International Journal of Scientific Research in Computer Science and Engineering
Vol.8, Issue.3, pp.01-05, June (2020)

Research Paper

E-ISSN: 2320-7639

Correlation study of New Cases, Deaths, Recoveries and Temperature with Machine Learning during COVID-19 spread in Saudi Arabia

Zafar Iqbal Khan[1*], Yasir Javed[2], Khurram Naim Shmasi[3]

[1]Dept. of Computer Science, College of CCIS/Prince Sultan University, Riyadh, Kingdom of Saudi Arabia
[2]Dept. of Computer Science, CCIS/Prince Sultan University, Riyadh, Kingdom of Saudi Arabia
[3]Dept. of Computer Science, Community College /King Saud University, Riyadh, Kingdom of Saudi Arabia

[*]*Corresponding Author: zkhan@psu.edu.*

Available online at: www.isroset.org

Received: 05/June/2020, Accepted: 22/June/2020, Online: 30/June/2020

***Abstract*—** Millions of people have been infected and killed by the recent outbreak of novel coronavirus throughout the world. It has affected 210 countries around the globe and two International conveyances [1]. It has evolved from epidemic to pandemic crossing all physical, socio economic and geographic barriers. The untraceable virus mutations can quickly effect hundreds of people before the antibodies are developed by the immune system. Since its first inception in Wuhan, the virus has rapidly influenced all nooks and corners of the tightly connected world. The imperative lethality of the virus varies from hot temperature to cold climates. There have been different impacts of COVID-19 on different races, geographical conditions and socio-cultural environments. It is evident that temperature is a critical factor for incubation period of pathogens. This research paper tries to find out the correlation among various factors such as infections, deaths, recoveries of the patients infected with COVID-19 with respect to the Temperature with the help of k-means clustering.

***Keywords*—** Coronavirus, COVID-19, k-means, clustering, Machine Learning, correlation

I. INTRODUCTION

Covid-19 virus belongs to a wider family of Coronaviruses that can further be classified into alpha, beta, gamma and delta. Out of these, 7 viruses are known to affect humans causing mild cold, catarrh and influenza to deadly respiratory problems [2]. Sometimes viruses found in animals undergo evolutions and affect Humans. Several virus epidemics viz. SARS, MERS, etc. were the result of such virus mutations [3]. A pneumonia of unknown type was reported by China to WHO in late December 2019 [4]. WHO issued notice regarding International Public Health Emergency on January 30, 2020 [5]. International Travelers and tourists spread the virus from China to more than 200 countries around the globe. By now, this Pandemic has caused staggering 207,813 deaths, with total reported cases exceeding 3011712 and total recovered cases 887982 [6].

Saudi Arabia reported its first corona casualty on March 02, 2020 [9]. Gradually disease spread to all major cities of the Kingdom. While Makkah, Jeddah and Riyadh shared most of the reported cases and deaths, other cities were also affected, some reported heavy while others fared lightly.

Use of Machine Learning techniques have seen gradual rise in medical science, particularly due to the availability of enormous amount of versatile data, like Alzheimer's disease Neuroimaging Initiative (ADNI), and UC Irvine Machine Learning Repository [10].

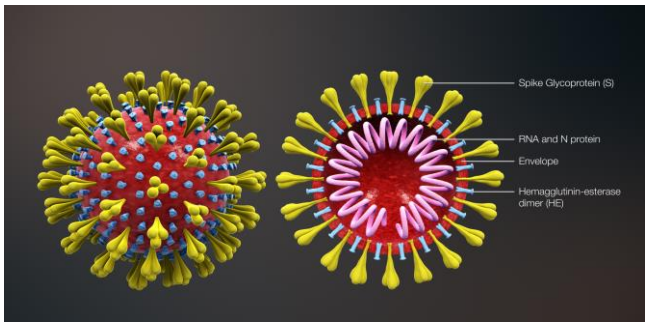

Figure 1. 3D image of Coronavirus, used from [7] under the Creative Commons licenses [8]. Showing structure Cross-Sectional view of nCov-19 the Spike S protein, HE protein, viral envelope, and helical RNA and N proteins.

Out of several Machine Learning techniques, Clustering is widely used for revealing structures in Data as it works for both labelled data as well as unlabeled data. The special feature of clustering is that it works very well on datasets where outcomes are not present, or simply relationship among data items is unknown [11].

Saudi Arabia, is a big country in Middle-East, mostly covered with desert although there are some exception of high mountains and Oasis also. Geographical and Natural structure makes Saudi Arabia a country with high temperature and relatively dry environment.

It was widely perceived that because of high temperature Coronavirus will not spread or will spread with relatively slow pace and will be less lethal. The work presented here aims to find correlation between Temperature and new COVID-19 cases reported and new Deaths reported in the Kingdom of Saudi Arabia.

Organization of paper comprises as following, Section II, discusses some other use of clustering techniques into Medical Data Mining. Section III, which mentions Methodology has two parts; part A discusses Data Collection phenomenon and part B shows experimental setup. Section IV, deals with result and discussion and Section V, is Conclusion and Future work and Section VI is References.

We used self-explanatory names for the data columns in our experimental setup, hence in all further mentions NewCases means New Cases reported, NewRecovery means patients recovered from the disease and NewDeaths means New Deaths.

II. RELATED WORK

Clustering Techniques have been used in a variety of medical diagnostics, especially diagnosis of diseases that leave effects on vital organs [12]. Nikas and Low [13] used clustering techniques for early diagnosis of Huntington's disease in mice. A variant of clustering algorithm viz. C-means was used by Polat to classify Parkinson disease [14]. Chen used clustering techniques along with feature selection for analyzing Clinical Breast Cancer data [15]. Yilmaz et al. used clustering techniques for diagnosis of Heart diseases and Diabetes [16]. Wu et al. presented use of Multiple fuzzy c-means Clustering algorithm in various medical diagnosis [17]. Trevithick and Painter, applied clustering techniques for data of Psychiatric patients and patients with mental ailment [18]. Nilashi and Ibrahim, used clustering techniques for early prediction of propagation of Parkinson's disease [19]. Gamberger et al. used clustering techniques to identify rapid and slow markers among male and female patients of Alzheimer's disease [20].

III. METHODOLOGY

The Ministry of Health, Kingdom of Saudi Arabia from its twitter handle @SaudiMOH tweets figure 2, city wise stats about daily new COVID-19 cases reported in the Kingdom along with new deaths reported, number of patients recovered from disease and total new cases in the Kingdom. These tweets were collected for the period of one month, and the relevant data was extracted from these tweets. Ten major cities were selected from this data for our study, geographically covering all four directions. AccuWeather website provides accurate and reliable temperature data for different locations [21].

A. Data Collection and Design

We followed @SaudiMOH and collected tweets for a period of over one month starting from March 22, 2020 - April30, 2020, and collected various information reported by the Ministry. These tweets had city wise daily new cases, deaths and recoveries of COVID-19 cases in the Kingdom figure 3. We added temperature data from the AccuWeather website for this period to the relevant data for the above mentioned time period. AccuWeather provide temperature data in the form of Maximum and Minimum Temperature for any given location with date.

Out of 10 cities for which data has been collected, we selected 6 major cities as number of COVID-19 cases in other cities were negligible and not regular. When this data was fed to weka, with a in-built feature of the tool, data was converted to an arff file figure 4.

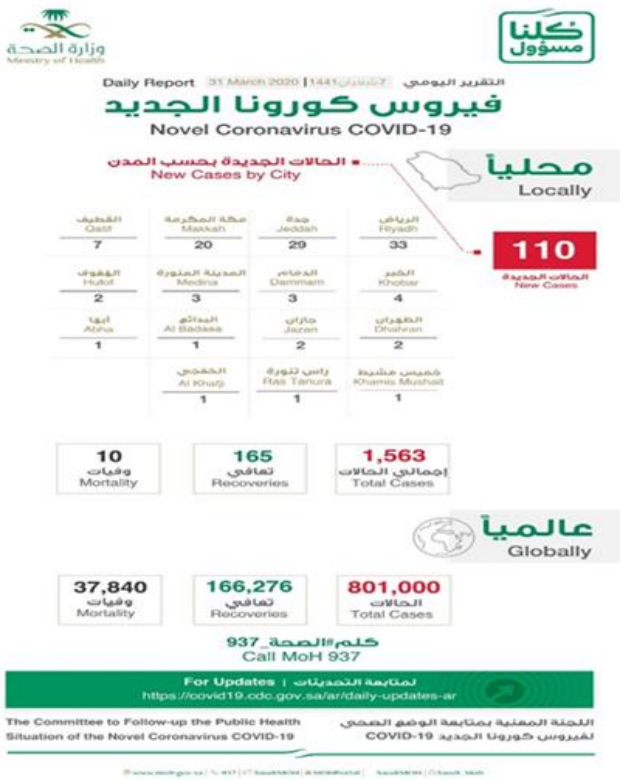

Figure 2: Image Tweeted by @SaudiMOH

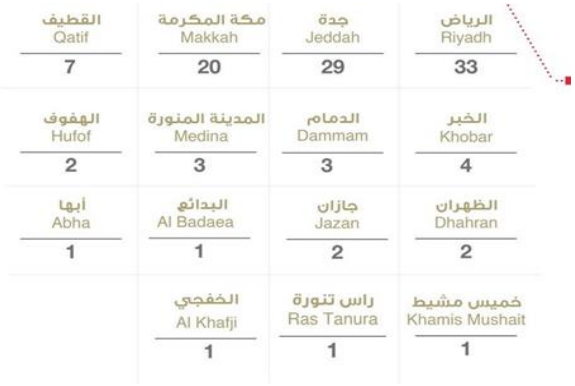

Figure 3: Magnified view of tweet from @SaudiMOH


```
@relation COVD-19CasesinSaudiArabiaLATEST

@attribute Date {22-Mar-20,23-Mar-20,25-Mar-20,26-Mar
Apr-20,2-Apr-20,3-Apr-20,4-Apr-20,6-Apr-20,9-Apr-20,1
Apr-20,17-Apr-20,18-Apr-20,19-Apr-20,20-Apr-20,21-Apr
Apr-20,28-Apr-20,29-Apr-20,30-Apr-20}
@attribute Makkah numeric
@attribute Madinah numeric
@attribute Jeddah numeric
@attribute Riyadh numeric
@attribute Dammam numeric
@attribute Jizan numeric
@attribute Tabuk numeric
@attribute Hofuf numeric
@attribute Qassim numeric
@attribute Khobar numeric
@attribute Taif numeric
@attribute Kharj numeric
@attribute MakHigh numeric
@attribute MakLow numeric
@attribute MadHigh numeric
```

Figure 4: Snapshot of Data in ARFF file

Here the city name corresponds to new COVID-19 cases reported in that particular city.

B. Setup

1) Clustering

There are several algorithms available for performing machine learning tasks. Clustering is such a technique exclusively used for unsupervised learning. Clusters are made out of given data without having any known relationship among them. The resultant clusters after the process have more intra cluster similarity than inter cluster similarities That means similar data items are more likely to be present in same clusters.

2) Simple k-Means Clustering

There are various clustering algorithms available but we applied simple k-Means clustering. It uses Euclidian distance based clustering mechanism and is comparatively faster than other clustering techniques on this data. Weka [22][23] was used as a tool for performing clustering on a Linux machine, with 16GB memory.

IV. RESULTS AND DISCUSSION

As the number of clusters are on experimenter's discretion, we found a reasonable run with number of clusters set to count 10. When clusters and cities were plotted, it showed nearly similar trends. Makkah, Madinah, Riyadh and Dammam are situated geographically apart and have reported high COVID-19 cases.

These cities have different temperatures but when clustering results were plotted for new cases, deaths and recoveries against temperature, approximately similar patterns were visible viz. figure 5, 6, 7 and 8.

The crest and trough visible in new cases, deaths and recoveries have obviously no dependencies on temperature which didn't changes significantly in the observation range.

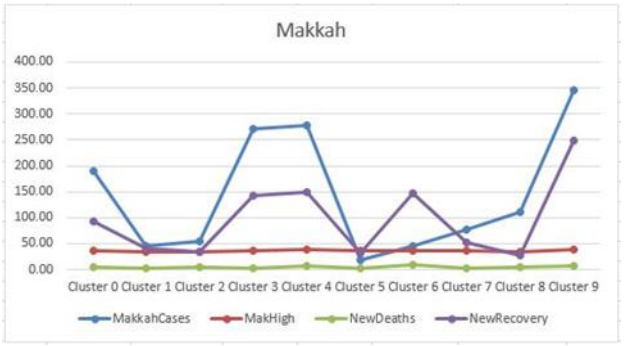

Figure 5: Clustering result for new cases, deaths and recoveries for Makkah

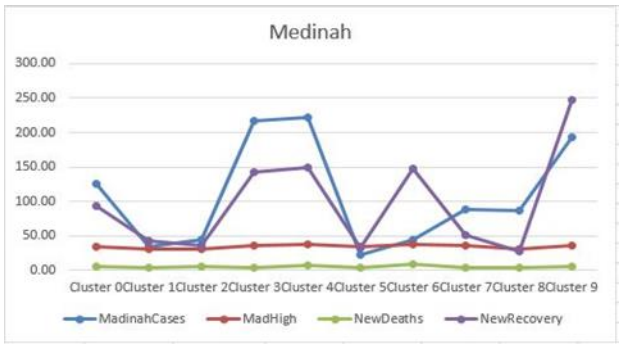

Figure 6: Clustering result for new cases, deaths and recoveries for Medinah

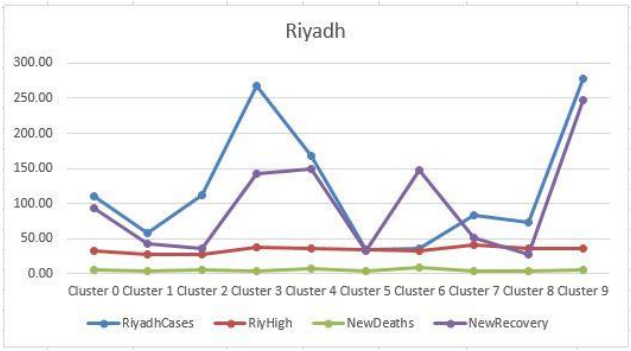

Figure 7: Clustering result for new cases, deaths and recoveries for Riyadh

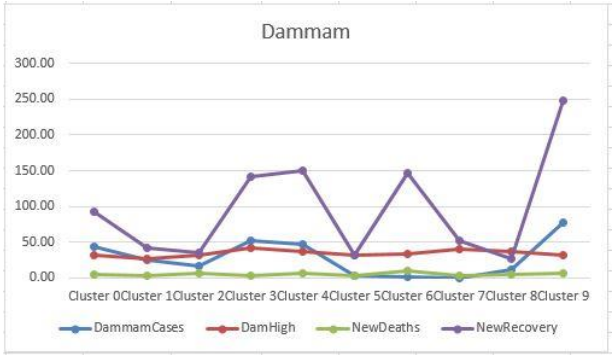

Figure 8: Clustering result for new cases, deaths and recoveries for Dammam

Clustering results, when tabulated and sorted for NewMortality, NewRecovery and TotalNewCases again do not show any adherence to dependency on temperature.

A. NewMortality and Temperature

Lowest NewMortality has values 2.78 and that is in cluster 1, where Makkah has highest value 26.94 and Riyadh 21.83 in *Table 4* for temperatures, while highest NewMortality value is 9 and that is in cluster 6 *Table 1*, Temperature values in cluster 6 for Makkah 30.50 and that of Riyadh 26.00.

B. NewRecovery and Temperature

Lowest NewRecovery has values 27.0 in cluster 8, corresponding temperature values in *Table 4* for Makkah and Riyadh are 28 and 31.25 while highest NewRecovery has values 247.67 in cluster 9, Temperature values from *Table 4* for Makkah and Riyadh are 30.83 and 28.83.

C. NewCases and Temperature

Lowest TotalNewCases is in cluster 5 with values 130.0, corresponding temperature values in *Table 4* for Makkah and Riyadh are 29.90 and 27.10 respectively whereas highest TotalNewCases is in cluster 9 with value 1249.67. Temperature values for Makkah and Riyadh are 30.83 and 28.83

Table 1: Cluster ID vs NewMortality

Cluster ID	NewMortality
Cluster 1	2.78
Cluster 5	2.80
Cluster 3	3.00
Cluster 7	3.00
Cluster 8	4.00
Cluster 0	4.66
Cluster 2	5.50
Cluster 9	6.00
Cluster 4	6.40
Cluster 6	9.00

Table 2: Cluster ID vs NewRecovery

Cluster ID	NewRecovery
Cluster 8	27.00
Cluster 5	32.00
Cluster 2	35.00
Cluster 1	41.65
Cluster 7	51.00
Cluster 0	93.13
Cluster 3	142.00
Cluster 6	147.00
Cluster 4	149.20
Cluster 9	247.67

Table 3: Cluster ID vs TotalNewCases

Cluster ID	TotalNewCases
Cluster 5	130.00
Cluster 6	214.50
Cluster 1	231.00
Cluster 2	262.50
Cluster 7	355.00
Cluster 8	373.00
Cluster 0	716.06
Cluster 3	1223.00
Cluster 4	1228.20
Cluster 9	1249.67

Maximum and Minimum temperature remain for short period of time, hence average temperature values *Table 4*, were taken to investigate the further correlation among parameters viz. NewMortality, NewRecovery and TotalNewCases.

Table 4: Clusters with Average Temperatures

Cluster ID	MakAvg	MadAvg	JedAvg	RiyAvg	DamAvg
Cluster 5	29.90	26.20	27.60	27.10	26.90
Cluster 6	30.50	30.00	29.75	26.00	25.75
Cluster 1	26.94	24.00	24.67	21.83	22.17
Cluster 2	26.75	24.00	25.50	23.25	26.75
Cluster 7	31.50	29.50	28.50	32.00	33.00
Cluster 8	28.00	24.75	26.25	31.25	31.00
Cluster 0	28.96	26.64	27.06	25.50	26.31
Cluster 3	30.00	29.00	29.00	31.50	36.50
Cluster 4	31.00	29.10	29.10	29.30	30.40
Cluster 9	30.83	28.83	29.00	28.83	27.17

Table 5: Correlation Matrix

	MakAvg	MadAvg	JedAvg	RiyAvg	DamAvg
Temp vs Mortality	0.24	0.41	0.46	-0.16	-0.26
Temp vs NewRecovery	0.58	0.69	0.71	0.23	0.11
Temp vs TotalNewCase	0.45	0.50	0.51	0.45	0.47

V. CONCLUSION AND FUTURE SCOPE

From all above discussions and data available from clustering results and distribution of correlation values, we can't establish a justified interdependencies of the studied parameters.

Interestingly, as opposed to initial consideration, no valid correlation could be found among New Cases reported, New Recoveries and New Deaths with Temperature. However, there might be some other factors e.g. Urban Planning, Socio-economic structures, Population density, morbidity history of patients etc. affecting these parameters and spread of disease, which needs further investigations.

REFERENCES

[1] COVID-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/ extracted on April 27, 2020

[2] The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. https://www.sciencedirect.com/science/article/pii/S0065352716300471?via%3Dihub

[3] Human Coronavirus Types. https://www.cdc.gov/coronavirus/types.html

[4] Rolling updates on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

[5] WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV). https://www.who.int/dg/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov)

[6] Coronavirus Worldwide Graphs. https://www.worldometers.info/coronavirus//worldwide-graphs/ extracted on April 27, 2020

[7] 3D medical animation corona virus. https://commons.wikimedia.org/wiki/File:3D_medical_animation_corona_virus.jpg

[8] https://www.scientificanimations.com (https://commons.wikimedia.org/wiki/File:3D_medical_animation_corona_virus.jpg), https://creativecommons.org/licenses/by-sa/4.0/legalcode

[9] MOH Reports First Case of Coronavirus Infection https://www.moh.gov.sa/en/Ministry/MediaCenter/News/Pages/News-2020-03-02-002.aspx

[10] Srivastav V, Issenhuth T, Kadkhodamohammadi A, de Mathelin M, Gangi A, Padoy N. (2018). MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. arXiv[Preprint].arXiv:1808.08180. Retrieved from: https://arxiv.org/pdf/1808.08180.pdf

[11] Alashwal H, El Halaby M, Crouse J J, Abdalla A, Moustafa A A. "The application of unsupervised clustering methods to Alzheimer's disease". *Front Comput Neurosci*, 13(issue): pp. **1-9, 2019**.

[12] Nithya N, Duraiswamy K, Gomathy P. "A survey on clustering techniques in medical diagnosis." *Int J Comput Sci Trends Technol*, 1(issue): **17-23, 2013**.

[13] Nikas J B, Low W C. "Application of clustering analyses to the diagnosis of Huntington's disease in mice and other diseases with well- defined group boundaries." *Comput Methods Programs Biomed*, **104(3): e133-e147, 2011.**

[14] Polat K. "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy 366 c-means clustering." *Int J Syst Sci*, **43**(issue): **597-609, 2012;**.

[15] Chen C -H. "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection." *Appl Soft Comput*, **20**(issue): **4-14, 2014**.

[16] Yilmaz N, Inan O, Uzer M S. "A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases." *J Med Syst*, **38**(**5**): **48, 2014**.

[17] Trevithick L, Painter J, Keown P. "Mental health clustering and diagnosis in psychiatric in-patients." *BJPsych Bull*, 39(issue):pp. **119-123, 2015**.

[18] Wu Y, Duan H, Du S. "Multiple fuzzy c-means clustering algorithm in medical diagnosis." *Technol Health Care*, 23(issue): pp.S519-S527, **2015**.

[19] Nilashi M, Ibrahim O, Ahani A. "Accuracy improvement for predicting Parkinson's disease progression." *Sci Rep*, 6(issue): 34181, **2016**.

[20] Gamberger D, Lavrac N, Srivatsa S, Tanzi R E, Doraiswamy P M. "Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease." *Sci Rep*, **7**(issue): pp.**63-67**, **2017**;.

[21] Siddiqui M K, Morales-Menendez R, Gupta P, Iqbal H, Hussain F, Khatoon K, Ahmad S. "Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis." *J Pure Appl Microbiol*, 14(Spl Edn.), **2020**.

[22] Eibe F, Mark AH, Ian HW. 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", *Morgan Kaufmann*, Fourth Edition, **2016.**

[23] Mark H, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H W. "The WEKA Data Mining Software: An Update." *SIGKDD Explorations*, 11(1): pp.**10-18, 2009**.

AUTHORS PROFILE

Zafar Iqbal Khan, pursed Bachelor of Science from Aligrah Muslim University, Aligrah, India in 2001 and Master of Science Aligrah Muslim University, Aligrah, India in year 2004. He is currently pursuing Ph.D. and currently working as Lecturer in Department of Computer Science, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Kingdom of Saudi Arabia since 2018. His main research work focuses on Network Security, Data Mining, and Deep Learning. He has 15 years of teaching experience.

Yasir Javed pursed Bachelor of Science and Master of Science from University of Management and Technology, and Ph.D. from Universiti Malaysia Sarawak, Malaysia and currently working as Assistant Professor in Department of Computer Science, College of Computer and Information Sciences, Prince Sultan University since 2012. He is a member of IEEE & IEEE computer society since 2013, a life member of the ACM since 2011. He has published more than 20 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. His main research work focuses on Robotics and IoT, Cyber Security, Big Data Analytics, Data Mining and Deep Learning. He has 15 years of teaching experience.

Khurram Naim Shamsi, pursed Bachelor of Science from Aligrah Muslim University, Aligrah, India in 1999 and Master of Science Aligrah Muslim University, Aligrah, India in year 2002. He is currently pursuing Ph.D. and currently working as Lecturer in Department of Computer Science, Community College, King Saud University, Riyadh, Kingdom of Saudi Arabia since 2008. His main research work focuses on Routing Protocols, Mobile Transactions and e-Payments. He has 17 years of teaching experience.

