall, S. Joshi, "Modeling and prediction of COVID-19 pandemic using Gaussian mixture model", *Chaos, Solitons and Fractals*, 138, 2020.

[3] L. Basnarkov, "SEAIR Epidemic spreading model of COVID-19", Chaos, Solitons and Fractals, 142, 110394, 2021.

[4] A. Şenol, Y. Canbay, M. Kaya, "Trends in Outbreak Detection in Early Stage by Using Machine Learning Approaches", Bilişim Teknolojileri Dergisi, 14 (4), 355-366, 2021.

[5] W. Getz, R. Salter, O. Muellerklein, H. Yoon, K. Tallam, "Modeling epidemics: A primer and Numerus Model Builder implementation", *Epidemics*, 25, 9-19, 2018.

[6] Adiga, A, Dubhashi, D, Lewis, B, Marathe, M, Venkatramanan, S, Vullikanti, A. "Mathematical Models for COVID-19 Pandemic: A Comparative Analysis", *Journal of the Indian Institute of Science*, 100(4), 793–807, 2020.

[7] O. Bjørnstad, K. Shea, M. Krzywinski, N. Altman, "Modeling infectious epidemics", *Nature methods*, 17(5), 455–456, 2020.

[8] S. Olaniyi, O. Obabiyi, K. Okosun, A. Oladipo, S. Adewale, "Mathematical modelling and optimal cost-effective control of COVID-19 transmission dynamics", *European Physical Journal Plus*, 135(11), 938, 2020.

[9] A. Isdory, E. Mureithi, D. Sumpter, "The impact of human mobility on HIV transmission in Kenya", *PLoS ONE*, 10(11), 2015.

[10] K. Al-Kindi, A. Alkharusi, D. Alshukaili, N. Al Nasiri, T. Al-Awadhi, Y. Charabi, A. El Kenawy, "Spatiotemporal Assessment of COVID-19 Spread over Oman Using GIS Techniques", *Earth Systems and Environment*, 4(4), 797–811, 2020.

[11] H. Unwin, S. Mishra, V. Bradley, A. Gandy, T. Mellan, et. al. "State-level tracking of COVID-19 in the United States". *Nature Communications*, 11(1), 1–9, 2020.

[12] J. Sousa, J. Barata, "Tracking the Wings of Covid-19 by Modeling Adaptability with Open Mobility Data", *Applied Artificial Intelligence*, 35(1), 41–62, 2021.

[13] F. Shahid, A. Zameer, M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM", *Chaos, Solitons and Fractals*, 140, 110212, 2020.

[14] C. Yeşilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm", *Chaos, Solitons and Fractals*, 140, 2020.

[15] N. Punn, S. Sonbhadra, S. Agarwal, "COVID-19 epidemic analysis using machine learning and deep learning algorithms", *medRxiv* 2020.04.08.20057679, 2021.

[16] V. Chimmula, L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks", *Chaos, Solitons and Fractals*, 135, 2020.

[17] A. Ramchandani, C. Fan, A. Mostafavi, "DeepCOVIDNet: An Interpretable Deep Learning Model for Predictive Surveillance of COVID-19 Using Heterogeneous Features and Their Interactions", *IEEE Access*, 8, 159915–159930, 2020.

[18] Kafieh, R, Saeedizadeh, N, Arian, R, Amini, Z, Serej, N, Vaezi, A, Javanmard, S. "Isfahan and Covid-19: Deep spatiotemporal representation", *Chaos, Solitons and Fractals*, 141, 110339, 2020.

[19] A. Rodriguez, N. Muralidhar, B. Adhikari, A. Tabassum, N. Ramakrishnan, B. Prakash, "Steering a historical disease forecasting model under a Pandemic: Case of Flu and COVID-19", 2020.

[20] S. Chang, E. Pierson, P. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, "Mobility network models of COVID-19 explain inequities and inform reopening", *Nature*, 589 (7840), 82–87, 2021.

[21] Y. Chen, Q. Li, H. Karimian, X. Chen, X. Li. "Spatio-temporal distribution characteristics and influencing factors of COVID-19 in China", *Scientific Reports*, 11(1):3717. PMID: 33580113, 2021.

[22] D. Balcan, V. Colizza, B. Gonçalves, H. Hud, J. Ramasco, A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases", *Proceedings of the National Academy of Sciences of the United States of America*, 106 (51), 21484–21489, 2009.

[23] A. Gatrell, T. Bailey, P. Diggle, B. Rowlingson, "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology", *Transactions of the Institute of British Geographers*, 21 (1), 256, 1996.

[24] M. Kulldorff, R. Heffernan, J. Hartman, R. Assun\ccão, F. Mostashari, "A space-time permutation scan statistic for disease outbreak detection", *PLoS Medicine*, 2 (3), 0216–0224, 2005.

[25] T. Ng, T. Wen, "Spatially Adjusted Time-varying Reproductive Numbers: Understanding the Geographical Expansion of Urban Dengue Outbreaks", *Scientific Reports*, 9 (1), 1–12, 2019.

[26] F. Rihan, H. Alsakaji, C. Rajivganthi, "Stochastic SIRC epidemic model with time-delay for COVID-19", *Advances in Difference Equations*, 2020(1), 502, 2019.

[27] A. Arenas, W. Cota, J. Gómez-Gardeñes, S. Gómez, C. Granell, J. T. Matamalas, D. Soriano-Paños, and B. Steinegger, "Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions." *Physical Review X*, *10*(4), 041055, 2020.

[28] M. Pinsky, and K. Samuel, **An introduction to stochastic modeling**, Elsevier Inc, 2010.

[29] Internet: Census Block Group Data, SafeGraph Data. https://docs.safegraph.com/docs/open-census-data, 05.17.2022.

[30] A. Viera, J. Garrett, "Understanding Interobserver Agrement :The Kappa Statistic", *Fam med*, 37(5), 360–363, 2005.