IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 20, 2021, accepted June 20, 2021, date of publication June 22, 2021, date of current version June 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3091655

Multiple Disease Hotspot Mining for Public Health Informatics in Resource Starved Settings: Study of Communicable Diseases in Punjab, Pakistan

FATIMA KHALIQUE AND SHOAB AHMED KHAN
National University of Sciences and Technology, Islamabad 46000, Pakistan

Corresponding author: Fatima Khalique (fathema.khalique@gmail.com)

ABSTRACT Communicable diseases remain a significant challenge for public health management. In particular, for resource scarce settings, it is important to understand the linkages and dynamics of multiple diseases that share common resources, space or time. We develop a framework, called Multiple Disease Management Framework (MDMF) based on machine learning approach for managing multiple diseases occurring in close space and time to identify locations that experience high disease burden rates. We use 8 water related disease incidence data in Punjab, Pakistan from year 2013 to 2019 to investigate interactions among hotspots of different diseases. However, the model is scalable and can be applied to any number of diseases. The hotspot analysis involves multi-level clustering and tagging of individual disease incidence streams that generates a distance based graph over a geographical area and is then integrated into a single stream in the framework to identify final sensitive locations called cluster alarms. The initial individual disease clustering yielded number of clusters as high as 24 clusters for each disease with up to 16 neighboring clusters of other diseases of similar sizes. The cluster tagging and multi-level clustering process was able to identify as low as 19 locations of cluster alarms across the whole province of 38 districts. The identification of high disease hotspots and their dynamics with the neighboring hotspots of multiple diseases allows identification of locations with higher need of related public health resources. This identification is very critical for national health agencies for optimal allocation of resources and devising an effective intervention programs.

INDEX TERMS Cluster analysis machine learning, multiple diseases, public health informatics, public health management, spatio temporal analysis.

I. INTRODUCTION

The increasing number of public health surveillance programs for both communicable and non communicable diseases has increased our understanding of health status of population. Traditional disease surveillance programs gather data from multiple sources and each program has its own data control variables and their particular formats. As a result, disease specific research studies are more common as the evidence base is expanding. In recent years, the availability of high resolution spatial and temporal big health data has opened avenues beyond traditional methods used in public health management. Considering, many of these diseases may not be occurring in isolation, for example, due to ever changing climatic and social conditions, the disease dynamics investigated in silos cannot be entirely relied on. It can be evidently concluded that in any given population, multiple diseases influence the burden and predictive certainty for policy and intervention in resource allocation.

In order to effectively mobilize the resources, it is important to identify critical areas. The term hotspot has been coined to mean any critical-situation location in context of security, disaster management, disease management or any other emergency situation. Data from multiple sources is now available in fine resolutions to understand the disease dynamics. While investigating disease patterns in silos helps

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

to understand a single disease dynamics in public health policy and resource allocation context, it is vital to understand the interactions and dynamics of multiple diseases through exploratory analysis. This understanding helps the evidence based decision making for public health management.

This study expands on the single disease hotspot analysis research concepts. It involves detecting and connecting dense hotspots of different diseases that share some common underlying factor, such as resources, symptoms, causes etc. The aim of the study is to integrate multiple disease surveillance information and combine it with machine learning based analysis approach to study multiple disease dynamics and yield hotspots or cluster alarms. We investigate the relationships among multiple diseases based on spatial data and are able to identify cluster alarms. We employ a clustering based approach combined with graph theory to identify high disease density area or hotspots and their dynamics with the neighbouring hotspots of different diseases. The framework also generates intermediate exploratory sub graphs that can be analyzed in details for investigations into sub population's health status for selected diseases. This approach allows application of solutions from multiple domains to study different dimensions of the resource allocation problem in spatio-temporal multiple disease regions. In addition, the presented framework is able to incorporate any number of diseases and allows easy injection of external factors such as climatic data, socio economic factors, and environmental factors etc. The presented analysis is helpful in multiple disease management in resource scarce settings for guiding decision making process. Many diseases share common resources in case an outbreak occurs and for a resource scares settings, we are able to identify cluster alarms where resources can be optimally utilized. In addition, this work serves as an exploratory mechanism for public health authorities to manage multiple diseases over a geographical in a population or sub population. The integrated approach towards using multiple disease fills the overlapping clusters over an area that are missed by single source or single disease analysis. The work aims to contribute towards developing a data driven multiple disease model that can be used for achieving public health targets through identification of geographical locations that are more resource sensitive given the distribution and dynamics of multiple diseases.

II. RELATED WORK

Increasing prevalence of communicable and non-communicable diseases all over the globe has made public health policy management and resource allocation a challenging task [1]. Particularly in resource-limited locations, as common in lower and middle income countries, it becomes of high importance that appropriate policies and intervention programs be designed to target areas fairly and efficiently [2]. Traditional diseases surveillance programs are an effective way to collect such information [3]–[5]. However, the one disease per program generates repeated data that remains under utilized for a potential broad view on public health status over a geographical area [6]. The availability of electronic health care records is unlocking the potential for new approaches towards disease modelling and their interactions or dynamics with other diseases. An insight into the interplay between different disease networks can be conducted to explore individual and combined interactions among multiple disease prevalence in a given population. Diseases with common underlying sources, causes in environment, symptoms or other features can be studied in relation to one another to represent the complex dynamics of their combined networks.

Data analytics approaches in general and machine learning (ML) methods in particular have improved insights into single disease investigations and its use in public health has a promising future to revolutionize the evidence based policy making and designing intervention programs [7]–[9].In addition several examples in literature exist for using machine learning approaches in epidemiology context for public health [10], [11]. Reference [12]. Studying disease clusters in public health context has helped guide public health agencies in devising an appropriate response to developing disease situation [13], [14]. Many surveillance systems support early warning techniques for public health authorities [15], [16]. Existing work is available using statistical and AI based methods to detect spatio-temporal disease hotspots including techniques eigenspot method [17], ScanStatistics [18], M-statistic [5] and using machine learning algorithms [19]. The term, cluster alarms is also used to mean location with accessive disease intensity. Hotspot or cluster alarm identification is crucial to early warning disease surveillance systems.

This work presents a multiple disease management framework that is intended to aid public health authorities in understanding the disease interactions in case of multiple outbreaks by representing their relationship with one another using machine learning approaches. The integrated approaches to study multiple diseases in context of public health management and resource utilization using advents of modern research in artificial intelligence and machine learning techniques is an under studied potential research area. The aim of this work is to create an exploratory analysis framework that allows investigations into different aspects of single and multiple diseases over space and time for public health informatics using machine learning techniques. This information is specially important when resources are scarce and there are common resource utilizing diseases that are affecting the population in a given space.

III. METHODS

A. MULTIPLE DISEASE MANAGEMENT FRAMEWORK (MDMF)

We present a machine learning based multiple disease management framework (MDMF) for disease analysis in public health management perspective and efficient resource allocation in a resource competitive settings. The MDMF acquires,

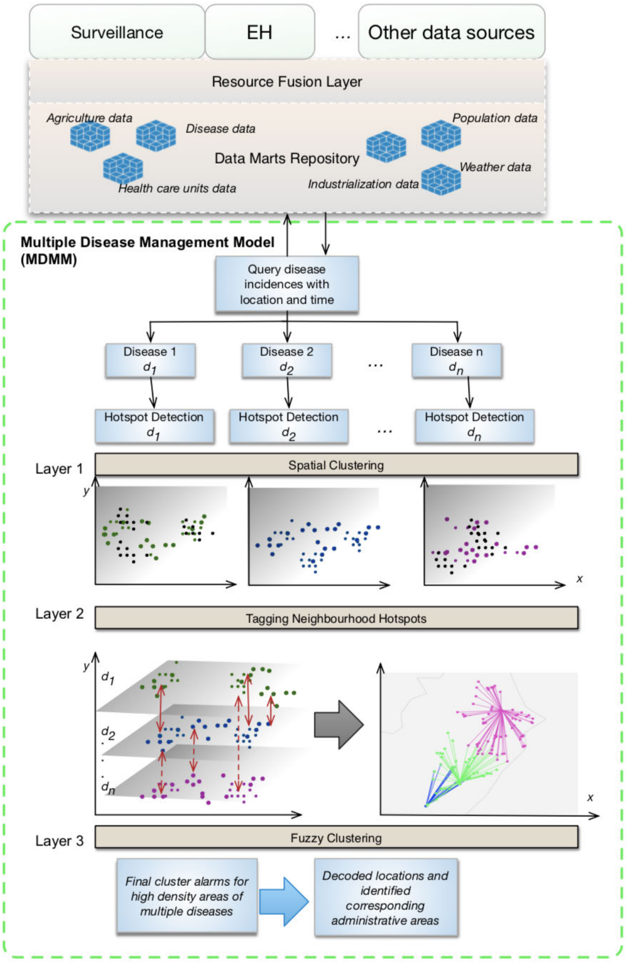


FIGURE 1. **Multiple Disease Management Framework (MDMF) with three main layers 1) Layer1; where spatial disease hotspots are identified individually 2) Layer2; neighbourhood hotspot detection among different diseases 3) Layer 3; all disease hotspots are integrated to identify cluster alarms as resource sensitive locations using fuzzy clustering.**

integrates and distribute health data based on public health requirements. Fig.1. outlines the stages and strategical placement of multiple disease representation and analysis framework within the over all scheme.The architectural aspects of the framework are discussed in detail in [20]. The analysis presented resides at the analytical layer that extracts data from the data marts and run its 3-stage algorithm for identifying locations that are resource sensitive.

1) HOTSPOT IDENTIFICATION

Formally, each disease incidence data is a geo-referenced record defined as tuple e

$$e_i = (x_i, y_i, t_i) \tag{1}$$

where x, y and t represent the longitude, latitude and timestamp of disease i^{th} incidence event. Thus each disease event is represented in terms of spatial features and each event is considered independent of the other. For a set of diseases $d_1, .., d_D$, where D is the total number of diseases. we select the events corresponding to each disease d_i for the study time period T.

A disease d_i data stream is clustered into n clusters C_j where $j = 1, ..n$. We define cluster C_j by a tuple such that

$$C_j = (id_j, H_j, e_i{}^j, m_j) \tag{2}$$

where id is the cluster identification, H_j is the centroid or hotspot in the cluster C_j, $e_i{}^j$ is all events i assigned to cluster j and m_j is the number of disease events in cluster C_j. Each hotspot is represented by a latitude and longitude, that is,

$$H_j = (x_j, y_j) \tag{3}$$

In this study we employ k-means for the cluster formation and centroid identification. However, other hotspot identification algorithms can also be applied to compliment one another for an investigative analysis. The hotspots vary in the density as well as population distribution. Therefore each hotspot is assigned a weight based on population density.

2) L2 TAGGING

In the second layer, as a measure of diseases dynamics in presence of other diseases, we find the nearest hotspots for all diseases in relation to one another. This concept of nearness can be calculated based on multiple factors, for example, geographical location, time, similarity of climatic conditions and socio-economic factors etc. In this article, we find the spatially neighbour hotspots of multiple diseases. For this purpose, a distance matrix M_{ij} is created between hotspots $HS_{ij}, HS_{i'j'}$ for every two diseases d_i and d_i' where $i \neq i', i' > i$ and n, m is the number of clusters of disease d_i and d_i' respectively. Therefore, $size(M_{ij}) = n \times m$. The distance matrix is created based on the Euclidean distance between longitude and latitude of hotspots of each cluster of every disease. Since for this particular analysis, we are only interested in hotspots in terms of spatial locations of events, it makes sense to define nearest hotspot in terms of geographical distance as well. However, generally in the framework, the distance can be defined to mean multiple features of disease distribution dynamics.

The distance matrix $M_{ii'}$ defines a weighted graph G consisting of non empty set of hotspots represented as nodes. The disease hotspot association graph consists of nodes representing the hotspot and edges based on the nearest disease neighbour hotspot as shown in Fig. 2.

The neighbours are tagged based on threshold applied over M_{ij}. The value of threshold can be user defined based on the requirements. The threshold is a set of constraint conditions that can be based on cluster size, distance, population density of the centroid district or union or intersection combination of these factors. For example, for smaller spatial units, a fix value based on radii around the hotspots identified may be specified or a weight or intensity based threshold be set for the hotspot zones. Hotspot can vary from disease to disease. For example, for dengue, a 50 cases maybe low but for

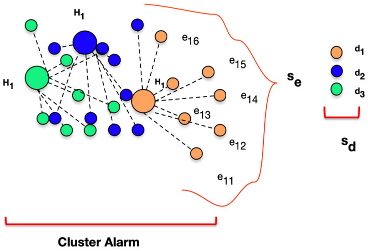


FIGURE 2. Hotspots for disease d_1, d_2, d_3 connected to each other based on the spatial distance. A cluster Alarm A representing 3 disease hotspots with multiple disease events e in each hotspot.

CoVID, it will represent a high value. In this research, we use population density based average inter-district distance for defining the threshold in order to create a better precision for neighbourhood selection. This threshold gives an approximation of neighbourhood with the high density related disease hotspots.

3) L3 CLUSTER ALARMS

As the final layer of MDMF for identifying the most resource starved locations for multi disease management, we find the cluster alarms based on second iteration of hotspot identification and calculate clustering coefficient of our threshold based pruned graph that gives the local density of multiple diseases. At this level, the clustering algorithm is run to determine geographical connectedness of the initial hotspots of individual diseases as well as identifying locations and their administrative districts within the province. The resulting cluster alarms identified can be further analyzed in epidemiology context by throwing in more parameters inside the model. The individual disease hotspots fulfilling the threshold are clustered collectively to select multiple disease hotspots and generate centroids that represent the cluster alarms A. The new clusters obtained are defined by (4).

$$C' = (id, A, H', s) \tag{4}$$

where id is the cluster identification, A is the centroid or cluster alarm of the cluster C', $H' \subset C$ is the set of hotspots of different diseases that form the new cluster and s is the total number of H in C'.

The new clusters are analysed based on number of disease hotspots in the cluster s_h and total number of disease incidences belonging to all clusters. Each A is assigned a low, medium, high based on number of disease clusters and number of disease incidences in a disease cluster. Therefore, cluster alarm is defined as (5)

$$A = (id, s_d, s_e) \tag{5}$$

where

$$s_e = \sum_{i=1}^{u} e'_{h(i)} \tag{6}$$

and id is the cluster id, s_d is the count of multiple disease hotspots participating in the cluster alarm, and s_e is the sum of all disease incidences of every disease in corresponding hotspots of the cluster alarm. Every A is categorized as low, medium and high based on two inputs s_d and s_e as shown in Fig. 2. The first input, that is, number of diseases belonging to a particular C' are assigned membership low, medium and high. Similarly, second input, s_e, is assigned low, medium, and high. For a two input tagging process, nine rules are obtained elaborated by an expert that classifies each alarm A as low, medium and high. For example, if numbers of diseases is low in a cluster and total elements in the cluster is low, A is low and the corresponding location is a less sensitive area. However, a high s_d and high s_e will also give a high A that implies and highly sensitive resource allocation location. These A locations are mapped geographically and their corresponding districts are identified for public health related decisions and interventions. Additionally a step back in the process, allows exploratory graph analysis for different disease interactions.

TABLE 1. Variables and their description used in the model.

Symbol	Description	Symbol	Description
d	disease	x	Longitude
e	disease event	y	Latitude
C	Cluster	n	number of clusters
H	Hotspot	C'	All hotspot cluster
m	$\sum e \subset C$	H'	$H \subset C'$
e'	$e \subset C$	A	Cluster Alarm
M	Distance matrix	s	$\sum H \subset C'$

B. STUDY AREA AND DATASET

We applied our method for studying eight water related diseases in Punjab, Pakistan. Punjab is the largest province of Pakistan with estimated population of 110,012,442 according to 2017 census results. In addition, as compared to other provinces of Pakistan, Punjab is most urbanized with 40% urban population as well as industrialized province of Pakistan. It consists of 38 districts and has major share of its budget spending in health sector (84.8%). According to 2017-2018 report in Pakistan, the doctor to population ratio, is 1:95, the dentist to population ratio is 1:9730 and bed to population ratio is 1:1580 where as this ratio is as high as 13.05 and 8.05, per 1000 people for Japan in Asia, Russia in Europe respectively. Therefore, in a province with highest population, industrialization and health expenditure in the country, it is imperative to identify most affected and needy areas to optimally utilize the scarce available resources for effective public health interventions.

The data set was obtained from the passive disease surveillance system operated by Punjab Information Technology Board (PITB) [21], where data is reported through all levels of health authorities in Punjab, online portal, mobile application and WHO reports. The data from year 2013 to 2019 was analysed for eight individual diseases namely, Typhoid, Influenza, Dengue, cholera, Acute Viral Hepatitis (AVH), Scabies, Gastroenteritis, and Malaria. Table.2. shows the disease incidences for each disease across the province

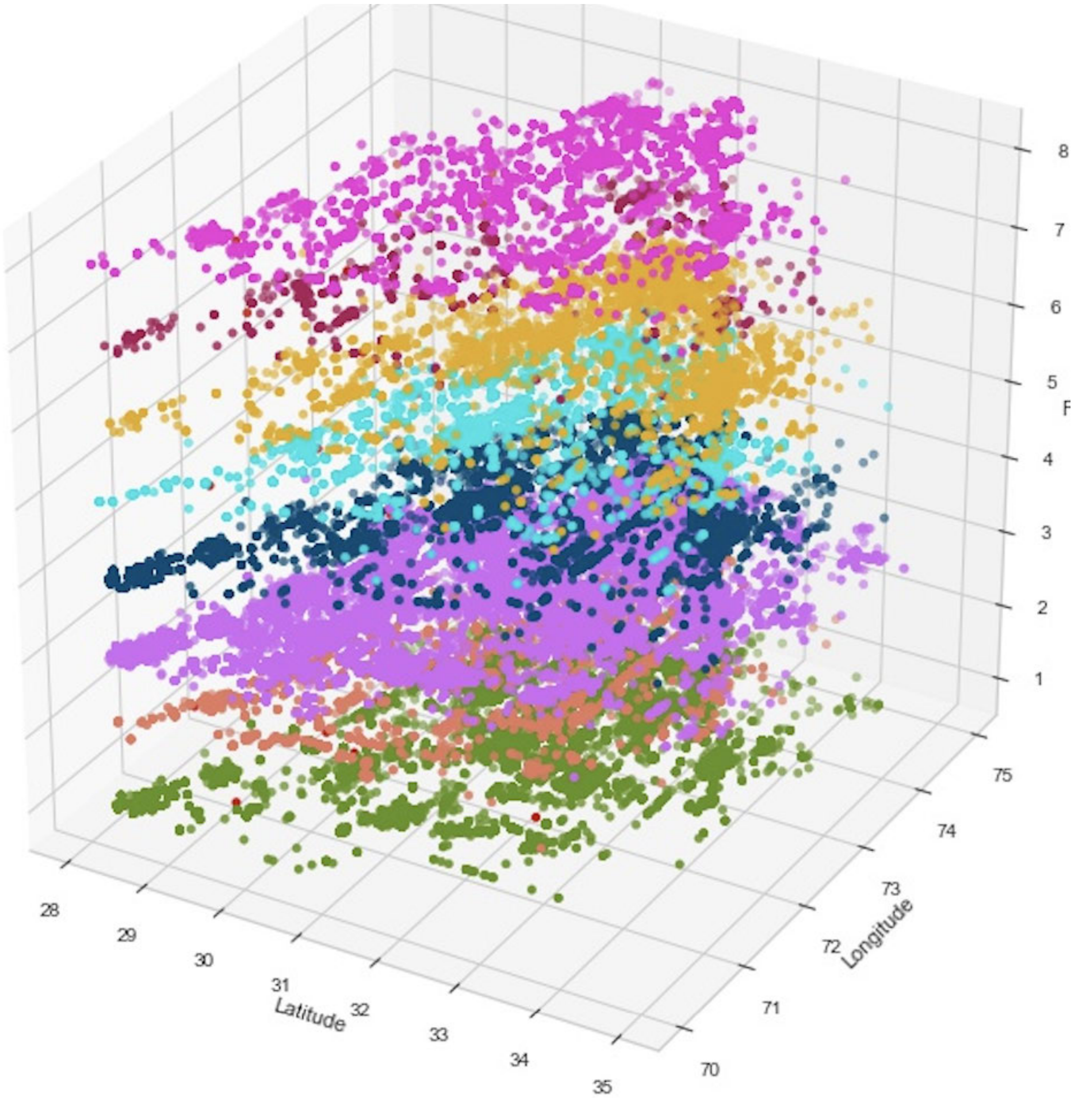


FIGURE 3. Spatial distribution of eight selected diseases across Punjab, Pakistan.

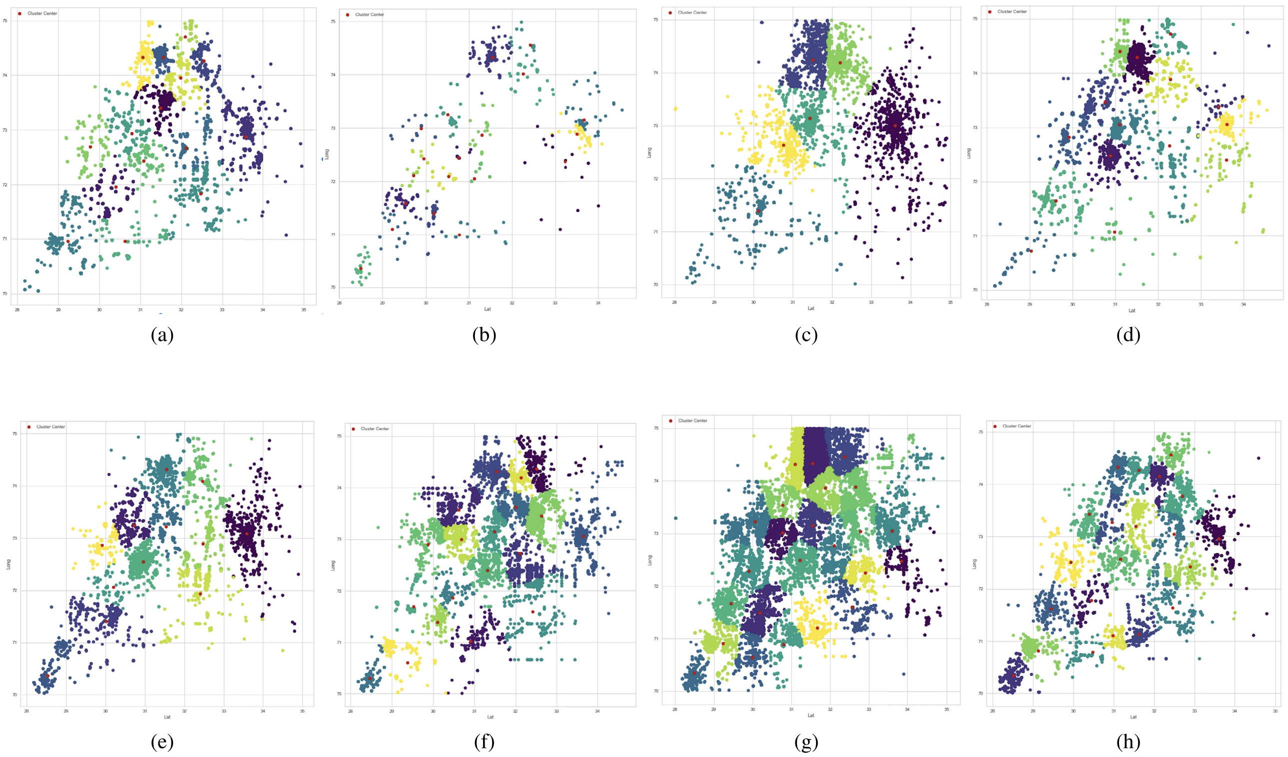
during the study period. For each disease and its reported incidence, the location of incidence in terms of latitude and longitude, and time of incidence in terms of day of the month of year is extracted. The latitude and longitude of all incidences are geo coded to be analysed on a polygon layer of administrative map of Punjab.

IV. RESULTS

The three layered model algorithm is applied to the eight disease point incidence data set. Fig. 3 shows the spatial distribution of selected diseases across the study area. The number of cases for each disease during study time period is shown in Table.2. For this study, we are interested in the spatial features of the selected diseases. For layer one, the hotspots are identified for each disease individually, independent of other diseases. Table2 shows the number of clusters identified during layer 1 clustering for each disease. Fig.4 shows the spatial clustering results when k-means is applied on each disease incidence data individually. For the first layer of clustering, 6 clusters were found for dengue, 12 for AVH, 15 for typhoid and cholera, 19 for scabies and influenza, 22 for malaria and 24 for Gastroenteritis disease. Each cluster

TABLE 2. Number of disease cases during study period in Punjab, Pakistan and number of hotspots identified in the region based on individual disease streams.

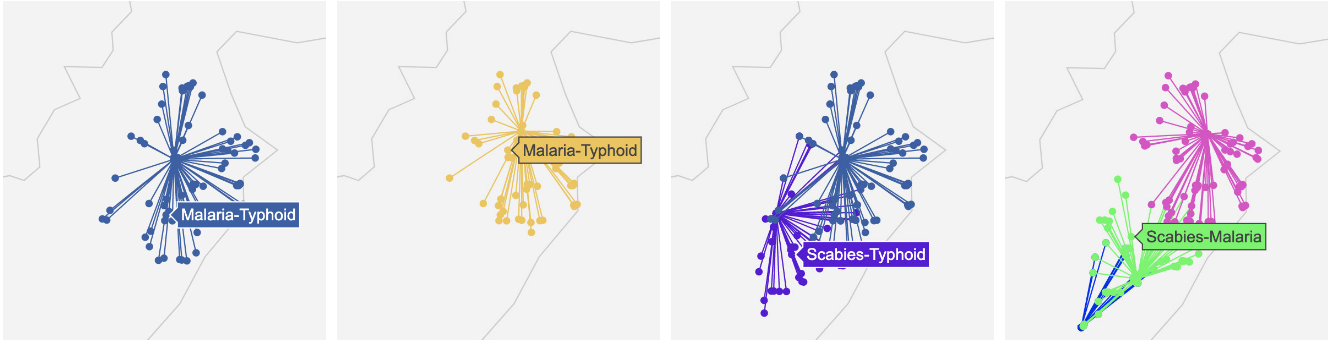
	Disease	Cases	Hotspots
1	Typhoid	34860	15
2	Influenza	3681	19
3	Dengue	28372	6
4	Cholera	40332	15
5	AVH	40208	12
6	Scabies	161907	19
7	Gastroenteritis	908062	24
8	Malaria	26254	22



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 4. K-means clustering results when applied on incidence data for: (a) Typhoid (b) Influenza (c) Dengue (d) Cholera (e) AVH (f) Scabies (g) Gastroenteritis (h) Malaria. Clusters are distinguishable based on colour with cluster heads in red.



(a) A single hotspot of malaria and its dynamics with other diseases' hotspots

(b) Another hotspot of malaria and its dynamics with all other diseases' hotspots

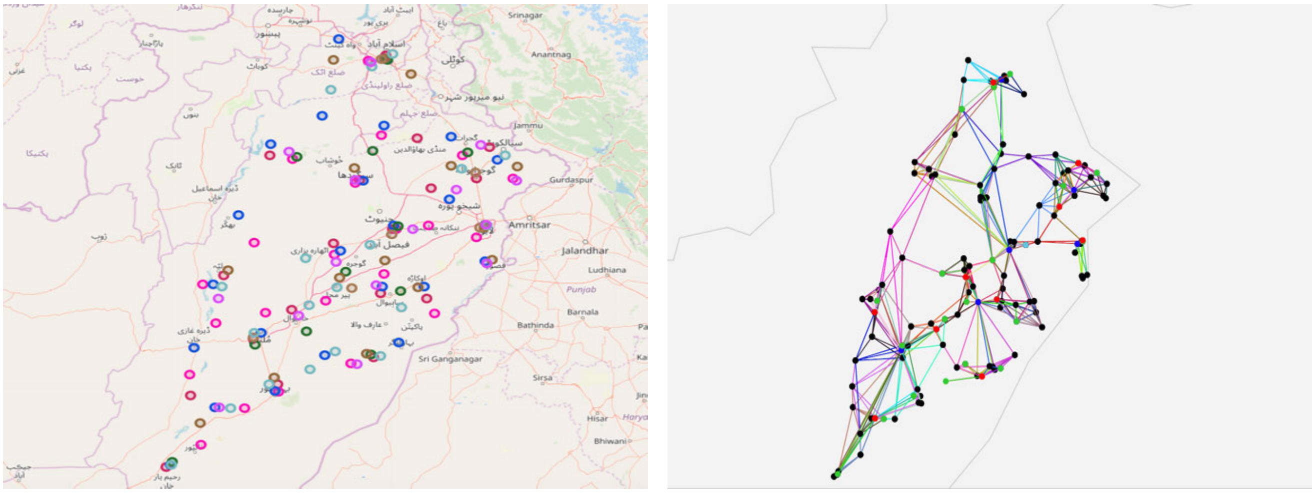
(c) A single hotspot of malria and scabies dieases and their dynamics with other diseases and each other

(d) Multiple hotspots of scabies and their dynamics with other disease hotspots

FIGURE 5. Dynamics of single and multiple hotspots of one or more diseases in an interactive graph. The graphs can be zoomed to multiple levels with interaction of multiple diseases and their selected hotspots.

head or centroid represents the respective disease hotspot.The figure shows that there is no indiscernible pattern that can be seen among the diseases over the region when hotspots are identified individually. It does however, identifies individual disease spatial hotspots for further analysis. Fig.6a shows the distribution of each disease hotspots across Punjab province after first layer of clustering.

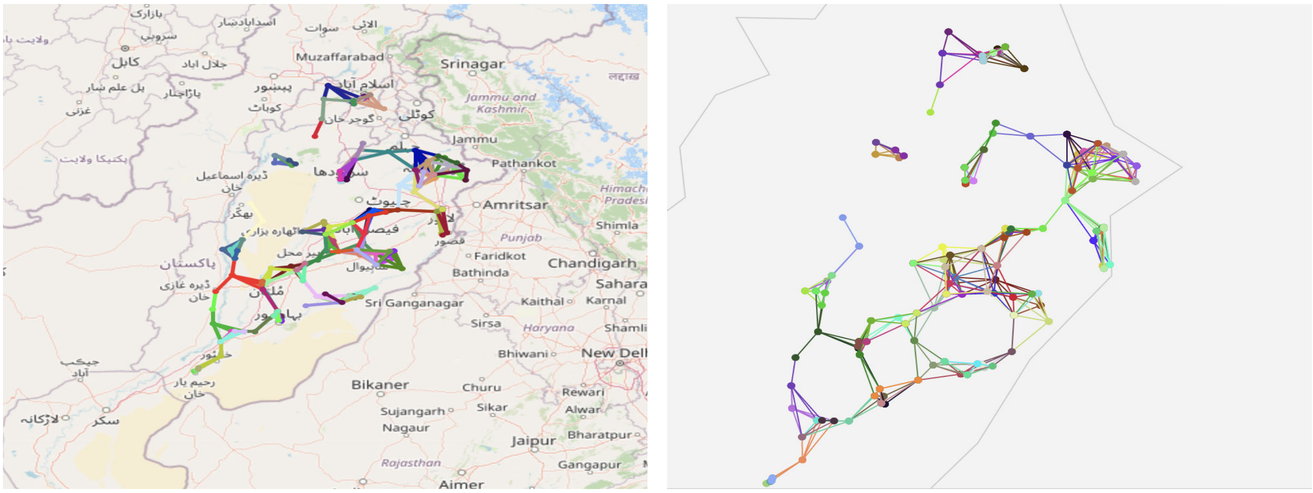
Initial analysis shows almost all disease hotspots near to each other, however detailed study in the areas show, that the hotspot locations are far apart and in some cases lie in a different district altogether. The distance graph at this stage is a mesh of interconnected hotspots with no meaningful pattern at a first glance. We create an interactive graph, that can be used to study interesting sub graphs at different zoom levels. For example, Fig. 5a shows one hotspot of malaria disease and its interaction with other all other diseases' hotspots whereas Fig. 5b shows another hotspot of malaria disease and its interaction with all other disease hotspots. Fig. 5c shows interaction of one cluster of malaria and scabies each and their interaction with each other and Fig.5d shows three clusters of scabies disease and its interaction with other diseases. Similarly all hotspots for every disease and its dynamics with



(a) Multiple disease hotspots after first clustering layer across Punjab

(b) Multiple disease hotspots inter linkages with closest every other other disease hotspot

FIGURE 6. Multiple disease hotspots in relation to one another.



(a)

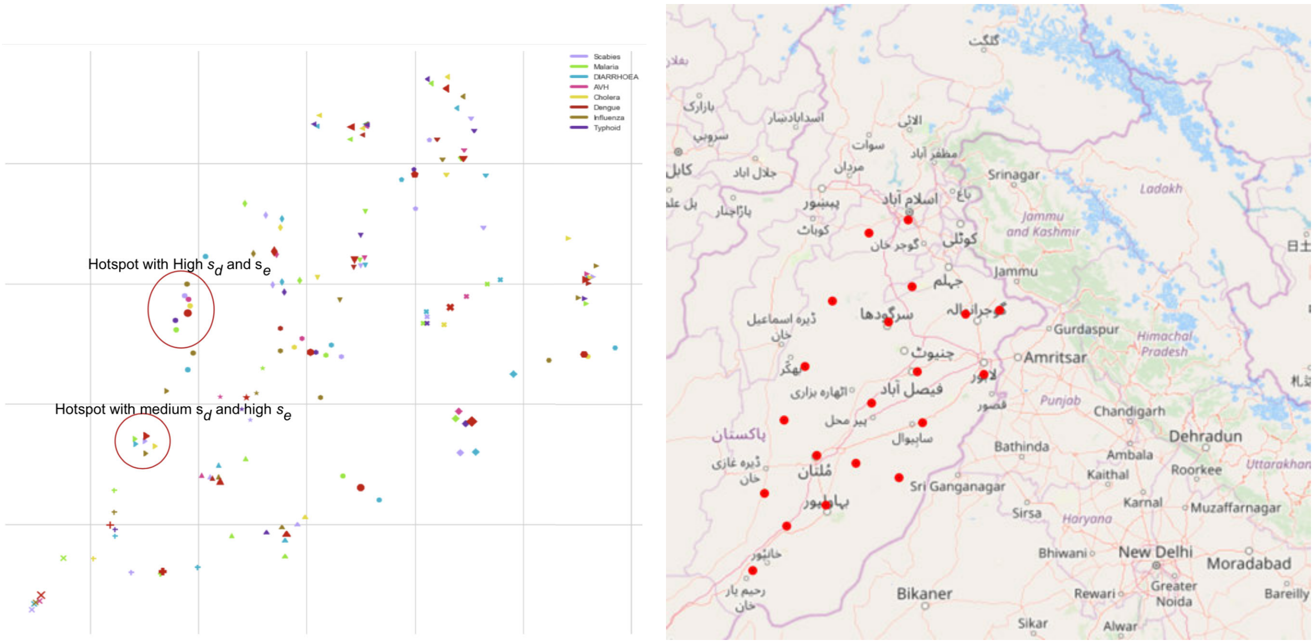
(b)

FIGURE 7. Multiple disease hotspots inter linkages graph pruned based on two different threshold values.

other disease hotspots can be viewed on the graph at different zoom levels.

It can be seen in Fig.6b, some of the disease hotspots are very close to each other, however, to give a quantitative measure to this concept,the algorithm is proceeded to apply neighbourhood tagging. During tagging, distance between hotspots of every disease is calculated with hotspots of every other disease. 28 distance matrices are generated for all 8 diseases. C_{1_1} represents cluster 1 for d_1 and H_{1_1} represents hotspot 1 for cluster 1 of d_1. Similarly, M_{1_2} represents distance matrix for hotspots of d_1 and d_2. The hotspots within each cluster is then tagged to create a distance based graph that is pruned based on threshold value. The tagging is done based on distances between each disease hotspot relative to other disease hotspots. The pairwise distance generates a heat map where threshold is applied to selected top m values. For this work, we have selected the threshold to be average inter district distance in the Punjab province. All disease edges with their corresponding nodes(hotspots) with the distance less than the inter district distance are excluded. This gives an interconnected graph of multiple diseases with hotspots of each disease connected to the closest hotspot of every other disease. If a disease hotspot does not have any other disease's hotspot in its neighbourhood, that is, not within threshold distance, it is removed from the graph.Fig.6b shows the pruned graph of hotspots based on threshold distance. Fig.7 shows the linkages among different disease hotspots when threshold is applied based on inter district distance.

The algorithm then steps into final layer of clustering where the clusters with H fulfilling the threshold constraints



(a) Layer 3 results for generating cluster alarms of medium and high intensities

(b) Geographical locations corresponding to cluster alarms

FIGURE 8. Layer 3 models for selecting resource starved locations by clustering selected individual disease hotspots and identifying resulting cluster alarms geographical locations.

are selected for every disease and clustered as a single file to generate a layer of cluster alarms. The cluster alarms are centroid of cluster of different diseases as its members. k means is applied at this layer and cluster alarms are classified as low medium and high and corresponding locations are identified for medium and high resource sensitive areas for critical public health intervention. Fig.8 shows the results of final layer of the model that includes the medium A values based on values of ordered pair (s_d, s_e) for (medium, high), (high, medium), (high,high), (low,high) through expert elaboration.

V. DISCUSSION

With disease data available from EHR, surveillance and other sources in fine resolution of space and time, it is imperative to study public health policy outcomes and their impact on population based on evidence. Identifying disease intensive areas has long been investigated using computer science techniques in general and machine learning in particular. While single disease streams dynamics with population exposure variables has proved beneficial for clinical settings, we are able to illustrate the significance of integrating the multiple diseases incidences individual streams through our novel approach in our framework.

The advantage of initiating the multi disease analysis with single disease is that it allows to retain original arrangement of events in the network. This means that we are able to represent the spread of disease locally, in terms of number of disease incidences, irrespective of other diseases. This information is especially useful when considering communicable diseases. In addition, as can be seen in Fig.3, events are widely variable among diseases and merging all spatial events for hotspot identification a single large cluster with high burden disease dominating and pulling the lower burden disease into them.

We identify the areas most affected by the communicable diseases from 38 districts. Table 3 details the 19 final cluster alarms and their corresponding districts. The size of cluster alarms is represented by the number of hotspots within each cluster alarm as well as number and type of disease distribution for each cluster alarm.

The adjacency matrix ensures the spatial integrity of the analysis. The algorithm takes into account, size and shape of cluster of identified cluster alarms based on number of diseases per cluster and total number of incidence on a geographical area as represented in Table 3. In health context, this information is valuable for determining causes in the environment, for example, presence of water bodies, construction work, and industrialization near the identified cluster alarms. In addition the intensity of cluster alarm is defined as low, medium or high in terms of the number of disease incidence individually as well as integrally.

The identification of location with multiple disease hotspots in proximity of each other is very critical for public health authorities for optimal allocation of resources as well as for identification of factors that are causing a particular disease spread or features common to most diseases based on correlation analysis.The model also reveals intermediate stages of disease dynamics with other diseases through interactive graphs. This analysis can significantly help in identifying most affected areas from multiple diseases so that further investigations into the causes present in environment, epidemiology and population characteristics can be conducted.

TABLE 3. Layer 3 Cluster alarms specification for the eight selected diseases including their container districts, population density of the district, number of hotspots in each cluster alarm with number of individual disease hotspots.

S.N.	District	Population Density/km^2	Hotspots	Scabies	Malaria	Gastroenteritis	AVH	Cholera	Dengue	Influenza	Typhoid
1	Rawalpindi	636	14	1	1	3	1	3	1	3	1
2	Multan	838	6	1	1	1	1	0	1	1	0
3	Lahore	3566	11	1	1	2	1	2	1	1	2
4	Sahiwal	576	7	1	1	1	0	1	1	1	1
5	RajanPur	90	7	1	2	1	0	1	0	1	1
6	Mianwali	181	5	1	1	1	1	0	0	0	1
7	Layyah	178	6	1	1	1	0	1	0	1	1
8	Bahawalnagar	232	8	1	1	1	1	1	0	2	1
9	Sialkot	903	11	2	1	2	1	1	1	1	2
10	Rahimyar khan	264	5	1	1	1	1	0	0	1	0
11	Gujranwala	939	9	2	2	2	1	1	0	0	1
12	Tobatek Singh	499	8	1	1	1	1	1	0	2	1
13	Nankana Sahib	458	7	1	2	1	1	0	1	0	1
14	Khanewal	476	6	1	1	1	1	0	0	1	1
15	Narowal	541	6	1	1	1	0	1	0	1	1
16	Okara	510	6	1	1	1	1	1	0	1	0
17	Lodhran	422	6	1	1	1	0	1	0	2	0
18	Dera Ghazi Khan	238	2	0	1	1	0	0	0	0	0
19	Bhakkar	129	2	0	1	1	0	0	0	0	0

In addition, in resource starved locations, it is significantly able to identify the areas that are hit by multiple problems.

VI. CONCLUSION

Using disease management models with ML based analysis for high density population and resource starved locations can reduce health expenditures. In order to evolve intervention programs based on research and technology advancement for cost effective disease management, it is important to allocate resources and device policies based on significant evidence. The presented model is flexible and can be scaled to include more parameters to represent other factors in disease outbreak in a given population. In future, more factors can be added to create other AI and ML based predictive models for multiple diseases to investigate their relationships in context of external factors. However, this work effectively provides a base model for managing multiple diseases to design better targeted intervention program in public health context. In addition to public health, the cluster alarm identification and analysis technique presented can be applied to other areas such as security, plant disease detection, environmental studies, social and sensor networks, and other domains with proper modifications.

REFERENCES

[1] D. J. Barnett, H. A. Taylor, J. G. Hodge, and J. M. Links, "Resource allocation on the frontlines of public health preparedness and response: Report of a summit on legal and ethical issues," *Public Health Rep.*, vol. 124, no. 2, pp. 295–303, Mar. 2009.

[2] N. Winters, S. Venkatapuram, A. Geniets, and E. Wynne-Bannister, "Prioritarian principles for digital health in low resource settings," *J. Med. Ethics*, vol. 46, no. 4, pp. 259–264, Apr. 2020. [Online]. Available: https://jme.bmj.com/content/early/2020/01/16/medethics-2019-105468

[3] L. Simonsen, J. R. Gog, D. Olson, and C. Viboud, "Infectious disease surveillance in the big data era: Towards faster and locally relevant systems," *J. Infectious Diseases*, vol. 214, no. 4, pp. S380–S385, Dec. 2016.

[4] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch, "Realtime influenza forecasts during the 2012–2013 season," *Nature Commun.*, vol. 4, no. 1, p. 2837, Dec. 2013.

[5] Y. Tseng and Y. Shih, "Developing epidemic forecasting models to assist disease surveillance for influenza with electronic health records," *Int. J. Comput. Appl.*, vol. 42, no. 6, pp. 616–621, 2019, doi: 10.1080/1206212X.2019.1633762.

[6] M. E. Woolhouse, A. Rambaut, and P. Kellam, "Lessons from ebola: Improving infectious disease surveillance to inform outbreak management," *Sci. Transl. Med.*, vol. 7, no. 307, Sep. 2015, Art. no. 307rv5.

[7] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.

[8] Z. S. Y. Wong, J. Zhou, and Q. Zhang, "Artificial intelligence for infectious disease big data analytics," *Infection, Disease Health*, vol. 24, no. 1, pp. 44–48, Feb. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2468045118301445

[9] T. B. Rodrigues, D. P. Salgado, M. C. Cordeiro, K. M. Osterwald, T. F. Filho, V. F. de Lucena, E. L. Naves, and N. Murray, "Fall detection system by machine learning framework for public health," *Procedia Comput. Sci.*, vol. 141, pp. 358–365, May 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918318398

[10] T. L. Wiemken and R. R. Kelley, "Machine learning in epidemiology and health outcomes research," *Annu. Rev. Public Health*, vol. 41, no. 1, pp. 21–36, Apr. 2020, doi: 10.1146/annurev-publhealth-040119-094437.

[11] S. K. Greene, E. R. Peterson, D. Kapell, A. D. Fine, and M. Kulldorff, "Daily reportable disease spatiotemporal cluster detection, New York City, New York, USA, 2014-2015," *Emerg. Infect. Dis.*, vol. 22, no. 10, pp. 1808–1812, Oct. 2016.

[12] E. C. Lee, J. M. Asher, S. Goldlust, J. D. Kraemer, A. B. Lawson, and S. Bansal, "Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference," *J. Infectious Diseases*, vol. 214, no. 4, pp. S409–S413, Dec. 2016.

[13] M. Rezaeian, "The concept of disease clustering for public health specialists," *Middle East J. Family Med.*, vol. 7, pp. 25–27, Oct. 2009.

[14] S. F. Olsen, M. Martuzzi, and P. Elliott, "Cluster analysis and disease mapping—Why, when, and how? A step by step guide," *BMJ, Brit. Med. J.*, vol. 313, no. 7061, pp. 863–866, 1996. [Online]. Available: http://www.jstor.org/stable/29733060

[15] K. M. Carley, D. B. Fridsma, E. Casman, A. Yahja, N. Altman, L.-C. Chen, B. Kaminsky, and D. Nave, "BioWar: Scalable agent-based model of bioattacks," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 36, no. 2, pp. 252–265, Mar. 2006.

[16] H. Zhang, L. Wang, S. Lai, Z. Li, Q. Sun, and P. Zhang, "Surveillance and early warning systems of infectious disease in China: From 2012 to 2014," *Int. J. Health Planning Manage.*, vol. 32, no. 3, pp. 329–338, Jul. 2017.

[17] S. Ullah, H. Daud, S. C. Dass, H. Fanaee-T, and A. Khalil, "An Eigenspace approach for detecting multiple space-time disease clusters: Application to measles hotspots detection in Khyber-Pakhtunkhwa, Pakistan," *PLoS ONE*, vol. 13, no. 6, Jun. 2018, Art. no. e0199176.

[18] K. Takahashi and H. Shimadzu, "Multiple-cluster detection test for purely temporal disease clustering: Integration of scan statistics and generalized linear models," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0207821.

[19] F. Khalique, S. A. Khan, W. H. Butt, and I. Matloob, "An integrated approach for spatio-temporal cholera disease hotspot relation mining for public health management in punjab, pakistan," *Int. J. Environ. Res. Public Health*, vol. 17, no. 11, p. 3763, May 2020, doi: 10.3390/ijerph17113763.

[20] F. Khalique, S. A. Khan, and I. Nosheen, "A framework for public health monitoring, analytics and research," *IEEE Access*, vol. 7, pp. 101309–101326, 2019.

[21] G. O. T. P. Punjab Information Technology Board. *Digital Punjab: Disease Surveillance System*. [Online]. Available: https://pitb.gov.pk/dss

FATIMA KHALIQUE received the M.S. degree in computer science from Uppsala University, Sweden, in 2007. She is currently the Ph.D. Scholar with the National University of Sciences and Technology (NUST), Islamabad, Pakistan. She has worked in industry and academia. She worked as a Lecturer with the National University of Sciences and Technology (NUST) and the National University of Modern Languages (NUML), Islamabad, Pakistan. She worked as a Software Developer with Zhonxing Telecom Engineering (ZTE) Islamabad. She is an Oracle Certified Professional. Her research interests include data mining, health informatics, artificial intelligence, data analytics, and machine learning algorithms.

SHOAB AHMED KHAN received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He has more than 22 years of industrial experience in companies in USA and Pakistan. He is currently a Professor of computer and software engineering (C&SE) with the NUST College of EME. He has founded the Center for Advanced Studies in Engineering (CASE) and the Center for Advanced Research in Engineering (CARE). CASE is a primer engineering institution that runs one of the largest post graduate engineering programs in the country and has already graduated 50 Ph.D. and more than 1800 M.S. in different disciplines in engineering, whereas CARE, under his leadership, has risen to be one of the most profound high technology engineering organizations in Pakistan developing critical technologies worth millions of dollars for organizations in Pakistan. CARE has made history by winning 13 PASHA ICT awards and 11 Asia Pacific ICT Alliance Silver and Gold Merit Awards while competing with the best products from advanced countries like Australia, Singapore, Hong Kong, and Malaysia. He is an inventor of five awarded U.S. patents and has more than 260 international publications. His book on *Digital Design* is published by John Wiley & Sons and is being followed in national and international universities. He served as a board member for Governance of many entities in the Ministry of IT and Commerce. He served as a member for the National Computing Council and the National Curriculum Review Committee. He has been awarded the Tamgh-e-Imtiaz (Civil), the Highest National Civil Award in Pakistan, the National Education Award 2001, and the NCR National Excellence Award in Engineering Education. He has served as the Chairman for the Pakistan Association of Software Houses (P@SHA).