Breast Cancer Diagnosis using Sequential Pattern Mining

Fokrul A. Mazarbhuiya, M. Y. Alzahrani, Ahmad H. Alahmadi

Abstract*: Cancer is a disease very common to both rural and urban peoples. It is the abnormal growth of some body cells which then destroy the normal functioning of surrounding cells. Cancer has different stages and can be cured easily if diagnosed earlier. Breast Cancer is widespread among the women of different age groups which results untimely death of so many ladies. As the cause of every cancer occurs before its actual appearance in the human body, sequential pattern from cancer datasets can be useful for determining the cause of Breast Cancer before its actual occurrence in the women body. In this article, we put forward a technique for digging out such patterns from Breast Cancer data. The effectiveness of the proposed technique is demonstrated by the experimental studies made with a real Breast Cancer dataset.

Keywords*: Mining Sequence, Frequent Sequence, Maximal Frequent Sequence, Disease, Symptoms of a disease, Signs of a disease, Breast Cancer.

I. INTRODUCTION

Cancer has been considered as a fatal disease. It is very much prevalent among people of different ages. In USA itself the cancer mortality rate is 163.5% as from 2011 to 2015 as National Cancer Institute data [1]. And the mortality rate is much higher among women, 198%. Out of these Breast Cancer contributes huge number of deaths. Breast Cancer is identified by its symptoms / signs like lump or mass, swelling, exasperation or dimpling in skin, soreness in breast or nipple, nipple retraction, reddish colour, scaly or thickened or discharge nipple. It has been found that if cancer is diagnosed earlier the possibility of prognosis increases proportionately. In the case of Breast Cancer the possibility is much more [2]. As the cause of a disease appears earlier than the actual occurrence of the disease, the same thing happens for Breast Cancer also. A Breast Cancer data is a collection of file of records of patients collected over a period of time where each record is consisting of symptoms / signs, the results of different pathological tests containing features of cancer, such as genetic mutations, cell signatures, DNA methylation on the patient's body etc. besides the patient's personal information and the records are put together in accordance with the date of visit of patients in the Cancer Hospital. Each file is associated with one patient. From such dataset if we are able to find the cause of occurrence Breast Cancer, the possibility of better prognosis will increase in manifolds. The sequential patterns can be successfully used for this purpose.

Sequential pattern from big data deals with finding order of occurrence events (a set of items). It is proposed in 1995, by *Agrawal* and *Srikant* [3]. In [4], authors put forwarded an efficient algorithm for finding sequential patterns. In [5], the authors have proposed a method of discovering sequential patterns from medical datasets. In [6], the authors have used classification rule based method for improving managerial efficacy in better dealing with cancer patients. The rules extracted by the method [6], can be directly incorporated to the healthcare system. In [7], the authors have proposed a method called differential sequential pattern mining to support insulin therapy. In [8], the authors have used DNA micro-array data to identify cancer patients.

In this article, we recommend an algorithm for the extraction of sequential patterns from Breast Cancer datasets which can be used for prediction of the Breast Cancer. The efficiency of our method is demonstrated with the experimental studies using a Breast Cancer dataset [9].

The article is structured in the following manner. In section-2, we present some related recent developments in this area. In section-3, we discuss the terminology utilized in the algorithm. In section-4, we present the algorithm proposed in this paper. In section 5, we talk about our findings and the conclusion is given in section-6.

II. RELATED WORK

Breast Cancer is the disease common to the ladies of different age groups and there are lots of untimely deaths of women due to Breast Cancer. In [10], authors have proposed a mining technique in DNA to identify the Breast Cancer. In [8], the authors have proposed a nice method of identifying the Cancer Patients. In [8], the classification based methods are introduced for the diagnosis of Cancer patients. In [11], the authors have made comparative studies on K-nearest neighbor, Naïve Baysian method and SVM based methods. In [12], R. J. Oskouei *et al* have made a comprehensive analysis of several works related to the applications of data mining on breast cancer diagnosis. In [13], authors have proposed a SVM based method which increases the accuracy of breast cancer diagnosis and minimizes errors.

Manuscript received on February 28, 2020.
Revised Manuscript received on March 22, 2020.
Manuscript published on March 30, 2020.

* Correspondence Author

Fokrul Alom Mazarbhuiya*, Department of Mathematics, School of Fundamental and Applied Sciences, Assam Don Bosco University, Assam, India, Email: fokrul_2005@yahoo.com

Mohamed Y. AlZahrani*, Department of Information Technology, AlBaha University, KSA, Email: imohduni@gmail.com

Ahmad H. Alahmadi, Department of Computer Science and Information,Taibah University, Medina, KSA, aahmadio@taibahu.edu.sa

In [14] authors have conducted a comparative study of applications of classification schemes for early detection of breast cancer. Similar works were done in [15], where it have been found that J48 decision tree is quite efficient in comparison to other classification based techniques. In [16], the authors have proposed a framework named XBPF for breast cancer diagnosis. In [17], V. Kumar *et al* proposed a differential classification based method for the predictions of malignant and benign breast cancer.

The sequence mining problem is proposed by in 1995 *Agrawal* and *Srikant* [3] while working super-market data. A couple of algorithms have been developed for this purpose and GSP [4] is one of them. It was then applied in many other fields like health care, education, text mining, telecommunication, intrusion detections etc [18]. Similarly it has been used for finding mining sequence pattern from medical datasets [5]. In [6], authors have applied classification rule and sequential patterns for improving managerial efficacy in healthcare system for better dealing with cancer patients. In [7], the authors have proposed a method for providing support to insulin therapy. In [19], the authors proposed a method for finding time-constrained sequential pattern from medical datasets. In [20], an analytic study on the navigation behavior integrating multiple source of website is done. In [21], the authors have discussed the technique of discovering actionable marketing intelligence in E-commerce scenario. In [22], the authors have developed a combined classifier with an extended tree structure and decision tree which can predict any desired customer event. In [23], the authors have tried to address the problem text categorization and proposed an approach called SPaC (sequential pattern for Categorization). In [24], the authors have proposed a system for the detection of misuse from attack data by enhancing the functionalities of Snort network-based detection system which has the ability to find out the sequential intrusion behavior. In this article we are proposing the sequential pattern mining method to identify the cause of breast cancer before its actual occurrence in the women's body. In other words, we are going find a model which can be used for prediction of breast cancer by providing it causes before its occurrence on the women's body.

III. TERMINOLOGY USED

In below we discuss some of terms, notations and notations used in this article.

Let us consider a set *S* = {ṩ$_1$, ṩ $_2$,… ṩ$_m$} consisting of *m* distinct symptoms / signs. A non-empty, unordered collection of one or more symptoms / signs is associated as disease. We denote a disease by (ṩ$_1$, ṩ$_2$,… ṩ$_k$), where ṩ$_j$ is a symptom / sign in *S*. For our convenience we take value of the symptoms or sign as binary. Also we will use symptoms, signs and diseases alternately.

A transaction corresponds to a collection of binary values where each value determines the presence or absence of a symptoms / sign. Thus each patient is associated with an ordered transaction sequence and is input to the algorithm. In below we discuss some terms of sequence mining of medical data which are used in our article.

A. Sequence

An ordered set of symptoms / signs is termed as sequence say ṩ and is represented by ṩ =($\delta_1\rightarrow$ $\delta_2\rightarrow\ldots$ $\rightarrow\delta_n$), where symptom/sign of a disease is denoted by δ_i. A sequence ṩ is an *n*-sequence, if $\sum|$ṩ$|$ is equal to *n*. By deleting some symptoms/signs/diseases from a sequence a subsequence can be constructed. If a sequence *s* is a subsequence of another sequence ṩ*'* then we say ṩ**'** supports ṩ.

Let us consider the dataset of input sequences, Π, where each sequence is a set of ordered transactions.

B. Frequency of a sequence

The total number of input sequence of Π which supports the sequence ṩ, is said to be the frequency of ṩ.

C. Maximal Sequence

A sequence whose frequency exceeds some user-specified threshold is called frequent sequence. Also a maximal frequent sequence is frequent sequence which is not contained in any other frequent sequence.

The justification behind the frequent sequences lies in extracting precedence and causal relationships within the diseases or symptoms which make them statistically notable.

IV. ALGORITHM PROPOSED

Sequential patterns (diseases) extraction process is composed of the following steps. A breast cancer data is in the raw form. It contains categorical, integer, real and also sometimes fuzzy attributes. Data preprocessing is applied at the beginning to change the raw form into a suitable type. It is to be mentioned here that we will consider all those parameters from the data which may contribute to any symptoms of breast cancer. Here each transaction will be a set of binary values where each value would correspond to a symptom/sign of breast cancer. So a transaction can be considered as a set of symptoms / signs which may indicate the presence or absence of breast cancer disease. Therefore each sequence of transactions corresponds to a record in Π. Thus actually the algorithm takes as input the sequences of set of symptoms / signs. The system architecture of the proposed whole process is given below in figure1.

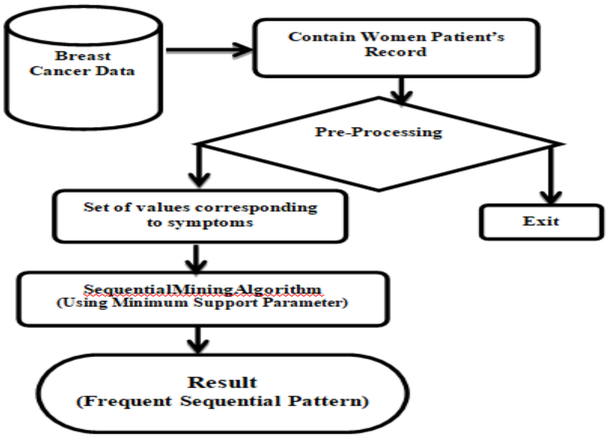
Breast Cancer Data
Contain Women Patient's Record
Pre-Processing
Set of values corresponding to symptoms
Exit
SequentialMiningAlgorithm (Using Minimum Support Parameter)
Result (Frequent Sequential Pattern)

Fig1: The system architecture

The algorithm or process goes through many passes. The pass1 finds out the support of each symptoms/ signs separately. At the end of pass1, the output will be frequent 1-sequence.

The pseudo code of frequent 1-sequence for finding algorithm is given below.

Algorithm1 (*For finding FS[1]= the frequent sequent of size-1*)

```
$CS[1]=\{ \delta[k];\ k=1, 2, ...n\}$ where $\delta[k]$ = kth symptoms/signs
for ($k=1;\ k\leq n;\ k++$)
set $\delta_{count}[k]=0$
for each t, input sequence in the dataset
for($k=1;\ k\leq n;\ k++$) do
{ if ($\delta[k]\subseteq t$) then
 $\delta_{count}[k]++$
}
 else if $\delta_{count}[k]\geq\theta$)
$\delta[k] \in FS[1]$
```

Every symptom / sign is having a support count δ_{count} whose value is 0 initially. If the symptom / sign is present in a transaction, then δ_{count} is increased. A frequent 1-sequence consists of a symptom/sign for which the sum total of δ_{count}, is greater than or equal to a pre-defined threshold. The course of action will be continued for all the symptoms/signs existed in the datasets.

Then a candidate generation process is used to generate candidate 2-sequence. After that the frequencies of 2-sequences are computed by making passes through the dataset and the procedure is repeated for higher sequences too. The pseudocode for *k*-sequence mining procedure is given below.

Algorithm2 (*For finding FS[k]= the frequent sequence of size-k*)

```
FS[1]=the sequence frequent sequence of size-1..
$k=2$
do while FS[k-1]!=Null
Generate candidate sequences CS[k]= set of candidate k-sequences, k-sequence is an ordered lists of symptoms/signs of the form $ṩ=(\delta[1]\rightarrow \delta[2]\rightarrow....\rightarrow \delta[k])$.
Prune (CS[k])
for all input sequence t in the dataset $\Pi$
do
increment count of $ṩ \in CS[k]$ if ($s\subseteq t$)
FS[k]=( $ṩ \in CS[k]$ such that frequency(s) surpass the threshold)
```

The algorithm has two essential sub-processes i) Generation of Candidate and ii) Pruning. The method of generating candidate for next level is quite analogous to one discussed in [3]. For this, any pair of frequent sequence is joined for which all the symptoms/signs upto previous but the last symptoms/signs are same. Likewise pruning sub-process checks the availability of all subsequences of the candidate in the previous level. If any of the subsequences is found to be missing in the prior level, the candidate will be removed. The pseudocodes for the two sub-processes are given below.

Candidate sequence generation with the given FS[k-1]

```
gen_candidate_sequences(FS[k-1])
$CS[k] := \varnothing$
  for all (k-1)-sequences $ṩ_{k-1} \in FS[k-1]$
   for all (k-1)-sequences $ṩ'_{k-1} \in FS[k-1]$
    if $\delta_{k-1}[1]= \delta'_{k-1}[1]\rightarrow \delta_{k-1}[2]= \delta'_{k-1}[2] \rightarrow...\rightarrow\delta_{k-1}[k-2]= \delta'_{k-1}[k-2] \rightarrow\delta_{k-1}[k-1]\neq \delta'_{k-1}[k-1]$
      then $s_k =(\delta_{k-1}[1] \rightarrow\delta_{k-1}[2] ... \delta_{k-1}[k-2] \rightarrow \delta_{k-1}[k-1]\rightarrow \delta'_{k-1}[k-1])$
           and $ṩ'_k =(\delta_{k-1}[1] \rightarrow\delta_{k-1}[2] ... \delta_{k-1}[k-2] \rightarrow \delta'_{k-1}[k-1]\rightarrow \delta_{k-1}[k-1])$
$CS[k] := CS[k] \cup \{ ṩ_k, ṩ'_k\}$
```

Pruning

```
prune(CS[k])
    for all $s_k\in$ CS[k]
    for all (k-1)-subsequences $s_{k-1}$of $s_k$ having same starting disease
      do
           if $s_{k-1} \notin FS[k-1]$
           then CS[k]= CS[k] $\sim\{s_k\}$
```

The above method supplies the set of frequent sequences where every frequent sequence is an ordered list of one or more symptoms/signs. Here a frequent sequence may fully or partially correspond to breast cancer or even may not correspond to breast cancer.

V. RESULTS AND DISCUSSIONS

For this purpose, we have used a Breast Cancer dataset [9] and the dataset is described in Table-I as follows.

TableI: Description of Datsets

Dataset	Dataset charc	Attribute charc	No. of instances	No. of attributes	Missing values	Area	Date of donation
Breast Cancer	Multivariate	Categorical	286	9	yes	life	15/05/1988

The Breast Cancer dataset contains 286 female patients of different categories based on class, age group, menopause, tumor-size, inv-nodes, node-caps, deg-malign, breast, breast-quad, irradiate, most of them are categorical attributes. Initially the data preprocessing is applied to make the dataset into sequences symptom / sign which are input to the process. Then the aforesaid algorithm is applied to the input sequences of varied sizes by taking the threshold value 0.4 (θ = 0.4).

The limited observation of the investigated outcome is articulated in tabulated form using Table-II, then graphically in Figure2 and finally using bar diagram in Fiqure2.

Table -II: Frequent sequences vs Datasets sizes

Dataset (Patient's record's no.)	Frequent sequential patterns no.
0	0
5	1
10	2
50	4
100	5
200	8
286	10

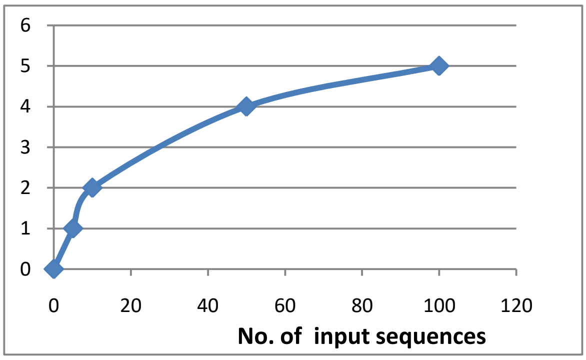

Fig2: Frequent sequence vs sizes of breast cancer datasets

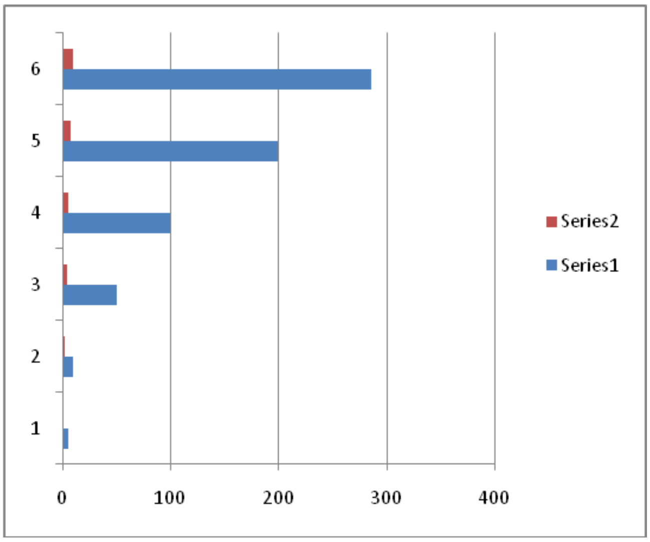

Fig3: Frequent sequence vs sizes of breast cancer datasets

From Table-II, it is examined that the number of frequent sequence is directly proportional number of patients. However, the rate of increase is steady after certain stage with respect to the rate of increment of number of the patients. The observation is expressed graphically in figure2 and figure3. In figure3, we consider have series1 for number patients given by green bar and series 2, for number frequent sequence given by red bar.

VI. CONCLUSION

An algorithm is discussed in this paper used for extracting patterns in the form of sequence from breast cancer datasets. A Breast cancer data is a gathering of records containing the breast cancer related information, about the patient from her first visit in the hospital to the last visit. The gathered records are ordered in according to the time or date of visit of patients in the hospital. The sequential patterns will let us know causal relationship of different symptoms / signs of disease in certain specified number of women breast cancer patients. The algorithm follows an A-priori based method where the sequences are extended by one symptom /sign after each iteration. Finally, the algorithm is implemented with a breast cancer dataset available in UCI machine repository.

The frequent sequence can be used to determine the causes of breast cancer before its actual occurrence on a women body. In other words it would be helpful in the prediction of breast cancer. This in turn would assist the medical practitioners to treat the breast cancer patient in a better way which would raise the chance survivability of the patients.

To extract the sequential patterns it's better to use a dataset collected over a long period time. However a breast cancer dataset may consists of the information about the patients after the occurrence of breast cancer in patient's body, so we can integrate the health data of the breast patients collected from the previous couple of years. It is not problematic as most of the advanced countries, the ministry of health is maintaining a centralized database of its people collected from the childhood or from their birth. If this can be done then more precise sequential patterns can be extracted leading to the better treatment of the disease. In future, the works can be done in the following directions:

- Use of healthcare data with the breast-cancer dataset.
- Seeking approaches other than level-wise approach.

REFERENCES

1. https://www.cancer.gov/research
2. https://www.cancer.org/cancer/pancreatic-cancer/detection-diagnosis-staging/survival-rates.html
3. R. Agrawal, and R. Srikant, Mining sequential patterns, *In Proc. of 11th Int'l Conf. on Data Engineering*, IEEE, (1995), pp.3-14.
4. R. Srikant and R. Agrawal; "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proc. of the 5th International Conference on Extending Database Technology (EDBT'96), Springer-Verlag, London UK, (1996).
5. M. Y. AlZahrani and F. A. Mazarbhuiya, "Discovering Sequential Patterns from Medical Datasets", Proceedings of 2016 International Conference on Computational Science and Computational Intelligence, IEEE Explore, Las Vegas, USA, (2016), pp. 70-73.
6. K. Choi, S. Cheng, H. Rhe and Y. Su, "Classification and Sequential Pattern analysis for improving Managerial efficiency and providing better medical services in public healthcare centre", Healthcare Informatics Research Vol. 16(2), (2010), pp. 67-76.
7. R. Deja, W. Froelich, and G. Deja, "Differential sequential patterns supporting insulin therapy of new onset type 1 diabetes", BioMedical Engineering online Vol. 14(1), (December 2015), pp. 2-11.
8. Z. S. Zubi, M. A. Emsaed, "Identifying Cancer patients using DNA micro-array Data in Data Mining Environment", Journal of Science and Engineering, Vol. 3, (2013), pp. 63-75.
9. R. S. Michalski., I. Mozetic, I. Hong, & N. Lavrac, The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann (1986).
10. S. Jawahar, and P. Sumathi, "An efficient Contiguous Pattern Mining Technique to Predict Mutations in Breast Cancer for DNA Data Sequences", International Bioinformatics and Biological Science, Vol 16 (n l p), (June 2018), pp. 35-41.
11. M. Shahbaz, S. Faruq, M. Shaheen, S. A. Masud, "Cancer Diagnosis Using Data Mining Techniques", Life Science Journal, Vol. 9(1), (2012).
12. Rozita Jamili Oskouei, Nasroallah Moradi Kor, and Saeid Abbasi Maleki, "Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges", Am J Cancer Res. (2017); 7(3): 610–627.
13. Medhat Mohamed Ahmed Abdelaal, Hala Abou Sena, Muhamed Wael Farouq and Abdel-Badeeh Mohamed Salem, "Using data mining for assessing diagnosis of breast cancer", Proceedings of IMCSIT 2010, IEEE Explore, (October 2010).
14. M. K. Keles, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study", Tehnicki Vjesnik 26(1): (February 2019), pp. 149-155.
15. Kehinde Williams, Peter Adebayo Idowu, Jeremiah Ademola Balogun, Adeniran Oluwaranti "Breast cancer risk prediction using data mining classification techniques", Transactions Networks and Communications, Vol. 3(2), (April 2015).

16. Ravi Aavula, R. Bhramaramba, "XBPF: An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, (June 2019).
17. Vivek Kumar, Brojo Kishore Mishra, Manuel Mazzara, Dang N. H. Thanh, and Abhishek Verma, "Prediction of Malignant & Benign Breast Cancer: A Data Mining Approach in Healthcare Applications", Proceedings of ICDSM 2019 To be published with-Springer, Lecture Notes on Data Engineering and Communications Technologies series, (2019).
18. M. Gupta, and J. Ha, "Applications of Pattern Discovery Using Sequential Data Mining", Pattern Discovery Using Sequence Data Mining: Applications and Studies edition, IGI Global, (2012) pp. 1-26.
19. M. Y. AlZahrani, and F. A. Mazarbhuiya, "Discovering Constraint-based Sequential Patterns from Medical Datasets", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, (November 2019), Vol. 8(4), pp. 724-728 .
20. B. A. Berendt, "Analysis of navigation behaviour in web sites integrating multiple information systems", The VLDB Journal, 9(1), (2000), 56-75.
21. A. G. Buchner, and M. D. Mulvenna, "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", SIGMOD Record, 27(4), (1998), pp. 54-61.
22. F. Eichinger, D. D. Nauck, and F. Klawonn, "Sequence Mining for Customer Behaviour Predictions in Telecommunications", Proceedings of the ECML/PKDD Workshop on Practical Data Mining (2006), pp. 3-10.
23. S. Jaillet, a. Laurent, and m. Teisseire, "Sequential Patterns for Text Categorization", Intell. Data anal., 10(3), (2006), pp. 199-214.
24. L. C. Wuu, c. H. Hung, and s. F. Chen, "Building Intrusion Pattern Miner for SNORT Network Intrusion Detection System", Journal of Systems and Software, 80(10), (2007), pp.1699- 1715.

AUTHORS PROFILE

Dr. Fokrul Alom Mazarbhuiya received his Ph.D. degree in Computer Science from Gauhati University (2007), India. He had been working in the College of Computer Science and IT at King Khalid University, and then Albaha University, Kingdom of Saudi Arabia since 2008 to 2018. He is currently working in Department of Mathematics, School of Fundamental and Applied Science, Assam Don Bosco University, India. His research interest includes Data Science, Information security, and Fuzzy logic.

Dr.Mohammed Y. Alzahrani has received his B.Sc. in Computer Engineering from Albaha Private College of Science and M.Sc. in Information Technology from Heriot Watt University, Edinburgh, UK. After this he has obtained Ph. D. degree in Computer Science from Heriot Watt University, Edinburgh, UK. Currently he is working as the Dean of College of Computer Science and Information Technology at Al Baha University, Al Baha, Kingdom of Saudi Arabia. His research interest includes Model Verification, Data Mining and Information Security.

Dr. Ahmed H. Alahmadi is an Assistant Professor in the Department of Computer Science and Information at Taibah University, Saudi Arabia. He has obtained his Ph. D. in Computer Science & Engineering from La Trobe University in 2014. Since then he has published couple of research papers in various peer-reviewed National and International Journals/Conferences. He also has a demonstrated history of working in higher education industry; he has worked as a Dean of Computer Science and IT College, Albaha University, KSA. Currently he is working as a Dean, Khaybar Community College, Taibah University. His research interest includes E-health, software engineering, business process modeling, requirements engineering, process mining, Information Security, and Machine Learning.