Reduce Prediction Time for HAI-Central Line Blood Stream Infection Using Big Data Mining Model

Omar Baeissa[1], Amin Y. Noaman[2], Abdul Hamid M. Ragab[3], Asmaa Hagag[4]

[1,2,3]Faculty of Computer Sciences & Information Technology, King Abdulaziz University Jeddah, SA
[4]Director of Infection Control & Environmental Health, King Abdullah Medical City Makkah, SA

Summary
This paper focuses on reducing prediction time for Central Line Associated Blood Stream Infection as one of the main types of Healthcare Associated Infection through a big data analytics model. There is 30,100 Central Line Associated Blood Stream Infection yearly in the US only. It is a severe infection that increases the mortality rate. Big data raises the bar as a result of additional features. It is mainly characterized by a tremendous amount of data that is composed of different forms. It also deals with the rapid data flow rate that is generated from multiple sources, and to top it off the quality of the data is questionable. There has been an increase in the infection rate of HAI during the past few years. Furthermore, the Centers for Disease Control and Prevention updated the definition. Prediction time reduction enables early intervention by clinical staff, which speeds up the recovery time and minimizes harm to the patient. Data mining approach consumes significantly less time, provides higher accuracy, and prevents personal subjective decisions. This paper compares seven data mining algorithms using real patient data of more than 28,000 cases from multiple sources. Naïve Bayes shows top accuracy result among other techniques.

Key words:
Big Data Analytics, Data Mining, Healthcare Associated Infections, Central Line Associated Blood Stream Infection.

1. Introduction

Large data sets that cannot be processed through traditional means are referred to as big data. Big data encompasses five main dimensions commonly known as the 5Vs. These are volume, velocity, variety, veracity, and value. Volume refers to the amount of data. In healthcare, for instance, data is growing exponentially. Patients' information, medical imaging, genomic details, 3D imaging, biometric data and medication information all account for enormous amounts of data. Newer technologies such as cloud computing aid in the management of this huge amount of data. The KPMG reports that healthcare data reached 150 Exabytes in 2013 and is growing at an annual rate of 1.2 Exabytes.

Different data sources and types are referred to as Variety. Healthcare related data originates from multiple sources. Traditionally, the main sources of data are hospital information systems, which encompass multiple sub-systems. Each one is dedicated to a different healthcare discipline. These sub-systems include Laboratory Information and Radiology Reporting systems that generate data in various formats and from various sources. Recent advancements have ushered personal wearable devices that record data and transfer it to clouds. Another rich source of data is social media. These sources generate data in varying formats. Generally, healthcare data is generated in structured, unstructured and semi-structured formats. Structured formats, which are composed mainly of clinical data, are easy to store, manage and analyze. Data formats characterized as unstructured and semi-unstructured include medical imaging, patient reports and physician notes. The rate of which the data is generated is referred to as Velocity, which in the case of healthcare is at classily high speed. Decision making in healthcare is dependent on the relevance of the data in hand. Data update frequency is an essential aspect in this regard. This illuminates the need for big data analytics for the early detection and prevention of infections or to simply minimize the side effects of a certain disease. Veracity refers to the quality of the data. Higher quality data is needed to obtain the highly accurate outputs needed for decision making in healthcare. Healthcare data is generated in various forms with different quality attributes. Value refers to the cost-effectiveness of generating the data. Is the data valuable to the customer? In most cases in healthcare, the value of the data is time dependent and may change from one day to the next [1].

The analysis of massive amounts of data to obtain valid, novel, potentially useful and understandable correlations and patterns is known as Data Mining. The demand for data mining in the healthcare sector has increased due to several reasons. The first of these reasons is the massive amounts of data generated in healthcare. This staggering amount of data is generated from high-throughput sequencing platforms, real-time imaging, the point of care devices, wearable computing and the workflow changes from paper-based processes to paperless systems. The second reason is the digitization of healthcare processes. This is a sound investment since it saves time as well as providing more intuitive data management tools. Data mining is an essential tool for the development of

Manuscript received May 5, 2019
Manuscript revised May 20, 2019

automated healthcare systems. It is also utilized for making better decisions and providing better treatments. Through data mining, disease detection accuracy increases while detection costs decrease [2].

Different data mining techniques applied across the field of healthcare are classification, regression, and clustering. Classification is the most popular technique used in healthcare. It is based on creating targeted classes from existing data sets and assigning new instances to these classes based on similarity. Regression is a mathematical technique that defines functions that find the correlation between multiple variables. It classifies the data into linear and non-linear sets. One of the limitations of this technique is its unsuitability for categorized data. Weighted Support Vector Regression (WSVR) is an example of applying the regression technique in healthcare. Another example is the Regression Decision Tree, which is used for predicting the number of hospitalizations. Clustering technique is the third approach for data mining in the healthcare sector. It is an unsupervised learning process that defines independent variables. Pre-defining the desired classes is not a mandatory step, unlike the supervised classification technique where this step is obligatory. Bypassing this step renders this technique suitable for use for exploring large amounts of data with minimal understanding. Although clustering works similarly to classification in that it groups data based on similarity, it differs in the fact that it is descriptive rather than predictive [3].

There are multiple ways to enhance data mining techniques accuracy and performance. The first example combines multiple techniques during a classification process. The overlap created by the combination of techniques ensures that each limitation is compensated for. This approach was applied in [4] where the author combined Naïve Bayes and Decision Tree algorithms to predict a medical condition. The error percentage was reduced due to the overlap of methods. Another method of enhancing is the attribute ranking and reduction. This method is used to increase accuracy using feature selection to reduce lower ranked attributes. It was applied by the author in [5] where the accuracy was enhanced by the reduction of attributes from 14 to 10. This improve3ment in accuracy indicated the negative impact lower ranked attributes have and their misleading effect on the classification process.

Healthcare-Associated Infections (HAI) are infections acquired during the patient's hospitalization period as a result of conducting one or more treatment procedures. They are also known as nosocomial or hospital-acquired infections. HAI has a negative impact on the patient as well as the hospital. It may increase the patient's period of hospitalization, which would imply additional expenditure of effort and resources. The most common types of HAI are Central Line Associated Blood Stream Infections, (CLABSI), Surgical Site Infections (SSI), Catheter-Associated Urinary Tract Infections (CAUTI), and Methicillin-Resistant Staphylococcus Aureus (MRSA) infection. The early detection of these infections and the pre-defining of susceptible patients affords a chance to apply the necessary protocols proactively. This can be achieved through effective clinical surveillance programs that predict suspected cases through verifying patient data [6]. According to the latest CDC report [7], 5 to 10% of inpatients in the US acquired HAI; this amounts to approximately 1.7 million cases. It causes the death of 99,000 patients with treatment cost reaching $20 billion. Although there is a tremendous decrease in CLABSI cases, still there is 30,100 Central Line Associated Blood Stream Infection yearly in the US only. It is a severe infection that increases the mortality rate [11]. In January 2019, the CDC defines CLABSI criteria as follow:

1. An eligible catheter in place for >48 hours
2. LCBI: Laboratory Confirmed Bloodstream Infection that is not related to any other infection

2. Related Works

This section aims to provide a critical analysis of previous publications that tackle similar objectives. It includes the utilization of different data mining methods applied in HAI prediction.

The author in [8] proposes the automation of surveillance and diagnosis of associated healthcare infections by applying a rule-based reasoning system. The system is based on static and dynamic rules. The static group is designed as prior knowledge defined by domain expertise with rules expressed in a Decision Tree format. The second group is dynamically generated by the system during the learning process which uses PART for rules induction. The issue of processing unstructured data was overcome by employing Naïve Bayes algorithms. The model was assessed over a period of 10 months on the Spanish National Health System using 2569 samples belonging to 1800 patients. It evaluated three different configurations using experts rules stand-alone, dynamic rules, and a combined version. The model deals with 3 of the 5 Vs of big data; a massive amount of data (Volume), different types of data (Variety) and data speed (Velocity). The acceptable prediction results were limited to urinary infections only (accuracy 93.4%, Sensitivity 97.6%, Specificity 93%). Additionally, the author indicates that the model requires additional data resources.

As an alternative to the most commonly used decision tree algorithm, the author of [9] proposes a multivariable regression model. The model is based on the weighted regression formula that estimates the infection probability based on an acceptable threshold defined by the user. The model reaches a sensitivity of more than 80% using five

variables with reasonable performance and 33% reduction of manual prediction workload. This technique has the advantage of considering multiple factors simultaneously. Also, higher predictive values are ensured by introducing weighted factors. Finally, it offers flexibility in terms of balancing sensitivity and efficiency because of the probability threshold. The paper highlights multiple challenges such as the need for high-quality data, ongoing monitoring of the data quality, and handling missing data. The solution depends on microbiology and pharmacy results, which are not covering every aspect of the CDC definition such as the admission period, temperature, catheterization period, and surgical procedure details. Finally, one of the main decisive factors of the model is the adjustability of the probability threshold that is also a significant limitation since it requires advanced programming skills that are not usually available in healthcare professionals. The author of [6] developed a surveillance system to detect CLABSI HAI based on a data mining technique AdaBoost. The detection reached an accuracy of 89.7% through the utilization of the AdaBoost method. The solution was assessed with a dataset from the US National Healthcare Safety Network and Consumer Survey. This system provides efficient infection control within healthcare facilities and improves infection detection accuracy. It was developed based on the cross-industry standard process CRISP data mining methodology. The CSP application fulfills two functions; it monitors HAI indicators and alerts healthcare practitioners to the possibility of infection. The solution incorporated 48 attributes designed for the prediction of HAI. It uses Rapid Miner tool to evaluate six data mining algorithms; these are Logistic Regression, Naïve Bayesian Inference, Multilayer Perceptron, Support Vector Machine, Random Forest, and AdaBoost. The accuracy was measured in terms of absolute relative error (ARE) where AdaBoost scored the highest result of 89.7%.

The author of [10] represents a real-time rule-based prediction model for HAI. The assessment took place in the Beijing hospital with a capacity of 3500 beds and 270 surgical operations per day. 12,000 patients per month average of ten days per stay. The data used includes laboratory results, serological and molecular testing, imaging reports, and fever history. The results exhibited 98.8% sensitivity, 93% specificity, and timesaving efficiencies were about 200 times those of traditional methods. One of the significant drawbacks of the solution was the exclusion of additional critical factors contained in medical records. Additionally, the study did not include all 3,500 patients nor did it detect mild infections during the early stages.

Table 1 includes the features and types of all related works reviewed in the previous section. Although the first work intended to predict all HAI types and managed to work with 3 Vs of big data, it showed acceptable results for CAUTI only and required additional resources to improve prediction. The second model covered the scope of multiple data sources, but the size of the dataset, as well as other relevant factors, were not considered. The third algorithm required larger datasets to validate the result reliability. The last approach indicated perfect overall accuracy levels while neglecting early stages of the infection as well as not utilizing the whole dataset.

Table 1: related works summary

LR	type	Algorithm	Dataset	Big Data	Eval
LR1	HAI	PART and NB	2569 samples 1800 patients	Volume, Velocity, Variety	Acc 93%
LR2	HAI	Multivariable regression	ND	Multiple sources	Sen 98%
LR3	CLBSI	AdaBoost	US national healthcare	ND	Acc 89.7%
LR4	HAI	Rule-Based	3,500 patients, 270 surgery/day	Volume, Variety	Sen 98.8% Spec 93%

3. Data Set

This section describes the data set used to build and assess the model. The dataset used consists of real patient data collected from the Infection Control and Environmental Health Department at King Abdulla Medical City (KAMC) in Makkah. KAMC operates with a total capacity of 550 beds with an emphasis on cardiology, neurology, oncology, and specialized surgeries. It contains 28,972 cases dating from January 2013 to November 2018 while also covering data for three main types of HAIs; CAUTI, CLABSI, and SSI. The data originates from 4 primary sources: hospital information system, laboratory information system, radiology information system, and physician notes. It contains 36 attributes such as admission date, location, sample type, lab results, radiology findings, and date of each event.

4. Data Mining Model

This section describes the performed data mining modeling process and assessment parameters.

In the data preparation phase, the attribute construction method was used to create additional attributes to calculate the period between each event and admission date, while another attribute showed the time between sample collection and date of sample collection. Other data preparation methodologies were applied such as hierarchy concept and numerosity reduction to improve the data quality. In the end, the total number of attributes is 47 attributes. A subset of attributes was selected to perform the assessment according to the CDC defined criteria for SSI, CAUTI, CLABSI and clinical needs as well as multiple trials of adjustment.

Based on the survey for the most suitable algorithms in healthcare and previously used techniques to predict HAI seven DM algorithms were nominated for evaluation. The decision tree was the top classifier in previous works. Random Forest (RF) and Gradient Boosted Trees (GBT) are different forms of DT. AdaBoost was used previously to assess CLABSI. Generalized Linear Model (GL) is one of the regression algorithms. Additionally, Naïve Bayes and Deep Learning (one of the neural network type) are suitable but have never been used for such cases. Finally, optimization performed for all the seven algorithms each one based on its internal criteria.

Table 2 demonstrates the summary of the performance results. Six measures were selected to evaluate the models. The accuracy of prediction, which counts the correctness of classified cases, False negative (FN) measures how many true positive cases predicted as negative, while false positive (FP) shows the opposite. Classification error (Class.E) that is calculated by the total number of errors divided by the total number of cases. Kappa shows how much the classification is done by chance, which is represented via a figure from zero to one where one represents no classification by chance. Finally, the run time of the algorithm in seconds.

5. Comparison

During the attribute selection process the models evaluated with two sets of attributes. The first set contains the final result of blood culture while the second did not. The aim of excluding the lab result is to improve the prediction time since a blood culture sample required five days to show an outcome.

The demonstrated assessment results in this section according to the final set of attribute that omits the culture output. Figure 1 illustrates the accuracy of each one of the algorithms. Naïve Bayes (NB) is the most accurate technique with 97.87% comparing to the most previously used DT (14.89%). In the healthcare domain, it is very crucial to avoid missing positive cases, which is represented by the false negative parameter. It is demonstrated in figure 2 for all of the assessed algorithms. NB shows the least percentage of missing positive cases (2.12%), and while DT (85.1%). On the other hand, NB shows the highest false positive results 0.94%, which is an acceptable result. For the medical field confirming false positive cases is more favorable than missing a positive case. NB kappa measures are more than 0.6, which indicate less prediction by chance while less than 0.2 in case of DT. Finally, the run time first place goes to NB since it shows the minimal runt time (0.27 seconds).

Table 2: Performance Summary

Algorithm	Accuracy	FP%	FN%	Class.E%	Kappa	Run T
ADB	14.89	0.1	85.1	0.78	0.173	0.58
DL	87.23	0.75	12.76	0.85	0.624	1.06
DT	14.89	0.1	85.1	0.78	0.173	0.28
GBT	76.6	0.32	23.4	0.51	0.715	0.81
GL	76.6	0.44	23.4	0.63	0.666	0.38
NB	97.87	0.94	2.12	0.95	0.627	0.27
RF	74.47	0.39	25.5	0.59	0.672	2.00
ADB-Opti	21.28	0.1	78.72	0.73	0.272	6.23
DL-Opti	91.49	0.7	8.51	0.76	0.66	1.47
DT-Opti	0	0	100	0.8	0	10.53
GBT-Opti	76.6	0.32	23.4	0.51	0.708	21.00
GL-Opti	76.6	0.44	23.4	0.63	0.65	3.63
NB-Opti	97.78	0.64	2.12	0.66	0.698	0.59
RF-Opti	74.47	0.38	25.53	0.58	0.655	35.90

As a conclusion, the data mining model build with excluding the final lab culture results. This improves the overall prediction time with five days less comparing to other DM models using this attribute. This enables early intervention by physicians and minimizes the impact of the infection. In the assessment of the suggested DT models, Naïve Bayes outperforms other algorithms including the most commonly used to detect HAI and in healthcare in general. NB shows the highest accuracy with a less false negative result in minimal time.

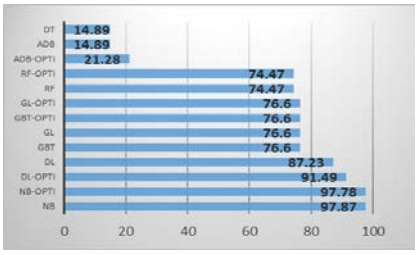

Fig. 1 Accuracy

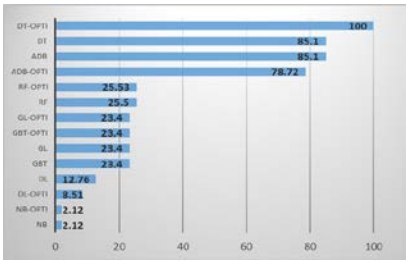

Fig. 2 False Negative

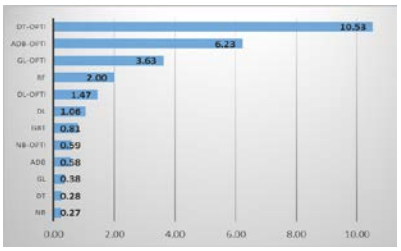

Fig. 3 Run Time

6. Summary

This paper reviews different data mining techniques applied to predict HAI-CLABSI. It discusses the strengths and weaknesses of each of the models. Then, utilizes real patient data to assess several DM algorithms including the most in use. As a result, the suggested NB algorithm superiority the other algorithms. Additionally, the suggested model reduces the prediction time five days in comparison to other DM models that use the laboratory results. Next step is to implement the model in the hospital environment and evaluate the prediction on the real data stream.

Acknowledgment

Acknowledge to King Abdulaziz University, Faculty of Computer Sciences and Information Technology for support and providing the required environment. Also, acknowledge to King Abdullah Medical City in Makkah, for their corporation and proving the data set to build and assess the model. Special thanks to Dr. Asmaa Mustafa, Mr. Tariq Awad, and Mr. Mohammed Alkam as co-investigator team.

References

[1] H. M. H. A. M. a. T. N. H. Asri, "Big Data in Healthcare: Challenges and Opportunities," in International Conference on Cloud Technologies and Applications, 2015.
[2] T. Anand, "Data Mining in Healthcare Informatics: Techniques and applications," in International Conference on Computing For Sustainable Global Development, INDIACom, 2016.
[3] P. Ahmad, "Techniques of Data Mining In Healthcare: A Review," International Journal of Computer Applications, vol. 120, no. 15, 2015.
[4] S. Alfisahrin, "Data mining techniques for optimization of liver disease classification," in International Conference on advanced computer science application and technologies, 2013, 2013.
[5] S. Sabab, "Cardiovascular Disease Prognosis Using Effective Classification and Feature Selection Technique," in International Conference on Medical Engineering, Health Informatics and Technology MediTec, 2016.
[6] A. Y. Noaman, A. H. Ragab, N. Farrukh, and A. Jamjoom, "improving Prediction Accuracy of (Central Line-Associated Blood Stream Infections) Using Data Mining Models," BioMed Research International, vol. 2017, no. Sep 2017, p. 12, 2017.
[7] CDC, "Preventing Healthcare-Associated Infection," CDC, [Online]. Available: https://www.cdc.gov/washington/~cdcatwork/pdf/infections.pdf. [Accessed 10 Mar 2018].
[8] H. Gómez-Vallejo, "A case-based reasoning system for aiding detection and classification of nosocomial infections," Decision Support Systems, vol. 84, pp. 104-116, April 2016.
[9] M. S. M. v. Mourik, "Automated Surveillance for healthcare-associated Infections: Opportunities for Improvement," Healthcare Epidemiology, vol. 57, pp. 85-93, 2013.
[10] M. Du, "Real-time automatic hospital-wide surveillance of nosocomial infections and outbreaks in a large," BMC Medical Informatics and Decision Making, vol. 14:9, 2014.
[11] CDC, "Bloodstream Infection Event (Central Line-Associated Bloodstream Infection and Non-central Line-Associated Bloodstream Infection)" CDC, [Online]. Available: https://www.cdc.gov/nhsn/PDFs/pscManual/4PSC_CLABScurrent.pdf. [Accessed 15 Jan 2019].

Omar Baeissa
Ph.D. Candidate in CS, King Abdulaziz University, SA
MSc in Information Technology, University of Stirling, UK
BSc in Medical Technology, King Abdulaziz University, SA
American Society of Clinical Pathology Board Certified (BB)